DIEGO DANTAS ALMEIDA

**Sequenciamento do genoma da serpente
Bothrops jararaca para caracterização da
estrutura gênica de toxinas**

Tese apresentada ao Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, para obtenção do título de Doutor em Biotecnologia.

São Paulo
2016

DIEGO DANTAS ALMEIDA

**Sequenciamento do genoma da serpente *Bothrops jararaca*
para caracterização da estrutura gênica de toxinas**

Tese apresentada ao Programa de
Pós-Graduação Interunidades em
Biotecnologia USP/Instituto
Butantan/IPT, para obtenção do
título de Doutor em Biotecnologia.

Área de concentração:
Biotecnologia

Orientador: Dr. Inácio de Loiola
Meirelles Junqueira de Azevedo

Versão original

São Paulo
2016

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)

Dantas Almeida, Diego

Sequenciamento do genoma da serpente Bothrops
jararaca para caracterização da estrutura gênica de
toxinas / Diego Dantas Almeida; orientador Inácio
de L. M. Junqueira de Azevedo . -- São Paulo, 2016.
126 p.

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

1. Genômica. 2. Toxinas. 3. Veneno. 4. Bothrops
jararaca. I. de L. M. Junqueira de Azevedo , Inácio
, orientador. II. Título.

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS**

Candidato(a): Diego Dantas Almeida.

Título da Tese: Sequenciamento do genoma da serpente *Bothrops jararaca* para caracterização da estrutura gênica de toxinas.

Orientador(a): Dr. Inácio de Loiola Meirelles Junqueira de Azevedo.

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:

COMISSÃO DE ÉTICA NO USO DE ANIMAIS
INSTITUTO BUTANTAN

Av. Dr. Vital Brazil, 1500, CEP 05503-900, São Paulo, SP, Brazil
Telefone: (55) (011) 2627-9585 - Fax: (55) (011) 2627-9505
ceuaib@butantan.gov.br

CERTIFICADO

Certificamos que o projeto intitulado "Sequenciamento do genoma da serpente *Bothrops jararaca* para caracterização da estrutura gênica de toxinas e seus elementos regulatórios", protocolo nº 1131/13, sob a responsabilidade de Inácio de Loiola Meirelles Junqueira de Azevedo e Diego Dantas Almeida – que envolve a criação e/ou utilização de animais pertencentes ao filo Chordata, subfilo Vertebrata (exceto o homem), para fins de pesquisa científica – está de acordo com os preceitos da Lei nº 11.794, de 8 de outubro de 2008, do Decreto 6.899, de 15 de julho de 2009 e de normas complementares, bem como está de acordo com os Princípios Éticos na Experimentação Animal adotado pelo Colégio Brasileiro de Experimentação Animal (COBEA), e foi aprovado pela COMISSÃO DE ÉTICA NO USO DE ANIMAIS DO INSTITUTO BUTANTAN (CEUAIB) em reunião de 9/10/2013.

Vigência do Projeto:	Nº de animais/espécie
10/2013 - 06/2014 Laboratório Especial de Toxinologia Aplicada	1 serpente <i>Bothrops jararaca</i> (M), 2 serpentes <i>Bothrops jararaca</i> (F)

São Paulo, 9 de outubro de 2013

Dr. Marcelo L. Santoro
Coordenador da CEUAIB

* Este protocolo foi aprovado, porém a liberação das serpentes está condicionada à entrega do número de SISBIO à CEUAIB, que deve ocorrer em um prazo de dois meses.

Aos meus pais, minha família, minha namorada e meus amigos.

Agradecimentos

Agradeço primeiramente aos meus pais, por todo o amor e apoio incondicional durante toda a minha vida, especialmente nesses anos fora de casa. À minha querida namorada Michelle pela paciência, pela dedicação, companherismo e pelo amor sincero.

À minha família por todo carinho transmitido até hoje, principalmente minha tia Neta, Juarez e meus primos que sempre me receberam tão bem na sua casa nos finais de semana em Mogi das Cruzes.

Ao meu orientador Dr. Inácio de L. M. Junqueira de Azevedo, pela oportunidade e por ter me integrado tão prontamente em seu laboratório.

À todos os profissionais e alunos que fazem o Laboratório Especial de Toxinologia Aplicada (LETA) do Instituto Butantan (IB).

À equipe do laboratório de proteômica do LETA: Dra. Solange Serrano, Débora A. Silva, Dra. Ana Helena P. Stunginski, Dra. Vanessa Rioli, Eduardo Shigueo, Carolina Brás, Paloma de Lima e Dra. Milene Cristina M. dos Santos, pela amizade e convivência diária.

Ao pessoal do laboratório de ciclo celular do LETA: Dra. Maria Carolina Sabbaga, Dra. Júlia Cunha, Dr. Matheus Henrique S. Dias, Raphael Pavani, Dr. Ricardo P. Llano, Dr. Marcelo Silva, Dr. Leo Iwai, Cecília Fonseca, Eduardo Lopes, Dr. Vincent Noël, Dr. Marcelo Reis, Dra. Simone Calderano, Loyze Lima, Daniel Ferreira, Mariana Lopes, Dra. Cristiane Araújo, Cristiane Castilho, Miryam G. Palomino, Teresa de Jesus e Fábio Montoni.

Aos técnicos Ismael Feitosa e Ivan Novaski, pela amizade e suporte aos serviços do laboratório e também ao amigo Bruno Duzzi. Aos outros técnicos: Regiane S. da Silva, Rosa Maria Carmo e Lidia A. da Silva; Às secretárias Maria José, Leda Maria G. Domingues e Leila Hashimoto Kusunoki e ao Isaías F. da Silva pelos serviços prestados ao laboratório.

Aos membros do laboratório de química de proteínas do LETA.

Um agradecimento especial aos membros do laboratório de Genômica e Transcriptômica:

À Dra. Ursula de Castro, Mariana S. L. C Morone e Dra. Nancy da Rós, pela amizade e por toda a ajuda na construção de bibliotecas de sequenciamento e outras etapas da bancada.

Ao Dr. Milton Nishiyama Júnior pela dedicação nas etapas de bioinformática do projeto.

À Dra. Pollyanna Fernandes Campos, pela parceria, amizade, conversas descontraídas e apoio durante todo o doutorado.

Ao Dr. Vincent Viala, que apesar de ter chegado recentemente já contribuiu bastante nas etapas de bancada do projeto e com as ilustrações da tese.

À Maria Andrea Camarano, a uruguaia (“Paraguaia”) mais doida que conheço (“calla-te!”), mas a mais legal também.

Ao professor Dr. Jesus Ferro e à Dra. Agda Facincani da UNESP-Jaboticabal pela doação do vetor para a construção da biblioteca de BACs.

À professora Dra. Claudia Vitorello da ESALQ-USP de Piracicaba, à professora Dra. Anete Pereira de Souza e Danilo Augusto do Laboratório de Análise Genética e Molecular (UNICAMP) pela ajuda técnica fundamental na construção da biblioteca de BACs.

Ao professor Dr. João Carlos Setubal e seu aluno George Condomitti, pela participação na montagem do genoma.

Às doutoras Mônica Lopes Ferreira e Carla Lima por disponibilizarem a sala de experimentos com animais para a dissecação das serpentes.

À todos os meus amigos de São Paulo e de Natal, que de alguma forma me apoiaram no doutorado.

Ao coordenador do CeTICS-FAPESP, Dr. Hugo Aguirre Armelin pelo apoio financeiro fornecido ao laboratório.

Ao CNPq e FAPESP (2013/07974-0) pelo suporte financeiro.

“O estudo em geral, a busca da verdade
e da beleza são domínios em que nos é
consentido ficar crianças toda a vida.”
Albert Einstein

RESUMO

ALMEIDA, D. D. **Sequenciamento do genoma da serpente *Bothrops jararaca* para caracterização da estrutura gênica de toxinas**. 2016. 126 f. Tese (Doutorado em Biotecnologia) – Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2016.

As serpentes do gênero *Bothrops* pertencem à superfamília Colubroidea, a qual engloba as diversas serpentes produtoras de veneno, incluindo a *Bothrops jararaca*, a serpente de maior importância médica no Brasil. Vários estudos foram realizados com o objetivo de caracterizar os componentes do veneno de serpentes e sua relação com a fisiopatologia do envenenamento. Tais estudos demonstraram que há uma enorme complexidade no repertório de toxinas ofídicas, embora ainda se conheça pouco sobre a origem dessa diversidade. Um dos fatores chave para o entendimento da composição do veneno e sua origem evolutiva é o sequenciamento do genoma de serpentes. Entretanto, a base molecular dos genomas das serpentes é pouquíssimo conhecida, mesmo considerando os avanços nas tecnologias de sequenciamento. Montagens de alguns rascunhos genômicos de serpentes peçonhentas foram disponibilizados recentemente nos bancos de dados, embora apenas um trabalho esteja disponível na literatura e com pouco aprofundamento acerca da estrutura genômica de toxinas. Além disso, o grupo “Squamata” (serpentes e lagartos) mantém-se particularmente pouco explorado quando comparado aos diversos grupos de vertebrados como aves, mamíferos e peixes que dispõem de vários genomas resolvidos.

Com o objetivo de obter um vislumbre da organização genômica dos genes de toxinas da serpente *Bothrops jararaca* foi realizado o sequenciamento com alta cobertura do genoma de um indivíduo dessa espécie. Para tanto, foram utilizadas várias estratégias: 1) a construção de bibliotecas tipo *shotgun* para realização de corridas de Sequenciamento de Próxima Geração (NGS) usando a tecnologia Illumina com leituras tipo *paired end* e tipo *mate pair* obtidas de bibliotecas construídas com fragmentos de diferentes tamanhos visando auxiliar a resolução de segmentos genômicos com longas repetições; 2) a construção de uma biblioteca de BACs da qual 768 *pools* de 12 BACs foram sequenciados e prospectados *in silico* para identificação de clones com genes de interesse; 3) obtenção de sequências complementares em equipamento PACBIO RS II da empresa Pacific Biosciences, a fim de se conseguir segmentos mais longos e contínuos

do genoma. Todas as sequências geradas foram montadas para produzir *contigs* e *scaffolds* do genoma. Foi possível obter um grande conjunto de segmentos genômicos, nos quais foram identificados genes de várias toxinas, entre elas Metaloproteinases, Serino Proteinases, BPPs, CRISPs e VEGF. Ainda foi possível depreender o contexto genômico de muitos destes genes e foram identificados os principais elementos repetitivos genômicos. Estes achados permitiram estabelecer novas hipóteses evolutivas para algumas classes de toxinas e reforçaram antigas hipóteses para outras. Os resultados nos levam a um melhor entendimento da função e evolução do sistema venenífero e podem servir de base para outros estudos futuramente. Além disso, o conjunto de dados de segmentos do genoma será disponibilizado publicamente, permitindo o acesso a qualquer um que esteja interessado em outros aspectos da biologia das serpentes.

Palavras chave: Genômica. Toxinas. Veneno. *Bothrops jararaca*.

ABSTRACT

ALMEIDA, D. D. **Genome sequencing of *Bothrops jararaca* snake for toxin gene structure characterization.** 2016. 126 p. Ph. D Thesis (Biotechnology) – Instituto de Ciências Biomédicas, Universidade de São Paulo, São Paulo, 2016.

Snakes of the *Bothrops* genus belong to the superfamily Colubroidea, which comprises the venom producing snakes, including the pit viper *Bothrops jararaca*, the most medically important snake in Brazil. Several studies were conducted in order to characterize the snake venom components and their relation to the pathophysiology of envenomation. Such studies have shown that there is a huge complexity in the repertoire of snake toxins, although we still know little about the origin of this diversity. One of the key factors for understanding the venom composition and its evolutionary origin is the availability of snake genome sequences. However, the molecular basis of snake genomes is yet poorly known, even considering the advances in sequencing technologies. Assemblies of some genomic drafts from venomous snakes have become available recently in the databases, even though there is only one publication in the literature without much information about the genomic structure of toxins. Furthermore, the "squamata" group (snakes and lizards) remains unexplored, particularly when compared to different groups of vertebrates such as birds, fish and mammals, which have several solved genomes.

In order to get a glimpse on the genomic organization of toxin genes from *Bothrops jararaca* snake, we conducted a high coverage genome sequencing of an individual belonging to this species. Therefore, several strategies were used: 1) the construction of shotgun libraries to perform Next Generation Sequencing (NGS) runs using Illumina technology with pair end and mate-pair reads, obtained from fragments of different sizes to aid in the resolution of genomic segments with long repeats; 2) the construction of a BAC library from which 768 pools of 12 BACs were sequenced and screened to identify clones containing genes of interest; 3) obtaining complementary sequences from a PACBIO RS II equipment in order to achieve continuous and longer segments of the genome. All generated sequences were assembled to yield contigs and scaffolds of the genome. It was possible to obtain a large collection of genome segments and multiple toxin genes were identified, including Metalloproteinases, Serine proteinases, BPPs, CRISPs and VEGF. In addition, it was possible to infer the genomic context related

to many of these genes and the main genomic repetitive elements were also identified. These findings allowed us to establish new evolutionary hypotheses for some toxin classes and to support old hypotheses for others. The results drive us to a better understanding of the venom system function and evolution, providing the basis for further studies. Besides, the set of the genome segments will be made publicly available, allowing access to anyone interested in other aspects of snake biology.

Keywords: Genomics. Toxins. Venom. *Bothrops jararaca*.

Lista de Ilustrações

Figura 1 -	Árvore filogenética, mostrando as relações entre diferentes táxons de serpentes.....	24
Figura 2 -	<i>Bothrops jararaca</i> (Fonte: wikipédia)	25
Figura 3 -	Cariótipo da serpente <i>B. jararaca</i>	30
Figura 4 -	Esquema mostrando o <i>workflow</i> do sequenciamento em plataforma Illumina. (Imagem: Stuart M. Brown).....	35
Figura 5 -	Estratégia experimental.....	38
Figura 6 -	DNA genômico purificado de <i>B. jararaca</i>	49
Figura 7 -	Distribuição de fragmentos da biblioteca <i>shotgun</i>	50
Figura 8 -	Distribuição de fragmentos a serem utilizados nas bibliotecas <i>mate-pair</i>	51
Figura 9 -	Distribuição de fragmentos das bibliotecas <i>mate-pair</i> quando finalizadas.	51
Figura 10 -	Distribuição de fragmentos da biblioteca <i>mate-pair</i> J6 (2kb-15kb) quando finalizada.	52
Figura 11 -	Gráfico gerado pelo programa fastQC, ilustrando a qualidade média das bases sequenciadas em uma das canaletas do lote 1.....	53
Figura 12 -	Gráfico gerado pelo fastQC, ilustrando a qualidade média das bases sequenciadas em uma das canaletas do lote 3, gerado no equipamento HiSeq1500.....	54
Figura 13 -	Gráfico gerado pelo fastQC, ilustrando a qualidade média das bases sequenciadas em uma das canaletas do lote 4, correspondente à biblioteca J2.....	55
Figura 14 -	Vetor pCUGIBAC1 digerido com a enzima HindIII. Foi feita uma corrida em PFGE, seguida de purificação do vetor pIndigoBAC536 do gel por eletroeluição.	63
Figura 15 -	Digestão parcial do DNA de <i>B. jararaca</i> com a enzima HindIII.....	64
Figura 16 -	Seleção de tamanho do DNA digerido com HindIII.	64

Figura 17 -	Determinação da concentração de DNA eletroeluído do gDNA de <i>B. jararaca</i>	65
Figura 18 -	Digestão analítica com a enzima Not I de 12 clones (canaletas 1 a 12) da biblioteca de BACs.	66
Figura 19 -	Esquema representando o BAC1 (biblioteca 4 da tabela 10) de 160 kb. ...	68
Figura 20 -	Esquema para identificação do clone de interesse no pool de BACs.	72
Figura 21 -	Exemplo de eletroforese de reações de PCR para <i>screening</i> de BACs contendo sequências de interesse.	73
Figura 22 -	Exemplo de confirmação da presença do <i>locus</i> genômico de interesse por PCR em colônias isoladas de um mesmo estoque de BAC.....	74
Figura 23 -	Perfil do sequenciamento de um lote de 23 células SMRT (de um total de 60) em sistema PacBio.....	75
Figura 24 -	Desenho esquemático mostrando as anotações na sequência de DNA mitocondrial de <i>B. jararaca</i>	77
Figura 25 -	Alinhamento entre o precursor rRNA45S de <i>B. jararaca</i> e o humano.....	78
Figura 26 -	Diagrama ilustrando o scaffold L2_002391, com tamanho total de 150 Kbp.....	81
Figura 27 -	Diagrama ilustrando um segmento do scaffold L8_000515, com tamanho total de 675 kb.	82
Figura 28 -	Contexto genômico de SVSPs e SVSPs-like em diferentes organismos.....	84
Figura 29 -	Diagrama ilustrando a posição dos íntrons de um gene de SVMP do tipo P-III encontrado em um dos BACs.	86
Figura 30 -	Diagrama ilustrando as sequências parciais do BAC 23G12 contendo dois genes de SVMPs.	88
Figura 31 -	Comparação entre os genes da ADAM28, Berythracivase-like (PIII) e Bothrostatina (PII) de <i>B. jararaca</i>	88
Figura 32 -	Mapeamento dos reads do transcriptoma das glândulas de veneno obtidos nesse trabalho em sequências de cDNA de referência de metalloproteínases do tipo P-I de <i>B. jararaca</i>	90
Figura 33 -	Esquema comparativo do <i>locus</i> genômico que contem o gene ADAM28 em diversos organismos.	92

Figura 34 -	A) Comparação entre o gene de uma SVMP-PIII e pseudogenes. B) Desenho esquemático das sequências de pseudogenes.....	93
Figura 35 -	Extremidade 5' terminal do BAC 23D05. Um gene de SVMP é interrompido no domínio metaloprotease por outro gene de SVMP.....	94
Figura 36 -	Diagrama ilustrando o gene do precursor dos BPPs.....	96
Figura 37 -	Alinhamento dos genes de CNPs e CNPs/BPPs.....	96
Figura 38 -	A) Diagrama ilustrando os genes do VEGF-A e VEGF-F de <i>B. jararaca</i> . B) Diagrama ilustrando os mesmos genes em <i>Protophthrops (Trimeresurus) flavoviridis</i> (Yamazaki et al., 2009).....	99
Figura 39 -	Esquema comparativo do <i>locus</i> genômico que contem o gene VEGF-F em diversos organismos..	100
Figura 40 -	Organização genômica do gene da Fosfolipase B de <i>B. jararaca</i>	101
Figura 41 -	Organização genômica do gene da Hialuronidase de <i>B. jararaca</i>	102
Figura 42 -	Organização e contexto genômico do gene da CRISP de <i>B. jararaca</i>	104

Lista de Tabelas

Tabela 1 -	Sequência dos pares de primers desenhados para amplificar por PCR segmentos dos genes de interesse a serem rastreados nos clones de BACs.....	47
Tabela 2 -	Extração de DNA genômico de <i>Bothrops jararaca</i>	49
Tabela 3 -	Concentração de cada biblioteca <i>mate-pair</i> para sequenciamento.....	52
Tabela 4 -	Dados gerais dos sequenciamentos na plataforma Illumina.	56
Tabela 5 -	Número de leituras em pares após processamento.....	57
Tabela 6 -	Estatísticas da melhor montagem (ALLPATHS-LGv2). Os valores de N50 de contigs e scaffolds estão destacados em vermelho.	58
Tabela 7 -	Composição dos conteúdos repetitivos do genoma da <i>B. jararaca</i> derivado das análises do RepeatMasker sobre o subconjunto resultante da montagem L1.....	59
Tabela 8 -	Sumarização da taxa de pares de leituras alinhadas na montagem do genoma.	61
Tabela 9 -	Sumarização do total de proteínas preditas pelo Augustus com similaridade em bancos específicos e com Uniprot.	62
Tabela 10 -	Estatísticas gerais de processamento das sequências de BACs.....	67
Tabela 11 -	Lista de BACs de interesse identificados por PCR, seus <i>pools</i> de origem e o gene de interesse.....	73
Tabela 12 -	Lista de estruturas gênicas das principais classes de potenciais componentes de venenos de serpentes do gênero <i>Bothrops</i> , recuperadas a partir do genoma de <i>B. jararaca</i> nas diferentes estratégias desse trabalho.....	79
Tabela 13 -	Comparação entre montagens de diferentes genomas de répteis.....	108

Lista de Abreviaturas e Siglas

ADAM	<i>A Disintegrin and Metalloproteinase</i>
BAC	Cromossomo Artificial de Bactéria
BPPs	Peptídeos Potenciadores de Bradicinina
CeTICs	Centro de Toxinas, Imuno-Resposta e Sinalização Celular
cDNA	DNA complementar
CRISP	Proteínas Secretadas Ricas em Cisteína
DNA	Ácido desoxirribonucléico
EDTA	Ácido Etilenodiamino tetra-acético
gDNA	DNA genômico
IPTG	Isopropil Tiogalactopiranosídeo
LB	Luria-Bertani
LINE	<i>Long Interspersed Elements</i>
mRNA	RNA mensageiro
NGS	Sequenciamento de Próxima Geração
nM	nanomolar
PBS	Tampão Fosfato Salino
PCR	Reação em cadeia da polimerase
PFGE	Eletroforese em Gel de Campo Pulsado
PMSF	Fenilmetilsulfonil fluoreto
qPCR	PCR quantitativo
RNA	Ácido ribonucleico
SINE	<i>Short Interspersed Elements</i>
SVMP	Metaloprotease de Veneno de Serpente
SVSP	Serinoprotease de Veneno de Serpente
TAE 1X	Tampão Tris 40 mM, Ácido acético 40 mM e 1 mM EDTA
TE	Tris 10 mM, EDTA 1 mM
UV	Ultra violeta
VEGF	Fator de Crescimento Endotelial Vascular

Sumário

1	INTRODUÇÃO.....	20
2	REVISÃO DE LITERATURA	23
2.1	RÉPTEIS E SERPENTES.....	23
2.2	PRINCIPAIS TOXINAS DE B. JARARACA	25
2.3	GENÔMICA DE RÉPTEIS E SERPENTES.....	29
2.4	ESTRUTURA GÊNICA DE TOXINAS DE SERPENTES	32
2.5	ESTRATÉGIAS EXPERIMENTAIS	34
3	OBJETIVOS	39
4	METODOLOGIA	40
4.1	ESPÉCIMES	40
4.2	OBTENÇÃO DE DNA	40
4.3	OBTENÇÃO DE RNA] E SEQUENCIAMENTO EM PLATAFORMA ILLUMINA	40
4.4	PREPARO E SEQUENCIAMENTO DE DNA EM SISTEMA ILLUMINA	41
4.5	ANÁLISE BIOINFORMÁTICA DAS SEQUÊNCIAS GENÔMICAS.....	41
4.6	PREPARAÇÃO DE PLUGS DE AGAROSE PARA CLONAGEM DE DNA.....	42
4.7	CONSTRUÇÃO DE BIBLIOTECAS GENÔMICAS EM BAC	43
4.8	SEQUENCIAMENTO DOS BACS	45
4.9	DESENHO DE PARES DE PRIMERS ESPECÍFICOS PARA SELEÇÃO DE BACS	46
4.10	IDENTIFICAÇÃO E VALIDAÇÃO DOS POOLS DE BACS DE INTERESSE.....	47
4.11	RESSEQUENCIAMENTO DE BACS EM SISTEMA ILLUMINA.....	47
4.12	SEQUENCIAMENTO EM SISTEMA PACBIO.....	48
5	RESULTADOS	49
5.1	EXTRAÇÃO DE DNA GENÔMICO PARA CONSTRUÇÃO DE BIBLIOTECAS DE SEQUENCIAMENTO	49
5.2	PREPARO DE BIBLIOTECAS GENÔMICAS PARA O SISTEMA ILLUMINA.....	50
5.3	SEQUENCIAMENTO GENÔMICO EM SISTEMA ILLUMINA.....	53
5.4	MONTAGEM PRELIMINAR DAS SEQUÊNCIAS	57
5.4.1	ANÁLISE DE ELEMENTOS REPETITIVOS NA VERSÃO 1.0 DO GENOMA	58
5.4.2	AVALIAÇÃO DA INTEGRIDADE DOS GENES DO GENOMA MONTADO	60
5.4.3	MAPEAMENTO DAS LEITURAS TRANSCRIPTÔMICAS NA VERSÃO 1.0 DO GENOMA	60
5.4.4	PREDIÇÃO DE GENES E ANOTAÇÃO DO GENOMA	61
5.5	PREPARO DE BIBLIOTECAS EM BAC	62
5.6	TESTE PILOTO DE SEQUENCIAMENTO ALEATÓRIO DE BACS EM POOLS.....	66
5.7	ESTRATÉGIA DE IDENTIFICAÇÃO IN SILICO DE GENES DE INTERESSE CLONADOS EM BACS	69
5.7.1	SEQUENCIAMENTO DOS POOLS DE BACS E SCREENING DE GENES.....	69
5.7.2	SELEÇÃO E RESSEQUENCIAMENTO DE BACS ESPECÍFICOS	71
5.8	SEQUENCIAMENTO EM SISTEMA PACBIO	75

5.9 MONTAGEM DO GENOMA MITOCONDRIAL	76
5.10 RNA RIBOSSOMAL	77
5.11 BUSCA E ANÁLISE DOS GENES DE TOXINAS	78
5.11.1. GENES DE TOXINAS RECUPERADOS	78
5.11.2 ORGANIZAÇÃO GENÔMICA DE SERINO PROTEINASES DE VENENO	80
5.11.3 CONTEXTO GENÔMICO DAS SVSPS	83
5.11.4 A ESTRUTURA GÊNICA DE UMA SVMP DO TIPO P-III.....	85
5.11.5 A ESTRUTURA GÊNICA DE UMA SVMP DO TIPO P-II.....	86
5.11.6 AUSÊNCIA APARENTE DE METALOPROTEINASES DO TIPO P-I.....	89
5.11.7 O CONTEXTO GENÔMICO DAS METALOPROTEINASES	90
5.11.8 POSSÍVEL PSEUDOGENE E GENES TRUNCADOS	92
5.11.9 MONTAGEM DO GENE DO PRECURSOR DOS BPPS/CNP A PARTIR DE READS PACBIO.....	94
5.11.10 GENE DO VEGF-F (SVVEGF) E UMA NOVA HIPÓTESE DE SUA ORIGEM	97
5.11.11 GENE DE FOSFOLIPASE B	101
5.11.12 GENE DE HIALURONIDASE	102
5.11.13 GENE DE CRISP	103
5.11.14 OUTROS GENES DE TOXINAS E COMPONENTES DO VENENO.....	104
6 DISCUSSÃO	106
7 CONCLUSÃO	116
REFERÊNCIAS*	119

1 INTRODUÇÃO

Há milênios o ser humano é fascinado pelos efeitos devastadores de venenos e peçonhas de animais, plantas e microrganismos, especialmente em relação às serpentes. No entanto, somente no início do século passado o homem começou a estudar e entender melhor os mecanismos pelos quais as peçonhas e venenos atuam, bem como a sua natureza. Daí em diante surgiu um ramo da ciência que estuda as toxinas, suas origens e mecanismos de ação. Essa área, conhecida atualmente como Toxinologia, evoluiu enormemente nos últimos anos, tendo passado por diferentes fases que acompanharam o desenvolvimento tecnológico. O conhecimento acumulado sobre a ação dos venenos permitiu o desenvolvimento dos soros antiveneno, já o isolamento de toxinas foi preponderante para o desenvolvimento de fármacos e o esclarecimento de vários mecanismos fisiológicos que ocorrem em mamíferos.

No começo do novo milênio até os dias atuais houve um interesse crescente na composição global dos venenos de animais (principalmente serpentes), que passou a ser investigada por meio da transcriptômica e proteômica (BRAHMA et al., 2015). Naturalmente, perguntas a respeito da origem evolutiva das toxinas e outros componentes de veneno foram surgindo. As informações geradas nos projetos “ômicos” serviram como base, por meio de análises filogenéticas de transcritos específicos, para um vislumbre de como os venenos se originaram.

Muitas hipóteses acerca da evolução das glândulas de veneno foram levantadas e ainda são alvo de intenso debate (CASEWELL; HUTTLEY; WUSTER, 2012; FRY et al., 2006; HARGREAVES et al., 2014; REYES-VELASCO et al., 2015). Um dos aspectos quase consensual é que muitas das toxinas são originadas por duplicações de genes com papéis fisiológicos em outros tecidos. No que concerne ao desenvolvimento das glândulas de veneno, a hipótese mais conhecida trata do aparecimento comum do veneno em um táxon que inclui determinadas espécies de lagartos e as serpentes (FRY et al., 2006), tal clado ficou conhecido como Toxicofera. O autor se baseou na ocorrência de transcritos homólogos a toxinas em glândulas de veneno ou salivares de diferentes espécies. Dessa forma, as espécies de lagartos e serpentes produtoras de veneno, mesmo possuindo glândulas com origens diferentes (os primeiros possuem glândulas de veneno mandibulares enquanto as segundas glândulas maxilares), teriam um ancestral comum do clado Toxicofera com genes semelhantes a toxinas expressos nos dois tipos de

glândulas, e enquanto as serpentes avançadas favoreceram as glândulas maxilares, reduzindo as mandibulares, algumas espécies de lagartos fizeram o oposto. Em ambos os grupos haveria perda secundária do veneno em algumas famílias (FRY et al., 2009). Porém, outros autores sugerem que as glândulas de veneno surgiram diversas vezes na história evolutiva dos répteis, contrariando veementemente a hipótese “Toxicofera” (HARGREAVES et al., 2014). Essa proposição é baseada no fato de que transcritos relacionados a algumas classes de toxinas podem ser expressos em baixos níveis em vários tipos de tecidos de serpentes peçonhentas e de lagartos, não sendo, portanto, necessariamente toxinas. Entretanto, outros autores mostram que transcritos reconhecidamente pertencentes a toxinas ofídicas foram encontrados em diferentes órgãos de serpentes peçonhentas (CASEWELL; HUTTLEY; WUSTER, 2012; JUNQUEIRA-DE-AZEVEDO et al., 2015) e mesmo em serpentes não peçonhentas (REYES-VELASCO et al., 2015). Isso levantou ainda mais questões, levando à hipótese de que genes de toxinas podem ser recrutados reversamente para os tecidos (CASEWELL; HUTTLEY; WUSTER, 2012).

Em meio a esse debate, há um elemento muito importante faltando: as sequências genômicas. Estas são peças fundamentais para uma compreensão mais completa sobre as diversas classes de toxinas existentes. Várias questões à respeito da estrutura gênica, modificações pós-transcricionais, número de parálogos, regulação da expressão, dentre outras, ainda estão abertas. Por exemplo, a grande quantidade de isoformas de toxinas encontradas no veneno são fruto de modificações pós-transcricionais? Ou são de fato diversos genes parálogos? O sequenciamento dos genomas ofídicos poderá esclarecer vários desses pontos.

Há pouco mais de 10 anos, a tarefa de sequenciar um genoma complexo exigia um esforço monumental de consórcios com várias equipes em diferentes centros de pesquisa. O tempo e os recursos necessários eram tremendos. Sendo assim, apenas poucos organismos modelos tiveram seus genomas sequenciados com a tecnologia vigente (baseada no método de Sanger). No entanto, com o advento do sequenciamento de próxima geração (NGS), os custos e dificuldades para tais projetos diminuíram consideravelmente, o que possibilitou o sequenciamento dos genomas de diversos organismos dos quais se sabia muito pouco. Dessa forma, muitos projetos conseguiram responder questões relevantes sobre a evolução de características peculiares de diferentes animais. Como exemplo temos: o genoma da ostra *Crassostrea gigas* mostrou

que a expansão de genes HSP70 e inibidores de apoptose foi importante para a adaptação à vida sésil e às condições estressantes das zonas entremarés (ZHANG et al., 2012); o genoma do crocodilo *Alligator sinensis* revelou que o gene de hemoglobina ligadora de bicarbonato foi duplicado e outros genes sofreram modificações regulatórias que permitiram a esses animais permanecerem longos períodos sem respirar embaixo da água. Outras modificações foram observadas em genes do sistema sensorial e imunológico (WAN et al., 2013); Já em *Gekko japonicus*, o sequenciamento genômico revelou que o surgimento das cerdas reponsáveis pela adesividade dos dedos das “lagartixas” é resultado da duplicação e diversificação de genes da β -queratina (LIU et al., 2015).

Da mesma forma, o sequenciamento genômico de serpentes peçonhentas e não peçonhentas pode esclarecer questões fundamentais sobre o desenvolvimento das glândulas de veneno e das toxinas. Até o momento, poucos genomas ofídicos foram sequenciados e muitos dos que estão atualmente disponíveis ainda apresentam-se bastante fragmentados, provavelmente resultado de repetições que dificultam a montagem dos mesmos. Nos últimos anos, entretanto, parece haver um movimento no sentido de sequenciamento dos genomas de serpentes, com esforços para melhores resultados serem atingidos.

Considerando as questões levantadas anteriormente, nós realizamos o sequenciamento com alta cobertura do genoma da serpente peçonhenta *Bothrops jararaca*, utilizando diferentes estratégias. Apesar das dificuldades inerentes a esse tipo de projeto, foi possível identificar vários genes de toxinas e componentes de veneno. Ainda estabelecemos o contexto genômico para algumas classes de toxinas o que permitiu inferências interessantes do ponto de vista evolutivo.

2 REVISÃO DE LITERATURA

2.1 RÉPTEIS E SERPENTES

Os répteis evoluíram a partir dos anfíbios labirintodontes aproximadamente 60 milhões de anos depois que os primeiros anfíbios surgiram. Desde o Permiano e ao longo do Cretáceo os répteis foram os vertebrados mais abundantes. Foram ainda os primeiros vertebrados que apresentaram requisitos para a vida plenamente terrestre, incluindo o aparecimento do ovo terrestre, as membranas embrionárias e um tegumento resistente à dessecação (HILDEBRAND, 2006). Os répteis são representados pelos clados Archosauria (crocodilianos e aves), Testudines (tartarugas) e Lepidosauria (tuataras e squamata) (VITT; CALDWELL, 2014).

As serpentes estão classificadas dentro da ordem Squamata (que também contém os lagartos e anfisbenídeos), representando o segundo grupo mais diverso dentre os répteis vivos, com cerca de 3400 espécies (<http://www.reptile-database.org/>). Assim como os lagartos, elas ocorrem em todos os continentes exceto Antártida. Todos os estudos recentes indicam que as serpentes divergiram primeiramente em Scolecophidia e Alethinophidia (o qual é subdividido nos clados Afrophidia e Amerophidia). No entanto, a classificação filogenética das serpentes ainda é objeto de discussão, especialmente com relação à definição de superfamílias/famílias/subfamílias das espécies mais derivadas. Tomando uma posição mais conservadora considera-se que a superfamília Colubroidea, a qual contém as espécies de serpentes peçonhentas conhecidas, é formada pelas famílias Xenodermatidae, Preatidae, Viperidae, Homalopsidae, Lamprophiidae, Elapidae e Colubridae (Figura 1)(VITT; CALDWELL, 2014). As serpentes avançadas (Caenophidia) são representadas pelo clado que contém a superfamília Colubroidea e a família Acrochordidae.

Figura 1 - Árvore filogenética, mostrando as relações entre diferentes táxons de serpentes. Abreviações: J: Jurássico, K: Cretáceo, Ng: Neogene. (Vidal et al., 2009; Vitt; Caldwell, 2014) .

A família Viperidae possui cerca de 330 espécies, dentre as quais estão as serpentes do gênero *Bothrops* (44 espécies) (<http://www.reptile-database.org/>). No Brasil existem cerca de 30 espécies do gênero *Bothrops*, sendo consideradas de importância médica. Entre os anos de 1990 e 1993, os acidentes ofídicos por *Bothrops* respondiam por mais de 90% dos casos. *Bothrops jararaca* é a espécie mais importante do ponto de vista médico, estando distribuída da região central da Bahia para o sul até o Rio Grande do Sul (FUNASA, 2001). Esta possui um padrão de coloração pouco variável, com manchas marrons escuras dispostas em faixas entremeadas por áreas claras, cinzas, oliváceas, amareladas ou beges (Figura 2) (CAMPBELL; LAMAR, 1989).

Figura 2 - *Bothrops jararaca* (Fonte: wikipédia)

Os sintomas mais frequentes do envenenamento por *B. jararaca* são distúrbios hemostáticos graves, com coagulopatia, hemorragia local ou sistêmica e danos locais ao tecido com consequências variáveis de acordo com a gravidade do acidente (KAMIGUTI et al., 1991; SANTORO; SANO-MARTINS, 2004). Estes sintomas resultam das três principais atividades do veneno botrópico: 1) proteolítica, com edema inflamatório no local da picada; 2) hemorrágica, com danos ao endotélio e hemorragia sistêmica; e 3) pró-coagulante, que conduz ao consumo de fatores da coagulação, agravando a hemorragia (MATSUI; FUJIMURA; TITANI, 2000).

2.2 PRINCIPAIS TOXINAS DE *B. JARARACA*

O veneno de *B. jararaca* é uma mistura complexa de várias classes de toxinas, enzimáticas e não-enzimáticas, sendo que a composição do veneno pode variar bastante intraespecificamente e de acordo com a dieta, idade e sexo, inclusive para indivíduos da mesma ninhada (AUGUSTO-DE-OLIVEIRA et al., 2016; DIAS et al., 2013; MENEZES et al., 2006; ZELANIS et al., 2011; ZELANIS et al., 2010). O veneno atua na imobilização e possivelmente na digestão da presa e na defesa contra predadores. As principais classes de toxinas encontradas em *B. jararaca* são:

Metaloproteínases (classe de toxinas mais abundante): As metaloproteases ou metaloproteinases (SVMPs) são os mais importantes componentes dos venenos botrópicos e são particularmente interessantes do ponto de vista evolutivo. São classificadas em três classes de acordo com a estrutura primária da proteína madura: PI possuem apenas o domínio metaloprotease; PII são compostas por domínio

metaloprotease e desintegrina; PIII são sintetizados com domínio metaloprotease, domínio tipo desintegrina e domínio rico em cisteína. Além disso, todas possuem um peptídeo sinal e um prodomínio. Cada classe pode ainda ser subdividida em subtipos definidos pela organização desses domínios após o processamento (BJARNASON; FOX, 1995). As SVMPS são mais filogeneticamente relacionadas com as enzimas transmembrânicas ADAMs de mamíferos (metaloprotease com domínio desintegrina) e ADAMTS (ADAM com thrombospondina motivo tipo-1) devido à similaridade de sequências do domínio metaloprotease (BODE; GOMIS-RUTH; STOCKLER, 1993; FOX; SERRANO, 2005; HUXLEY-JONES et al., 2007), sendo o gene da ADAM28 o mais próximo do ponto de vista filogenético (CASEWELL, 2012). A ADAM28 é constituída por oito domínios: peptídeo sinal, pró-domínio, metaloprotease, tipo desintegrina, rico em cisteínas, *epidermal-growth-factor* (EGF-like) C-terminal, transmembrana e cauda citoplasmática (FOX; SERRANO, 2005). Alguns desses domínios são ausentes nas SVMPS, características que conferem a essas proteínas uma alta diversidade funcional. Em sua maioria apresentam atividade hemorrágica através de degradação de proteínas da membrana basal na parede dos vasos sanguíneos, hidrólise de fatores da coagulação, inibição da agregação plaquetária e apoptose de células endoteliais (BERNARDONI et al., 2014).

Serino-proteinases de veneno (SVSPs): Enzimas estruturalmente semelhantes a calicreínas/quimotripsinas. Atuam em uma variedade de componentes da cascata de coagulação, clivando ligações específicas e ativando as proteínas envolvidas na fibrinólise, agregação plaquetária e sistema calicreína-cinina causando um desequilíbrio do sistema hemostático da presa (SERRANO; MAROUN, 2005). O mecanismo catalítico de SVSPs é definido por um resíduo de serina reativo (equivalente à serina 195 de quimotripsinogênio) que desempenha um papel fundamental na formação de um complexo transitório acil-enzima, o qual é estabilizado pelos resíduos His57 e Asp102 dentro do sítio ativo (BARRETT; RAWLINGS, 1995). Acredita-se que os genes que codificam essas enzimas sofreram um processo de evolução acelerada, o que resultou em várias isozimas capazes de interagir especificamente com substratos para causar um desequilíbrio da hemostasia (DESHIMARU et al., 1996).

Lectinas do tipo C: Existem duas classes principais de lectinas do tipo C em venenos de serpentes: as lectinas "verdadeiras", geralmente homodiméricas e apresentando atividades de ligação a sacarídeos de forma cálcio-dependente (XU et al.,

1999); e as proteínas semelhantes a lectinas do tipo C, geralmente heterodimérica e que se ligam a proteínas em vez de sacarídeos. A principal atividade relatada da primeira classe é hemaglutinação (DE CARVALHO; MARANGONI; NOVELLO, 2002; XU et al., 1999) enquanto os membros da segunda classe podem interferir nas vias de coagulação de várias maneiras, exibindo propriedades anticoagulantes ou efeitos moduladores de plaquetas (MORITA, 2005).

Peptídeos Potenciadores de Bradicinina: Os BPPs são oligopeptídeos de 5-14 resíduos de aminoácidos que geralmente apresentam um resíduo piroglutamil no N terminal e uma sequência Ile-Pro-Pro (IPP) no C-terminal. Estes peptídeos inibem a enzima conversora de angiotensina somática (sACE), a qual gera angiotensina II, um potente vasoconstritor, a partir do seu precursor (Angiotensina I). A sACE também é capaz de inativar o peptídeo hipotensor bradicinina, que é liberado no sangue de vítimas envenenadas por meio da ação das SVSPs. Portanto, os BPPs estão envolvidos com a hipotensão observada após o envenenamento por espécies do gênero *Bothrops*, e foram utilizados como modelos para o desenvolvimento de novas drogas anti-hipertensivas (CAMARGO et al., 2012). Alguns destes oligopeptídeos também induzem a produção de óxido nítrico no endotélio por diferentes mecanismos (GUERREIRO et al., 2009; IANZER et al., 2007), o que contribui para o efeito hipotensor. Os BPPs são codificados por um precursor que contém, além destes, um peptídeo sinal, uma região espaçadora e outra região codificando um Peptídeo Natriurético do tipo C (CNP) (MURAYAMA et al., 1997), o qual por si só pode ter algum efeito hipotensor, atuando em receptores natriuréticos (MADHANI et al., 2003).

Fosfolipases A2 (PLA2s): são enzimas amplamente presentes em venenos de serpentes, que desempenham papéis tóxicos importantes na imobilização e/ou abate das presas (KINI, 2003). Entre suas diversas atividades, a miotoxicidade é um efeito clinicamente relevante que pode levar a danos teciduais graves e sequelas associadas a envenenamentos. Dois genes ancestrais diferentes representam dois grupos de PLA2s, grupo I e grupo II, os quais foram recrutados e expressos de forma independente nas secreções da glândula de veneno de Elapídeos e Viperídeos, respectivamente (LOMONTE; GUTIERREZ, 2011). Através de um processo de evolução acelerada (NAKASHIMA et al., 1995), estes genes sofreram mutações e foram convertidos em toxinas, apresentando atividades neurotóxicas e/ou miotóxicas, o que parece representar um processo de evolução convergente.

Nos venenos de viperídeos, dois subtipos de PLA2s miotóxicas podem ser encontradas, as que possuem o aspartato na posição 49, o qual é um resíduo catalítico importante, e as que possuem a lisina na posição 49, sendo conhecidas respectivamente como variantes “Asp49” e “Lys49”. Este último, além da substituição do aspartato por lisina na posição 49, apresenta outras substituições de aminoácidos chave localizados no *loop* ligador de cálcio. Dessa forma, as PLA2s “Lys49” são enzimaticamente inativas, sendo portanto proteínas PLA2-like (DOS SANTOS et al., 2009; MARAGANORE et al., 1984). Apesar de não possuírem atividade enzimática, as PLA2s “Lys49” também apresentam atividade miotóxica *in vivo* (GUTIERREZ; LOMONTE, 1995; MONTECUCCO; GUTIERREZ; LOMONTE, 2008), induzindo mionecrose dependente de um conjunto de aminoácidos na região C-terminal que afeta diretamente a integridade do sarcolema (CINTRA-FRANCISCHINELLI et al., 2010) mecanismo este diferente das PLA2s “Asp49”.

Além destas, existem outros componentes de menor expressão no veneno, como: **proteínas ricas em cisteína (CRISPs)**, que são altamente distribuídas em venenos de serpentes (CHING et al., 2012; CIDADE et al., 2006), no entanto, conhece-se pouco à respeito de sua atividade. Atualmente sabe-se que elas podem inibir a contração da musculatura lisa e bloquear canais iônicos regulados por nucleotídeos cíclicos (OSIPOV et al., 2005; YAMAZAKI; MORITA, 2004). Recentemente foi demonstrado que uma CRISP isolada do veneno de *Echis carinatus sochureki* pode ter ação anti-angiogênica por interação direta com células endoteliais (LECHT et al., 2015); as **L-aminoácido oxidases (LAAOs)** de venenos de serpentes são citotóxicas, apoptóticas, inibem a atividade de agregação plaquetária e exibem atividade bactericida e anti-viral (ALVES et al., 2008; RODRIGUES et al., 2009); As **Hialuronidases** encontradas em venenos de serpentes podem potencializar a toxicidade do veneno e contribuir para danos no local da picada, pois afetam a integridade da matriz extracelular devido à degradação do ácido hialurônico (GIRISH et al., 2002); já o **Fator de Crescimento Endotelial Vascular (svVEGF ou VEGF-F)** é capaz de aumentar a permeabilidade vascular, atuando como um agente de dispersão de toxinas e provavelmente contribui para a ação hipotensora observada após o envenenamento (JUNQUEIRA DE AZEVEDO et al., 2001). Vários outros componentes protéicos minoritários são descritos em venenos viperídicos, mas sua real contribuição para o envenenamento é pouco compreendida. Inclui-se entre eles o nerve growth fator (NGF), 5'nucleotidase, cistatinas, fosfolipase B, fosfodiesterases, defensinas, dentre outros.

2.3 GENÔMICA DE RÉPTEIS E SERPENTES

A ciência genômica adquiriu na última década uma dimensão enorme e se estabeleceu como ferramenta preferida para análise genotípica de diversos organismos com as mais variadas finalidades. Na agricultura e pecuária, como ferramenta de conhecimento do background gênico das variantes mais produtivas (ELSIK et al., 2009; YAMAMOTO et al., 2010). Na genética médica, como método investigativo de polimorfismos associados à clínica, ao nível individual (REN; REN, 2016; ZHANG; LIU, 2016). Na microbiologia, como técnica de metaidentificação de espécies dentro de uma população complexa em quaisquer ambientes (RIESENFELD; SCHLOSS; HANDELSMAN, 2004). Entretanto, o uso da genômica para tais finalidades aplicadas só foi possível após a aquisição de um conhecimento básico sobre o genoma modelo das espécies estudadas.

O exponencial crescimento das variantes de técnicas genômicas (exoma, ChIPSeq, target sequencing, etc.) e das respectivas ferramentas de análise toma como premissa a existência de genomas referenciais e genes previamente anotados. Na verdade, nota-se um claro movimento de desenvolvimento da ciência genômica em torno de modelos preexistentes. Nesse contexto, o sequenciamento e descrição de novos genomas não são mais o foco dessa ciência; por outro lado constituem o pilar necessário para essas análises.

É surpreendente, portanto, saber que a sequência nucleotídica e a estrutura dos genomas de serpentes são pouquíssimo conhecidas. Atualmente, sabemos que os genomas das serpentes têm um grau de complexidade semelhante ao do humano. No gênero *Bothrops* é comum haver 36 cromossomos, sendo 10 pares de microcromossomos e 8 pares de macrocromossomos, incluindo os pares sexuais ZZ (macho) ou ZW (fêmea) (Figura 3). O valor-C estimado para o gênero varia entre 1,8 pg e 2,7 pg, sendo predito para *B. jararaca* como 2,25 pg, antevendo um genoma de 2,2 bilhões de pares de bases (ATKIN et al., 1965; GREGORY, 2016).

Figura 3 - Cariótipo da serpente *B. jararaca*.
(Becak; Becak; Nazareth, 1962)

Além disso, enquanto dezenas de gêneros de mamíferos, aves, anfíbios e peixes tiveram genomas completamente resolvidos, o grupo dos répteis *Squamata* (cobras e lagartos) mantém-se particularmente pouco explorado. Até pouco tempo atrás, o genoma do lagarto *Anolis carolinensis* (ALFOLDI et al., 2011) constituía o único exemplo de genoma disponível. Recentemente o genoma do *Gekko japonicus* (LIU et al., 2015) foi disponibilizado, o qual pertence à mesma família das “lagartixas” (Gekkonidae). Porém o grau de proximidade entre lagartos e as serpentes é algo semelhante entre o homem e o lêmure (mesma ordem, diferentes sub-ordens).

Ao final de 2013, foram publicados dois importantes trabalhos na mesma edição da revista PNAS descrevendo os *drafts* dos genomas das serpentes píton (*Python molurus bivittatus*) (CASTOE et al., 2013) e da cobra rei (*Ophiophagus hannah*) (VONK et al., 2013). Este último é particularmente relevante para a área toxinológica por se tratar de uma serpente peçonhenta. Assim, o trabalho aborda alguns aspectos da estrutura gênica das toxinas presentes nessa espécie e muitos aspectos evolutivos. Porém, a abordagem metodológica não permitiu a geração de grandes *scaffolds* (ordenação de segmentos montados) de genoma nem a localização em cromossomos. Com isso, pouco pôde ser observado sobre a estrutura e localização dos genes de toxinas. Na verdade, o foco deste estudo conduzido por um grupo de embriologistas, bem como no caso de *Python morulus* descrito acima, teve como mote o entendimento de importantes aspectos da biologia das serpentes, sendo a análise de toxinas um aspecto secundário.

Apesar dessa recente contribuição, *Ophiophagus hannah* pertence à família *Elapidae*, grupo mais comum no Velho Mundo, mas pouco diversificado nas Américas. No Brasil, a família *Viperidae* é a mais representativa dentre as serpentes peçonhentas, especialmente os gêneros *Bothrops* (jararacas), *Crotalus* (cascavéis) e *Lachesis*

(surucucu). Assim, as bases genéticas das serpentes brasileiras continuam pouco conhecidas.

Mais recentemente (2015 e 2016) foram disponibilizados no NCBI outros dois *drafts* de genomas de serpentes: o do colubrídeo *Thamnophis sirtalis*, submetido por “The Genome Institute at Washington University School of Medicine (WUGSC)” e o do viperídeo *Protobothrops mucrosquamatus*, submetido pela “Okinawa Institute of Science and Technology”. Ambos não possuem publicações associadas diretamente com a montagem genômica e nem com evolução de toxinas. Apesar disso, os scaffolds presentes nessas montagens são relevantes para estudos comparativos e alguns deles foram utilizados nesse trabalho.

A estrutura corpórea única das serpentes faz com que seu estudo genômico seja bastante relevante para o entendimento de aspectos críticos da biologia do desenvolvimento. São animais que sofreram perda de membros e alguns órgãos, ao passo que o corpo se alongou com múltiplas vértebras e órgãos assimétricos. Assim, constituem um modelo importante para o estudo dos genes de desenvolvimento, como o *hox* (BEJDER; HALL, 2002).

Do ponto de vista fisiológico, as serpentes apresentam regulação ectotérmica, reprodução variada (ovípara, ovovivípara, vivípara), troca de pele, etc. Além disso, possuem órgãos sensoriais especializados, como órgão de Jacobson (ou vomeronasal) que permite a quimiorrecepção olfativa de partículas conduzidas pela língua e fossetas loreais, termorreceptores de precisão que lhes permitem localizar a presa por sinais infravermelhos. A glândula de veneno é também um tecido interessante, visto que é um epitélio extremamente especializado. Possui um lúmen central em que o veneno é estocado, o que não ocorre em outras glândulas exócrinas (GOPALAKRISHNAKONE; KOCHVA, 1993). Após uma picada, as células do epitélio alteram sua forma cubóide para colunar, ocorre expansão das cisternas do retículo endoplasmático rugoso e o veneno é então sintetizado. A atividade máxima de síntese das células secretoras e a maior concentração de mRNAs são observados de 4 a 8 dias após a extração manual do veneno ou picada. Em seguida, o veneno é acumulado no lúmen da glândula, enquanto o epitélio retorna ao estado quiescente. Este ciclo está sujeito ao controle do sistema nervoso simpático, sendo este tipo peculiar de glândula exócrina um modelo interessante para a investigação dos mecanismos celulares de secreção proteica (YAMANOUYE et al., 1997).

2.4 ESTRUTURA GÊNICA DE TOXINAS DE SERPENTES

Do ponto de vista da ciência das toxinas, a ausência de dados genômicos deixa uma série de questões abertas e mal compreendidas quanto à organização estrutural de precursores de toxinas. Muitas dessas toxinas são compostas por múltiplos domínios cujos recrutamentos e substituições parecem ocorrer dinamicamente (CASEWELL et al., 2011; JUNQUEIRA-DE-AZEVEDO et al., 2006).

Atualmente a estrutura de alguns poucos genes de toxinas de serpentes são conhecidos. Itoh et al., (1988) revelaram a organização genômica do gene da batroxobina, uma serinoprotease do veneno de *Bothrops atrox*. O mesmo possui uma extensão de 8 kb, possuindo 5 éxons, sendo os resíduos catalíticos His-41, Asp-86 e Ser-178 codificados nos éxons 2, 3 e 5, respectivamente. A organização estrutural desse gene é semelhante à dos genes da calicreína e tripsina.

Outro gene codificante de toxinas conhecido é o svVEGF de *Trimeresurus flavoviridis* o qual possui 3.137 pb, contendo 6 éxons e 5 íntrons, possuindo um baixo grau de identidade com o gene do VEGF tissular (VEGF-A), o qual apresenta íntrons maiores e 8 éxons em vez de 6 (YAMAZAKI et al., 2009).

Há ainda evidências de genes de toxinas ocorrendo em clusters, como as Fosfolipases A2 de *Protobothrops flavoviridis*, com quatro éxons e três íntrons. Um clone de cosmídeo de 25.026 pb foi isolado e sequenciado, este continha 5 sequências PLA2-like (duas delas pseudogenes) *in tandem*. Quatro destas eram seguidas por uma sequência de transcriptase reversa incompleta cuja forma intacta se assemelha a um LINE do tipo CR1. Supõe-se que esse cluster foi gerado por duplicações consecutivas de uma unidade PLA-CR1 por *crossing over* desigual (IKEDA et al., 2010).

Com relação às SVMPs, a classe de toxina mais abundante no veneno de *B. jararaca*, até pouco tempo só se conhecia a organização genômica de um gene. Trata-se de uma metaloprotease do tipo PIII de *Echis ocellatus* (um viperídeo da subfamília viperinae) contendo 12 éxons e 11 íntrons. Retrotransposons do tipo LINE L2/CR1 e RTE/Bov-B, SINEs do tipo SINE/Sauria e um transposon do tipo Charlie-hAT foram identificados dentro dos íntrons 1, 3, 7 e 8 (SANZ; HARRISON; CALVETE, 2012). Este ano outro trabalho do mesmo grupo foi publicado mostrando a estrutura gênica de outras duas metaloproteases do tipo PII e PI de *Echis ocellatus* (SANZ; CALVETE, 2016). Todos estes trabalhos utilizaram amplificação por PCR para a montagem dos genes em questão.

Os transcritos correspondentes a estas metaloproteases apresentam certas diferenças estruturais em relação aos correspondentes em outros viperídeos da subfamília crotalinae (à qual a *B. jararaca* pertence).

Além das características peculiares dos precursores, os venenos de serpentes são marcados pela alta variabilidade estrutural e funcional das famílias de proteínas que compõem seu arsenal de toxinas. Os altos níveis de variabilidade levam a diversificação dos seus efeitos farmacológicos, gerando um leque abrangente de moléculas capazes de interagir com receptores de diferentes espécies, o que permite uma adaptação aos predadores e a uma maior variedade de presas.

As glândulas produtoras de venenos de serpentes têm se mostrado como um laboratório de fenômenos de evolução. Os fenômenos de duplicação gênica e evolução acelerada, inicialmente descritos por Ohno (1970), foram propostos em várias famílias de proteínas presentes em diferentes tipos de venenos (JUAREZ et al., 2008; LYNCH, 2007; MOURA-DA-SILVA; THEAKSTON; CRAMPTON, 1996; OGAWA et al., 2005). Por exemplo, as análises filogenéticas das metaloproteases de veneno (SVMPs) evidenciam a duplicação gênica e acúmulo de mutações substitutivas (MOURA-DA-SILVA; THEAKSTON; CRAMPTON, 1996) além de outros mecanismos de diversificação como a transposição (DOLEY; MACKESSY; KINI, 2009), perda (CASEWELL et al., 2011) ou recombinações pré- ou pós-transcrição de domínios funcionais (MOURA-DA-SILVA et al., 2011), ou ainda modificações pós-traducionais (FOX; SERRANO, 2008). Como resultado desse processo, as SVMPs apresentam grande diversidade estrutural e funcional, sendo componentes fundamentais para a ação dos venenos.

As hipóteses relacionadas à evolução e diversificação das toxinas ainda carecem de um suporte genômico para poderem ser melhor entendidas. Ainda se sabe pouco sobre o número de cópias gênicas e só recentemente têm sido disponibilizados segmentos longos do genoma de serpentes. Não se sabe tampouco a localização genômica e outras informações estruturais para a maioria das toxinas. Dessa forma, a organização gênica das toxinas é extremamente especulativa, pois se baseia quase integralmente em extrapolações feitas a partir da estrutura dos mRNAs (cDNAs) e proteínas e não da direta observação dos genes.

Assim, entendemos que a genômica de répteis é pouquíssimo estudada apesar dos mesmos constituírem um grupo muito representativo da fauna mundial e, em especial, da brasileira. Especialmente entre os ofídios, a falta de bancos de dados com

informações genômicas é limitante para o entendimento de sua biologia e da evolução de seu veneno. Com isso, a utilização das tecnologias mais avançadas de análise é prejudicada, uma vez que um genoma de referência é hoje praticamente uma regra para estudos moleculares.

Sendo *Bothrops jararaca* uma das espécies de serpente mais bem estudadas, endêmica de nosso país, presente em todo Estado de São Paulo, responsável pelo maior número de acidentes e cujo veneno já subsidiou o desenvolvimento concreto de um fármaco amplamente usado na medicina, a consideramos um modelo ideal para um estudo sobre genomas ofídicos.

2.5 ESTRATÉGIAS EXPERIMENTAIS

Atualmente a maioria dos projetos de sequenciamento genômico tem utilizado as tecnologias NGS, que se baseiam no sequenciamento massivo e paralelo de uma coleção de fragmentos de DNA. Dessa forma, é possível gerar uma quantidade imensa de dados em uma única “corrida” do aparelho. As plataformas de segunda geração mais conhecidas hoje são: Illumina, Roche 454 (a qual foi descontinuada recentemente pelo fabricante) e Ion Torrent. Todas estas variam no tamanho das leituras geradas, quantidade de dados e tecnologia empregada. As mesmas também exigem a preparação prévia do material a ser sequenciado, que consiste basicamente na fragmentação do DNA alvo, reparação das pontas, ligação de adaptadores e amplificação dos fragmentos. O produto final após esta etapa é o que chamamos de bibliotecas de sequenciamento.

A tecnologia Illumina baseia-se na criação de “clusters” de DNA, o que envolve a amplificação clonal de fragmentos de DNA ligados a uma superfície (“flow cell”) por PCR em ponte. O sequenciamento é feito pela identificação específica de nucleotídeos marcados com fluoróforos incorporados durante a síntese de uma nova fita em cada *cluster*. A cada ciclo o fluoróforo é removido permitindo o início de um novo ciclo. A *flow cell* apresenta bilhões de *clusters* e os sinais emitidos durante as incorporações em cada *cluster* são captados por uma câmera e as imagens são armazenadas (Figura 4).

Figura 4 - Esquema mostrando o *workflow* do sequenciamento em plataforma Illumina. (Imagem: Stuart M. Brown)

A Roche 454 utiliza um método que amplifica o DNA dentro de gotículas de água em uma emulsão de óleo (PCR em emulsão), com cada gotícula contendo um fragmento de DNA ligado a uma microesfera. Após a reação, há uma expansão clonal do DNA ligado a cada microesfera. O sequenciador contém um suporte com muitos poços, cada um contendo uma única microesfera e reagentes necessários para o sequenciamento. Aqui a técnica utilizada é o Pirosequenciamento, que utiliza a luciferase para gerar luz a cada liberação de pirofosfato quando há a incorporação correta de nucleotídeos individuais, que são adicionados um por vez. Os sequenciadores 454 tem a vantagem de gerar *reads* relativamente longos (até 700 pb), embora o volume de dados seja pequeno em relação aos outros sequenciadores.

A plataforma Ion Torrent também utiliza o PCR em emulsão de maneira similar ao 454, no entanto, o método de detecção é baseado na liberação de um íon H^+ a cada

incorporação de um nucleotídeo complementar à cadeia molde. Este íon é detectado por um sensor em um *chip* semiconductor, sendo o único dos sistemas citados que não é baseado em imagens.

De todas essas plataformas, a Illumina é a mais utilizada por ser bastante poderosa, podendo gerar até seis bilhões de reads por corrida, com alta qualidade, apresentando o menor custo por base e tendo relativa facilidade de utilização. Contudo, uma das deficiências do Illumina é a extensão reduzida dos *reads* gerados quando comparados aos de outras tecnologias, mas ainda assim úteis para um sequenciamento do tipo *whole genome*. Desse modo, este é um instrumento que atende aos requisitos de um projeto genoma e vem sendo amplamente utilizado para tal propósito (AMEMIYA et al., 2013; LIU et al., 2015; WAN et al., 2013).

O sequenciamento na plataforma Illumina (e em outras) permite determinar a sequência das duas extremidades dos fragmentos de uma biblioteca, sendo esse tipo de sequenciamento conhecido como *paired-end*. As bibliotecas podem ter tamanhos de fragmentos diferentes, mas em geral a média de tamanhos obtidos pode ser determinada por protocolos específicos. Bibliotecas longas ou curtas podem ser construídas de acordo com os objetivos do projeto e a informação da distância entre as extremidades é útil na montagem de *contigs* e *scaffolds*. Apesar disso, o protocolo *paired-end* não permite o sequenciamento de fragmentos muito longos e para vencer essa limitação há outro tipo de sequenciamento chamado de *Mate-pair*. Neste protocolo fragmentos longos de tamanhos definidos são selecionados e adaptadores biotinilados são ligados nas extremidades. Essas moléculas são então circularizadas e refragmentadas gerando moléculas menores. Os subfragmentos contendo biotina são enriquecidos, reparados e novos adaptadores são ligados nas extremidades. Daí em diante o protocolo torna-se parecido com o *paired-end*, com a diferença de que as extremidades sequenciadas correspondem à de um fragmento longo.

Além dos tipos de sequenciamento mencionados, novos métodos têm surgido recentemente e têm sido chamados de sequenciamento de terceira geração. Entre eles o SMRT (*Single-Molecule Real-time Sequencing*) é uma tecnologia promissora desenvolvida pela empresa Pacific Biosciences (PACBIO) e que vem ganhando espaço. As bibliotecas são construídas com adaptadores em forma de grampo ligados às extremidades do DNA molde, de maneira tal que a molécula resultante é estruturalmente uma dupla fita linear, mas os adaptadores formam bolhas criando uma

molécula topologicamente circular. O DNA é então imobilizado juntamente com uma DNA polimerase altamente processiva (Phi29) em um nanopóço de um suporte de vidro (*zero-mode waveguides* - ZMWs). A reação de polimerização se desenrola e nucleotídeos marcados com um fluoróforo no grupo fosfato são incorporados enquanto uma câmera de alta resolução no fundo do ZMW registra a fluorescência dos nucleotídeos sendo incorporados em tempo real. Como a molécula é circular, os insertos são lidos várias vezes no sentido senso e antisenso, o que gera uma sequência consenso (*circular consensus sequence*). Mas a redundância é dependente do tamanho do inserto, de forma que insertos pequenos apresentam alta redundância o que aumenta consideravelmente a acurácia da técnica. Todavia, para insertos longos o SMRT *sequencing* apresenta uma alta taxa de erros, o que diminui a qualidade dos reads e requer uma alta cobertura para montagem *de novo* de genomas. A tecnologia PACBIO já foi utilizada para montagens *de novo* de diversos genomas, especialmente de microrganismos, mas também para vertebrados, incluindo humanos (RHOADS; AU, 2015). Nesse caso, a alternativa mais comum é corrigir os reads PACBIO usando reads de tecnologias de segunda geração (KOREN et al., 2012).

Além do sequenciamento de próxima geração, uma antiga estratégia ainda se mostra eficiente. Trata-se das bibliotecas genômicas de fragmentos longos do tipo BACs (*Bacterial Artificial Chromosomes*). As bibliotecas de BACs são feitas basicamente pela ligação de fragmentos longos de DNA em um vetor baseado no plasmídeo F (*Fertility factor*) de bactérias (LUO; WING, 2003). Este plasmídeo suporta insertos em torno de 150 a 300 kb e os constructos resultantes são transfectados em bactérias *E. coli*. As bactérias transformadas com os BACs são selecionadas por um antibiótico e dão origem a clones com fragmentos específicos e enriquecidos de DNA exógeno. Os clones de interesse podem ser prospectados e sequenciados isoladamente, o que diminui sobremaneira a complexidade de montagem de segmentos determinados. Adicionalmente, os segmentos de DNA clonados podem ser usados como sondas para mapeamento genômico físico por técnicas como a hibridização *in situ* com fluorescência (FISH). Em muitos genomas de vertebrados foi utilizado o sequenciamento das extremidades de BACs (BAC-ends) para melhoramento da montagem e ancoramento nos cromossomos por FISH, por exemplo, na montagem do genoma de *Anolis carolinensis*. Apesar das vantagens, esse é um método tecnicamente difícil e laborioso por ser

necessário lidar com moléculas longas altamente instáveis e com equipamentos específicos.

Para o sucesso da montagem de um genoma complexo, é fundamental lançar mão de diferentes estratégias de sequenciamento, uma vez que cada uma possui vantagens e desvantagens características. Neste trabalho foram utilizadas diversas estratégias (Figura 5), que incluem: o sequenciamento *shotgun*, que gera bilhões de sequências curtas, porém com alta qualidade; e *Mate-pair*, que auxilia na ordenação dos contigs montados para geração dos scaffolds (ambos os tipos de sequenciamento foram realizados em plataforma Illumina); a construção de uma biblioteca de insertos longos em BACs (cromossomos artificiais de bactéria); e finalmente o sequenciamento em plataforma PACBIO, a qual gera reads longos, no entanto com baixa qualidade.

Além disso, foi feito o transcriptoma de vários tecidos de *B. jararaca* para mapeamento no genoma e predição de genes. Essa abordagem pode facilitar a identificação de elementos expressos do genoma, além de permitir uma análise comparativa em larga escala entre transcritos da glândula de veneno e de outros tecidos, possibilitando a identificação dos principais genes da glândula de veneno envolvidos em rotas celulares de produção do veneno (JUNQUEIRA-DE-AZEVEDO et al., 2015). Tais dados também foram utilizados como guia para facilitar a montagem da estrutura de genes de interesse.

Figura 5 - Estratégia experimental simplificada para o sequenciamento do genoma de *B. jararaca*.

3 OBJETIVOS

Este trabalho teve como objetivo o sequenciamento do genoma da serpente *Bothrops jararaca* visando uma análise focada em elementos relacionados ao seu veneno.

Objetivos específicos:

- Obter um conjunto de sequências, segmentos e mapas genômicos da serpente *B. jararaca*;
- Identificar neste genoma genes de toxinas conhecidas dos venenos de serpentes *Viperidae*;
- Descrever as estruturas organizacionais de alguns desses genes, incluindo íntrons e regiões flanqueadoras;
- Identificar elementos expressos por correlação com bancos transcriptômicos;
- Prover uma base de dados genômicos relevante ao estudo da biologia geral e comparativa de serpentes.

4 METODOLOGIA

4.1 ESPÉCIMES

Dois exemplares da serpente *Bothrops jararaca* (Viperidae), fêmeas, foram cedidos pelo Laboratório de Herpetologia do Instituto Butantan. Os animais foram mortos após anestesia com pentobarbital sódico 30 mg/kg por via subcutânea, seguindo protocolo 1131/13 aprovado pela Comissão de Ética no Uso de Animais do Instituto Butantan. Foram retiradas amostras de diversos tecidos, sendo que o sangue e fígado foram usados para extração de DNA. As carcaças dos exemplares foram tombadas na Coleção Herpetológica do Instituto Butantan com números de tombo IBSP84406 e IBSP 84408, como subamostra da Autorização de Acesso ao Patrimônio Genético No 010713/2013-1.

4.2 OBTENÇÃO DE DNA

Para fins de sequenciamento em sistema Illumina, a extração de DNA foi feita utilizando o DNeasy Blood and Tissue kit (Qiagen) a partir do sangue de um dos indivíduos de *B. jararaca* citados acima (IBSP84406), seguindo as instruções do fabricante. A concentração aproximada, a relação A260/280 e A260/230 foram verificadas através de nanofotômetro (NanoDrop). A integridade do DNA foi avaliada por eletroforese em gel de agarose 1% e a quantificação por fluorimetria usando-se o kit Quant-IT Picogreen (Life Technologies). O material a ser sequenciado também foi analisado por eletroforese capilar em chips do sistema Bioanalyzer (Agilent).

4.3 OBTENÇÃO DE RNA] E SEQUENCIAMENTO EM PLATAFORMA ILLUMINA

Os tecidos preparados nessa etapa (glândulas de veneno, fígado, músculos intercostais, pâncreas, cérebro, pulmão, músculos da mandíbula, estômago, rim, intestino e coração) foram triturados em homogeneizador Polytron PT3100 e o RNA total isolado com reagente TRIZOL (Invitrogen), uma modificação do método descrito por Chomczynski e Sacchi (1987). Os RNAs polyA⁺ foram obtidos por purificação com *beads* magnéticos (DYNAbean, LifeTechnologies). A quantidade de mRNA foi estimada por espectrofluorescência mediante o uso de um marcador fluorescente e curva padrão (Quant-iT RiboGreen, Invitrogen). A qualidade dos RNAs totais e RNAs polyA⁺ foi

avaliada por eletroforeses em chip picoRNA em sistema Bioanalyzer (Agilent Technologies).

As bibliotecas para sequenciamento em plataforma Illumina foram construídas utilizando o kit TruSeq RNA Sample Prep Kit v2 a partir de 300 ng de mRNA e sequenciadas em equipamento HiSeq1500 no Instituto Butantan, conforme recomendações do fabricante.

4.4 PREPARO E SEQUENCIAMENTO DE DNA EM SISTEMA ILLUMINA

Foram construídas no total seis bibliotecas para sequenciamento na plataforma Illumina, na Faculdade de Ciências Agrárias da UNESP, em Jaboticabal.

A primeira biblioteca é do tipo *shotgun*, onde o DNA é fragmentado em segmentos pequenos (~500pb) e ligado a adaptadores que servem como sítios de anelamento para sequenciamento aleatório em ambas as pontas. Essa biblioteca foi construída segundo o protocolo do TruSeq DNA LT Sample Prep Kit, Illumina, Inc.

As outras bibliotecas, J1, J2, J3, J4 e J6, foram do tipo *mate-pair*. Nesse caso, o DNA é fragmentado em segmentos maiores, no caso 2 kb, 5 kb, 10 kb, 15 kb e numa faixa de 2 a 15 kb, respectivamente. Em seguida, esses fragmentos são circularizados com um adaptador contendo uma molécula de biotina e fragmentados novamente em segmentos menores (~500 pb). Os segmentos contendo o adaptador são capturados por afinidade a estreptavidina, de forma que as sequências geradas de ambas as pontas dessas moléculas estarão a uma distância conhecida no DNA original. Para essas bibliotecas foi utilizado o protocolo Nextera Mate Pair Library, Illumina, Inc.

Os sequenciamentos foram realizados em dois equipamentos: Lote 1 e 2 no equipamento HiScanSQ da UNESP-Jaboticabal e lotes 3 e 4 no equipamento HiSeq1500 no Instituto Butantan, conforme recomendações do fabricante. Em todos os casos, foram realizados sequenciamentos do tipo paired-end com 100 bases de extensão.

4.5 ANÁLISE BIOINFORMÁTICA DAS SEQUÊNCIAS GENÔMICAS

As sequências genômicas dos quatro lotes foram analisadas individualmente quanto à qualidade e quantidade de sequências geradas, processadas e, em seguida, foram iniciados os testes de montagem das sequências, conduzidos inicialmente no Laboratório de Bioinformática do Instituto de Química da USP, sob coordenação do Prof.

Dr. João Setubal. A melhor montagem foi feita com o software ALLPATHS-LGv2, utilizando-se um subconjunto de reads de uma das lanes (L1) e foi considerada a versão 1.0 do genoma. Esta foi realizada pelo bioinformata do CeTICS Dr. Milton Nishiyama Jr.

Os principais parâmetros utilizados para pré-processamento das leituras pelo ALLPATHS-LG e montagem *De novo* foram PLOYDY=2, MAX_MEMORY_GB=512, MIN_CONTIG=1000, HAPLOIDIFY=True. A etapa seguinte, corresponde à melhora na extensão e criação de novos *scaffolds* a partir da montagem obtida com AllPaths-LG, por meio do software SSPACE, podendo incorporar leituras *paired-end*, *mate-pair* e contigs, com os parâmetros: “SSPACE_Standard_v3.0.pl -T 70 -l libraries.pe.mpe.bac.txt -s BjararacaAssemblyGenomeClean1.fasta -k 5 -v 1 -z 100 -a 0.7 -x 0 -m 35 -o 20 -b jararaca_scaffolds1”. Após a extensão e criação dos *scaffolds* com SSPACE, integrado ao *pipeline*, o software GapCloser foi empregado para fechamento dos gaps.

Para avaliar a integridade do genoma montado foi utilizado o CEGMA (*Core Eukaryotic Genes Mapping Approach*) (PARRA; BRADNAM; KORF, 2007), para acessar a integridade da montagem em termos de um conjunto altamente conservado de genes em eucariotos.

Para as buscas por elementos repetitivos e transponíveis foi utilizado o RepeatMasker versão 4.0.5, com RepBase database versão 20.05. As buscas foram efetuadas baseadas em elementos relacionados com o clado *Squamata*. Para o mapeamento dos *reads* transcriptômicos no genoma foram utilizados os softwares Tophat2 (KIM et al., 2013) e bowtie2 (LANGMEAD; SALZBERG, 2012).

Para predição e anotação gênica foram utilizados os softwares PASA2 e Augustus (mais detalhes junto aos resultados).

O software CLC Genomics foi utilizado para anotações manuais, curagem de anotações em scaffolds de interesse da versão 1.0 (também para sequências de BACs – ver adiante), para mapeamento de reads genômicos, dos BACs e de transcriptomas de tecidos em regiões de interesse. Reads PACBIO foram utilizados em casos isolados para facilitar o entendimento de estruturas genômicas.

4.6 PREPARAÇÃO DE PLUGS DE AGAROSE PARA CLONAGEM DE DNA

Para fins da construção da biblioteca de BACs, amostras de sangue (tecido escolhido por conter grande quantidade de hemácias nucleadas em suspensão) foram colhidas para preparação de plugs de imersão em agarose. Foi feita a contagem do

número de células sanguíneas totais, através do contador de partículas *Beckman Coulter* e por contagem em câmara de Neubauer. A seguir a amostra de sangue foi centrifugada a 800 g, 4 °C por 5 minutos. O sobrenadante foi descartado e as células foram ressuspensas em PBS (tampão fosfato). Esse processo foi repetido mais duas vezes. No final, a suspensão de células foi diluída para uma concentração aproximada de 5×10^7 céls/ml. Essa suspensão de células foi equilibrada a 45 °C por 10 minutos e misturada com Agarose Low-Melting 1% (InCert® Agarose-Lonza), previamente equilibrada a 45 °C, em uma proporção de 1:1. A mistura foi colocada em fôrma apropriada, e deixada no gelo para solidificação e obtenção dos blocos (plugs) de agarose.

Os plugs foram colocados em uma solução de EDTA 0,5M, n-lauroilsarcosina 1% e Proteinase K 200 µg/mL a 50 °C durante 48 h, trocando-se a solução após 24 h. Em seguida os plugs foram lavados em 50 mL de TE pH 8,0 (10 mM Tris, 1 mM EDTA) por 1h sob agitação a temperatura ambiente. O tampão TE pH 8,0 foi trocado e os plugs foram incubados por 1 h a 50 °C. O tampão foi substituído novamente por TE pH 8,0/PMSF 200 µM e os plugs foram incubados por 30 minutos a 50 °C. O processo foi repetido mais uma vez e a seguir foi feita nova lavagem com tampão TE pH 8,0 (incubados a temperatura ambiente por 1 h). Esta lavagem foi repetida mais uma vez e em seguida os plugs puderam ser armazenados em EDTA 0,5 M a 4 °C ou em etanol 70% a -20 °C (nesse caso deixou-se a 4 °C overnight e só então foram guardados a -20 °C).

4.7 CONSTRUÇÃO DE BIBLIOTECAS GENÔMICAS EM BAC

Os plugs foram lavados duas vezes com TE pH 8,0 sob agitação leve no gelo, por 30 minutos cada lavagem. Então o tampão foi trocado pelo tampão de equilíbrio da enzima HindIII (Tris-HCl 10 mM, NaCl 50 mM, DTT 1 mM), e foram realizadas duas lavagens com esse tampão, por 30 minutos cada (no gelo).

Os plugs a serem digeridos foram cortados ao meio e cada metade foi transferida para um tubo *eppendorf* de 1,5 mL contendo 1 mL de tampão de equilíbrio HindIII. Cada tubo correspondendo a certo tempo de digestão a ser avaliado (1 h, 2 h, 3 h e *overnight*). Foi preparada uma solução diluída a 0,1 U/µL da enzima HindIII. A seguir 10 µL dessa solução foram adicionadas aos tubos, os quais foram deixados equilibrando em gelo por 30 minutos. Após este tempo, foram adicionados 10 µL de uma solução de MgCl₂ 1 M, incubou-se por 30 minutos no gelo novamente. Finalmente os tubos foram deixados a 37 °C, e cada tubo foi removido em seu respectivo intervalo de tempo.

Primeira seleção de tamanho - O tempo de digestão escolhido foi o de 3 h por ter sido o que apresentou maior dispersão de fragmentos na faixa desejada (~200 kb). Dessa forma, seis plugs foram submetidos à digestão parcial. Após a digestão, foi feita a primeira seleção de tamanho através do PFGE. Os parâmetros utilizados foram: Gel de agarose 1%; tampão TAE 1 X; Voltagem: 6.0 V/cm; temperatura de 12 °C; 120° ângulo de orientação; tempo de corrida: 18 h; Pulso Inicial: 1 s; Pulso final: 40 s; Ramping factor: Linear.

São colocados dois marcadores de peso molecular dos lados esquerdo e direito do gel, a fim de servirem de guia para cortar a fração de 150 kb a 350 kb. Como o DNA não pode ser corado com Brometo de Etídeo e nem exposto à luz UV, apenas as porções mais externas do gel são coradas (as laterais do gel foram cortadas de forma a conter o marcador de peso molecular e parte do DNA separado). A região que contém DNA de 150 kb a 350 kb foi cortada e usada para a segunda seleção de tamanho, com o objetivo de eliminar fragmentos menores ainda presentes.

Segunda seleção de tamanho - O bloco cortado na primeira seleção de tamanho foi dividido em duas frações: A (menor peso ~150 a 250 kb) e B (alto peso ~250-350 kb). As frações foram dispostas no topo de um novo gel de agarose low-melting 1% (SeaPlaque® GTG™ agarose/Lonza) da esquerda para a direita, respectivamente. Os parâmetros utilizados na segunda corrida foram:

Tampão TAE 1 X; Voltagem: 5.0 V/cm; Temperatura de 12 °C; 120° ângulo de orientação; tempo de corrida: 13 h; Pulso Inicial: 3 s; Pulso final: 4.5 s; Ramping factor: Linear. Um bloco de agarose correspondendo a cada fração (onde o DNA está mais concentrado) foi cortado e imediatamente transferido para um tubo *Falcon* contendo 50 mL de TAE 1 X.

Eletroeluição - A eletroeluição foi feita em equipamento gentilmente cedido pela professora Anete Pereira de Souza, no Laboratório de Análise Genética e Molecular (UNICAMP). Os blocos correspondendo a cada fração foram submetidos à eletroeluição (*electroeluter 422* – *BioRad*) no mesmo dia.

Ligação e transformação - O DNA eletroeluído foi quantificado por densidade de pixels em gel de agarose utilizando o software ImageJ2 (SCHINDELIN et al., 2015) e utilizado para ligação no vetor pIndigoBAC536, preparado em nosso laboratório a partir do vetor pCUGIBAC1 (cujo clone foi gentilmente cedido pelo Prof. Dr. Jesus Ferro, FCAV UNESP-Jaboticabal) seguindo o protocolo de LUO e WING (2003). A reação de ligação foi

conduzida a 16 °C overnight, seguida por inativação da enzima a 65 °C durante 15 minutos. O produto da ligação foi aplicado em tubos de agarose 1%/Glicose 0,1 M para dessalinização, durante 1 h em banho de gelo. A seguir 13 µL da ligação e 100 µL de células eletrocompetentes TransforMax™ EPI300™ foram misturados e deixados em gelo por 10 minutos. Então foi feita a eletroporação a 13 kV/cm e as células eletroporadas foram ressuspensas em 2 mL de meio SOC. Cem microlitros correspondentes a cada fração foram plaqueados (Meio LB + cloranfenicol 12.5 µg/ml + X-Gal 40 µg/ml + IPTG 0.4 mM) e deixados a 37 °C overnight.

4.8 SEQUENCIAMENTO DOS BACs

Para a análise dos BACs, inicialmente foram selecionados aleatoriamente 37 clones, os quais foram preparados utilizando um protocolo padrão de lise alcalina, sendo que algumas amostras passaram por um filtro MultiScreen MAGVN2250 (Millipore). O DNA foi quantificado por fluorimetria em um equipamento FlexStation3 (Molecular Devices) usando o kit de quantificação de DNA “Quant-it™ PicoGreen® dsDNA Assay Kit” (Invitrogen).

Foram construídas 3 bibliotecas de sequenciamento para *pools* de 10 BACs (bibliotecas 1, 2 e 3) e outras 7 para BACs individuais. As bibliotecas foram construídas utilizando o *Nextera DNA Sample Preparation Kit*. O sequenciamento paired-end desses BACs foi realizado em equipamento Illumina MiSeq® do Centro de Estudos do Genoma Humano (IB-USP). Este é um equipamento mais simples adequado para o sequenciamento de pequenos genomas e que gera leituras mais longas de até 300 pb.

As sequências de BACs foram filtradas, removendo-se as de baixa qualidade, as relacionadas ao vetor, adaptadores e DNA genômico de *E. coli*. Posteriormente foram conduzidos testes de montagem com vários k-mers utilizando os softwares SGA (SIMPSON; DURBIN, 2012), ABySS (SIMPSON et al., 2009), Soapdenovo e CLC Genomics. Maiores detalhes são fornecidos junto aos Resultados.

Posteriormente, desenvolvemos uma estratégia de sequenciar *pools* de BACs em larga escala. Para isso, *pools* de 12 BACs foram usados para construir bibliotecas de sequenciamento no sistema Illumina com o kit *Nextera XT DNA library prep kit*. Cada biblioteca recebeu um adaptador específico (Illumina index) de forma que até 384 bibliotecas (derivadas de 384 *pools* de 12 BACs) foram reunidas para uma rodada de sequenciamento no modo Rápido de um equipamento Illumina HiSeq1500 (disponível

em nosso laboratório), na configuração *paired-end run* 2 x 250 bp, sendo 192 bibliotecas em cada uma das duas canaletas da *flowcell* deste modo de corrida.

Assim, em uma única rodada de sequenciamento podemos obter sequências de até 4608 BACs. Considerando que uma corrida no modo rápido gera em torno de 90 Gbp e considerando um tamanho médio de 150 kbp para os BACs, temos uma cobertura teórica de aproximadamente 130 X por BAC, calculado conforme abaixo:

1 corrida Rápida = 90 Gbp = 90.000 Mbp

90.000 Mbp / 384 *pools* de BACs = 234 Mbp por *pool*

234 Mbp / 12 BACs no *pool* = 19,5 Mbp por BAC

19,5 Mbp / 150 kbp por BAC = 130 X de cobertura por BAC

As sequências de cada *pool* foram então montadas automaticamente para se obter contigs (utilizando os softwares Soapdenovo e CLC genomics). Estes então foram utilizados para prospecção de genes de interesse. Uma vez identificados segmentos de um gene de interesse, o *pool* pode ser desmembrado e ressequenciado para determinação da sequência completa do BAC em questão. Como a estratégia se mostrou bem sucedida foi então repetida com 384 novos pools (4608 BACs).

4.9 DESENHO DE PARES DE PRIMERS ESPECÍFICOS PARA SELEÇÃO DE BACs

Os *primers* específicos para genes de toxinas das famílias SVMPS, SVSPs e svVEGF foram desenhados (Tabela 1) de acordo com as sequências obtidas na montagem dos pools de BACs e por comparação com sequências de cDNAs das respectivas toxinas. As condições ideais de temperatura das reações de PCR foram padronizadas por gradiente de temperatura em equipamento *Eppendorf Mastercycler ep Gradient S* (dados não mostrados).

Tabela 1 - Sequência dos pares de primers desenhados para amplificar por PCR segmentos dos genes de interesse a serem rastreados nos clones de BACs.

Gene alvo	Nome do primer	Sequência (5' → 3')
svVEGF	VEGF_F	CCAAGTTCCCCTTTGTCTGA
	VEGF_R	TGAAAGTCACGCAGATCCAG
SVSP 1	SVSP_F	GGAGTTCAATTSTTCTATAAGG
	SVSP_R	GGCAGATCACATCTTTATTTCC
SVMP	SVMP_F	CHTTATTGAACCCTTGRAGC
	SVMP_R	GAGTAAGATTTRACTGAGAG

4.10 IDENTIFICAÇÃO E VALIDAÇÃO DOS POOLS DE BACs DE INTERESSE.

Após rastrear os *pools* de BACs de origem das sequências genéticas de interesse, conforme explicado nos resultados, alíquotas destes *pools* foram coletadas nas placas de estoque, cultivadas em meio LB líquido seletivo a 37 °C por 12 h. Em seguida o DNA foi extraído e purificado por lise alcalina. As amostras de DNA dos *pools* foram amplificadas por PCR utilizando a enzima Taq DNA Polymerase recombinante (Invitrogen), nas condições estabelecidas anteriormente para cada *locus* de interesse, e a existência dos genes de interesse nos *pools* foi validada pela observação da banda do tamanho esperado em gel de eletroforese.

Após a identificação de um *pool* de BAC, o BAC individual de origem da sequência genômica de interesse foi investigado. Para isso, alíquotas dos clones que compõem este *pool* foram coletadas de sua placa de estoque e inoculadas individualmente em meio LB líquido seletivo a 37 °C por 12 h. O DNA foi extraído e purificado por lise alcalina. A identificação foi realizada seguindo o mesmo procedimento feito com o *pool*.

4.11 RESSEQUENCIAMENTO DE BACs EM SISTEMA ILLUMINA.

Após a identificação dos BACs de interesse, bibliotecas de sequenciamento Illumina foram preparadas para cada BAC segundo o procedimento a seguir. Os clones de interesse foram plaqueados em meio sólido LB/água seletivo, a 37 °C por 16 h. Para cada BAC, duas colônias isoladas foram coletadas e cultivadas isoladamente em meio de cultura líquido LB seletivo a 37 °C por 16 h. O DNA foi extraído e a presença do

fragmento genômico de interesse foi confirmada por PCR. Uma vez selecionadas para compor a biblioteca de sequenciamento, as amostras de DNA foram quantificadas conforme descrito anteriormente. Em seguida, as bibliotecas de sequenciamento foram preparadas usando o protocolo do kit “Nextera XT DNA Library Prep Kit” (Illumina) de acordo com as instruções do fabricante. As bibliotecas foram sequenciadas em uma plataforma de nova geração Illumina MiSeq (Programa Equipamento Multiusuário, processo FAPESP 2014/50148-7) no Centro de Pesquisas sobre o Genoma Humano e Células-tronco (Instituto de Biociências, USP, São Paulo/SP) e montadas com o software SPAdes.

4.12 SEQUENCIAMENTO EM SISTEMA PACBIO

Amostras de DNA de sangue e rim de *B. jararaca* foram extraídas conforme descrito no item 4.2. O DNA foi analisado em gel de agarose e quantificado por fluorimetria usando o kit Quant-iT PicoGreen dsDNA Assay Kit (Thermo). Três tubos contendo 4 ug de DNA do sangue e 1 tubo com 4 ug de DNA do rim de *B. jararaca* foram enviados para o *Genome Sequencing Shared Resource* da Duke University, em Durham, Carolina do Norte, EUA. A exportação do DNA recebeu autorização do IBAMA No 15BR017067/DF. Três bibliotecas de sequenciamento foram preparadas na *facility* (uma de rim e duas de sangue), selecionando fragmentos acima de 15 kb, com o intuito de gerar reads mais longos. As bibliotecas foram sequenciadas em 60 SMRT cells utilizando a química P6-C4 em plataforma PacBio RS II.

5 RESULTADOS

5.1 EXTRAÇÃO DE DNA GENÔMICO PARA CONSTRUÇÃO DE BIBLIOTECAS DE SEQUENCIAMENTO

As purificações de DNA para uso na construção de bibliotecas para sequenciamento tipo *shotgun* foram realizadas com kit comercial a partir de amostras de sangue e fígado, conforme descrito na metodologia.

A figura 6 apresenta o perfil do material purificado e a tabela 2 descreve as quantidades e pureza observadas.

Figura 6 - DNA genômico purificado de *B. jararaca*.

M – Marcador; F1 – Fígado (primeiro eluato); F2- Fígado (segundo eluato); S1 – Sangue (primeiro eluato); S2 – Sangue (segundo eluato).

Tabela 2 - Extração de DNA genômico de *Bothrops jararaca*.

Amostra	Concentração (ng/μL)	A260/280	A260/230	Utilização
Fígado (1º eluato)	499,9	1,71	1,11	Estoque
Fígado (2º eluato)	51,4	1,71	0,75	Estoque
Sangue (1º eluato)	111,0	1,90	1,88	Illumina Shotgun
Sangue (2º eluato)	69,2	1,85	2,19	Illumina mate-pair

Observou-se que apesar do DNA extraído do fígado possuir uma concentração bem mais alta do que a do DNA extraído do sangue, a qualidade deste último estava melhor em termos de pureza e integridade. Dessa forma, as amostras de sangue foram

utilizadas para construção das bibliotecas de sequenciamento em plataforma Illumina, conforme descrito a seguir. Este material também foi enviado aos EUA para sequenciamento em sistema PacBio, conforme mencionado no item Metodologia.

5.2 PREPARO DE BIBLIOTECAS GENÔMICAS PARA O SISTEMA ILLUMINA

A construção das bibliotecas foi realizada como serviço no Departamento de Tecnologia da Faculdade de Ciências Agrárias da UNESP de Jaboticabal, seguindo o protocolo descrito na metodologia.

Numa primeira etapa, foi construída a biblioteca *shotgun*. A figura 7 ilustra o perfil analítico da biblioteca terminada.

Figura 7 - Perfil eletroforético microcapilar em chip *HighSensitivity* do sistema *Bioanalyzer*, mostrando a distribuição de fragmentos da biblioteca *shotgun*.

Num segundo momento foram construídas as bibliotecas do tipo mate-pair, também na UNESP de Jaboticabal. Estas bibliotecas são importantes para a estratégia de montagem, pois permitem conhecer a sequência das extremidades de segmentos de tamanho definido, auxiliando o desenho de *scaffolds* nas montagens.

Definimos a construção de bibliotecas com os seguintes tamanhos: J1=2 kb, J2=5 kb, J3=10 kb e J4 ou J5=15 kb, conforme utilizado em outros trabalhos. A figura 8 mostra o perfil da eletroforese microcapilar das amostras fragmentadas e selecionadas, indicando a obtenção dos tamanhos de fragmento desejados.

Figura 8 - Perfil eletroforético microcapilar em chip DNA12000 do sistema Bioanalyzer (visto como simulação de gel de agarose) mostrando a distribuição de fragmentos a serem utilizados nas bibliotecas *mate-pair*.

Após a seleção de tamanhos, as bibliotecas de sequenciamento foram construídas resultando no perfil ilustrado na figura 9.

Figura 9 - Perfil eletroforético microcapilar em chip HighSensitivity do sistema Bioanalyzer, mostrando a distribuição de fragmentos das bibliotecas *mate-pair* quando finalizadas.

Por esse perfil, observou-se que a biblioteca J4 (15 kb) praticamente não continha material. O mesmo protocolo foi repetido, mas sem sucesso (biblioteca J5). Isso possivelmente está relacionado com a pouca quantidade de material recuperado após a seleção de tamanhos (vide figura 8). Segundo explicações do pessoal técnico da empresa Illumina essa faixa de tamanho é bastante difícil de ser obtida e por isso recomendaram,

alternativamente, a construção de uma biblioteca de faixa variada (2-15 kb), pois nesse protocolo não há purificação de material, diminuindo a perda. Dessa forma, foi feita uma nova biblioteca (J6), cujo perfil está ilustrado na figura 10. Como se pode notar, nesse caso foi obtido um perfil indicativo da presença de biblioteca suficiente para o sequenciamento.

Figura 10 - Perfil eletroforético microcapilar em chip HighSensitivity do sistema Bioanalyzer, mostrando a distribuição de fragmentos da biblioteca *mate-pair* J6 (2kb-15kb) quando finalizada.

A quantificação das bibliotecas finais foi feita em qPCR, que ponderado pelo tamanho médio dos fragmentos indicou as quantidades mostradas na tabela 3.

Tabela 3 - Concentração de cada biblioteca *mate-pair* para sequenciamento.

Biblioteca	Concentração (nM)	Tamanho selecionado (kb)
J1	1,95	2
J2	3,98	5
J3	1,53	10
J4	1,04	15
J5	0,50	15
J6	1,3	2-15

5.3 SEQUENCIAMENTO GENÔMICO EM SISTEMA ILLUMINA

A proposta inicial de execução do projeto previa o uso do equipamento HiScanSQ instalado na Faculdade de Ciências Agrárias da UNESP, onde as bibliotecas foram preparadas. De fato, duas corridas com a biblioteca *shotgun* foram realizadas nesse equipamento. A primeira utilizou 4 canaletas para uma primeira avaliação das bibliotecas. Isso resultou numa quantidade bruta de 45 Gb sequenciadas (lote 1), correspondente a uma cobertura bruta de 18 X do genoma de *B. jararaca*. Entretanto, após a análise primária, observou-se que a qualidade das sequências ficou abaixo do esperado, como pode ser observado na figura 11. Dessa forma, o número total de sequências aproveitáveis foi bem mais baixo, conforme sintetizado na tabela 4.

Figura 11 - Gráfico gerado pelo programa fastQC, ilustrando a qualidade média das bases sequenciadas em função da posição no *read* (em pb) em uma das canaletas do lote 1. Observa-se que, em média (a linha azul corresponde à média), apenas as 60 primeiras bases mantiveram qualidade (Q) >30 e com grande desvio. A linha vermelha corresponde à mediana, a barra amarela e os limites superiores e inferiores correspondem respectivamente a 25-75% e a 10-90% das bases naquela posição.

A segunda corrida no HiScanSQ utilizou 3 canaletas (lote 2). Nesse caso, os dados brutos (33 Gb) corresponderam a 13 X de cobertura bruta do genoma e foram um pouco melhores do que os do lote 1, mas ainda assim a qualidade média dos reads foi inferior ao esperado (dados não mostrados).

Os próximos sequenciamentos passaram então a ser realizados em nosso laboratório (Laboratório de Toxinologia Aplicada, Instituto Butantan), uma vez que adquirimos um equipamento Illumina HiSeq1500 que tem capacidade superior ao HiScanSQ.

O primeiro sequenciamento nesse equipamento, utilizando a mesma biblioteca *shotgun* construída em Jaboticabal, gerou 308 Gb de dados brutos, com excelente qualidade (figura 12), permitindo uma cobertura de 123 X do genoma, conforme discriminado na tabela 4.

Figura 12 - Gráfico gerado pelo fastQC, ilustrando a qualidade média das bases sequenciadas em uma das canaletas do lote 3, gerado no equipamento HiSeq1500. Observa-se que, em média, todas as 100 bases sequenciadas mantiveram qualidade (Q) >32.

Para o sequenciamento mate-pair também utilizamos o equipamento HiSeq1500. Nesse caso, as bibliotecas foram ordenadas da seguinte forma:

J1 (2 kb): lanes 1 e 2;

J2 (5 kb): lanes 3 e 4;
 J3 (10 kb): lanes 5 e 6;
 J4 (15 kb): lane 7
 J6 (2-15 kb): lane 8

Observou-se novamente uma excelente qualidade das sequências geradas, inclusive para a biblioteca J4 que indicava ter pouco material nos testes de bancada (figura 13). Essa corrida constitui o lote 4 que produziu 252 Gb (101 X de cobertura).

Figura 13 - Gráfico gerado pelo fastQC, ilustrando a qualidade média das bases sequenciadas em uma das canaletas do lote 4, correspondente à biblioteca J2. Observa-se que, em média, todas as 100 bases sequenciadas em cada read mantiveram qualidade (Q) >36.

Dessa forma, os sequenciamentos geraram uma cobertura bruta teórica total de 256 X o tamanho do genoma de *B. jararaca*. Após as filtrações de sequências de baixa qualidade, muito curtas, adaptadores, controles internos, etc., um lote final de trabalho de 562 Gb foi obtido, correspondente a 225 X o genoma de *B. jararaca* (tabela 4).

Tabela 4 - Dados gerais dos sequenciamentos na plataforma Illumina.

Lote	Lote 1		Lote 2		Lote 3		Lote 4		Total	
Local de Sequenciamento	UNESP-Jaboticabal		UNESP-Jaboticabal		I. Butantan		I. Butantan			
Equipamento e No de lanes usados	HiScanSQ (4 lanes)		HiScanSQ (3 lanes)		HiSeq1500 (7 lanes)		HiSeq1500 (8 lanes)			
Tipo da biblioteca e tipo de Seq.	Shotgun/pairend		Shotgun/pairend		Shotgun/pairend		MatePair/pairend			
Total de leituras geradas	453.572.561	100,0%	333.571.555	100,0%	3.081.125.996	100,0%	2.525.007.961	100,0%	6.393.278.073	100,0%
Total de leituras removidas (filtros abaixo)	95.119.287	21,0%	16.121.801	4,8%	103.057.718	3,3%	361.184.685	14,3%	575.483.491	9,0%
Muito curtos	74.990.983		3.370.581		54.874.597		211.639.190		344.875.351	
Baixa qualidade	622.123		223.451		5.920.008		5.146.248		11.911.830	
Homopolímeros	71.530		638.816		192.520		70.791.680		71.694.546	
Baixa complexidade	516.678		1.176.867		8.592.424		2.366.876		12.652.845	
Pureza de sinal (software CASAVA)	18.917.973		10.712.086		0		0		63.108.228	
Reads totais Phix (controle)	0		0		33.478.169		71.240.691		104.718.860	
Total de leituras boas	358.453.274	79,0%	317.449.735	95,2%	2.978.068.278	96,7%	2.163.823.282	85,7%	5.817.794.569	91,0%
Reads Trimados	344.780.113		169.210.595		432.641.718		530.573.015		1.477.205.441	23,1
Total de bases geradas	45.357.256.100	100,0%	33.357.155.500	100,0%	308.112.599.600	100,0%	252.500.796.100	100,0%		
Total de bases boas	23.948.359.522	52,8%	30.669.545.393	91,9%	292.199.751.363	94,8%	215.473.437.968	85,3%	562.291.094.246	88,0%
Cobertura bruta do genoma	18 X		13 X		123 X		101 X		256 X	
Cobertura efetiva do genoma	10 X		12 X		117 X		86 X		225 x	

Tabela 5 - Número de leituras em pares após processamento.

Lote	Equipamento	tipo do seq.	Reads bons	Reads em pares	Reads únicos	% Reads únicos
Lote 1	HScan1	pairend1	358.453.274	308.222.314	61.858.695	17,26
Lote 2	HScan2	pairend1	317.449.735	309.067.002	9.226.033	2,91
Lote 3	Hseq	pairend1	2.978.068.278	2.945.859.230	57.395.548	1,93
Lote 4	Hseq	mate J1	363.851.895	359.711.640	4.864.310	1,34
Lote 4	Hseq	mate J2	899.316.718	881.861.024	21.585.003	2,40
Lote 4	Hseq	mate J3	600.780.197	593.826.284	8.208.990	1,37
Lote 4	Hseq	mate J4	110.821.384	99.312.824	12.227.943	11,03
Lote 4	Hseq	mate J6	189.053.088	256.455.352	4.679.249	2,48

Considerando que todos os sequenciamentos foram realizados como paired-end, onde ambas as pontas são sequenciadas, avaliamos também qual a porcentagem de leituras que após o tratamento se mantém em pares ou tem uma das pontas descartada por baixa qualidade. Observamos pela tabela 5 que o Lote 1 terminou com muitas leituras desapareadas (17%). Porém nesse caso isso deve estar relacionado com a baixa qualidade do sequenciamento, visto que a mesma biblioteca apresentou porcentagens melhores em lotes diferentes. Já a biblioteca J4 também apresentou alta taxa de leituras desapareadas (11%), mas nesse caso, deve ser devido aos problemas no preparo dessa biblioteca reportados anteriormente, visto que as demais bibliotecas do lote apresentaram bons resultados.

5.4 MONTAGEM PRELIMINAR DAS SEQUÊNCIAS

Conforme descrito na metodologia, os testes de montagem das sequências foram conduzidos inicialmente no Laboratório de Bioinformática do Instituto de Química da USP, sob coordenação do Prof. Dr. João Setubal, sendo que a montagem da versão 1.0 foi realizada pelo Dr. Milton Nishiyama Junior. O software ALLPATHS-LGv2 apresentou os melhores resultados de montagem utilizando os *reads* Illumina shotgun e mate-pair (Tabela 6). Esse software foi criado para montagem *De novo* de genomas grandes e complexos, tendo como principais características: i) tratamento de regiões repetitivas, ii) correção de erros das leituras, iii) uso mais eficiente da memória, iv) tratamento com regiões de baixa cobertura de acordo com k-mer utilizado, v) normalização das leituras, vi) montagem de acordo com a ploidia, vii) montagem de contigs, viii) criação de *scaffolds*. O software SSPACE (BOETZER et al., 2011) foi utilizado para gerar os *scaffolds*, já o software GAPcloser é uma ferramenta utilizada para substituir regiões com N's

(Gaps) pela base nitrogenada correta. O N50 obtido para os contigs foi de 7,2 kb, valor pouco maior que o atingido para a serpente *O. hannah* (ver tabela 13 na Discussão). Dado um conjunto de *contigs* ou *scaffolds*, o N50 corresponde ao tamanho do contig em que o número de bases de todos os contigs mais curtos é igual ou próximo do número de bases dos contigs mais longos. O mesmo é uma métrica estatística comumente utilizada para avaliar montagens genômicas. Os dados serão disponibilizados no GenBank em breve, dentro do Bioproject PRJNA343256.

Tabela 6 - Estatísticas da melhor montagem (ALLPATHS-LGv2). Os valores de N50 de contigs e scaffolds estão destacados em vermelho.

Descrição	Montagem L1
Contig tamanho mínimo	1000
# contigs	281308
# contigs por Mb	158.7
# <i>scaffolds</i>	15748
Tamanho total contigs	1003762228
Tamanho total <i>scaffolds</i> com gaps	1399679231
N50 tamanho contig (kb)	7,2
N50 Tamanho scaffold (kb)	491
N50 tamanho scaffold (kb), com gaps	526
# <i>scaffolds</i> por Mb	11,25
Tamanho médio de gaps em <i>scaffolds</i>	740

5.4.1 ANÁLISE DE ELEMENTOS REPETITIVOS NA VERSÃO 1.0 DO GENOMA

Uma etapa a ser considerada no processo de anotação do genoma é a busca de elementos repetitivos e transponíveis. Uma das ferramentas mais adequadas para este tipo de busca é o RepeatMasker (TARAILLO-GRAOVAC; CHEN, 2009), um programa que busca por elementos repetitivos e regiões de baixa complexidade, este possui uma base curada de elementos repetitivos do RepBase (BAO; KOJIMA; KOHANY, 2015).

A taxa de repetição esperada para um genoma viperídico é de cerca de 40% (CASTOE et al., 2011) e foram identificadas aproximadamente 102 Mb de sequências repetitivas (10% do genoma). Uma das etapas da montagem *De novo* com o ALLPATHS-LG é a remoção de leituras com alta representatividade e cobertura, para não prejudicar as etapas da montagem, deste modo é muito provável que grande parte destas regiões

de elementos repetitivos e transponíveis já tenham sido removidos na fase inicial do ALLPATHS-LG e por isso apenas 10% são observados. Os elementos predominantes foram os retroelementos (correspondendo a 6.79%), dentre os quais o que predomina é o LINE (elemento nuclear intercalado longo) L2/CR1/Rex com cerca de 4.51% (tabela 7), o qual já foi descrito na literatura como muito abundante em serpentes (CASTOE et al., 2011).

Tabela 7 - Composição dos conteúdos repetitivos do genoma da *B. jararaca* derivado das análises do RepeatMasker sobre o subconjunto resultante da montagem L1.

Dados gerais			
sequências:	15450		
Tamanho total:	1409216134 bp	1060082498 bp excl N/X-runs	
Nível GC:		37.57 %	
bases mascaradas:	101622329 bp	9.59 %	
Repetições	Número de elementos	tamanho	% das sequências
Retroelementos	271764	71958071 bp	6.79 %
SINEs:	40601	5359496 bp	0.51 %
Penelope	257	71945 bp	0.01 %
LINEs:	222715	61663883 bp	5.82 %
CRE/SLACS	0	0 bp	0.00 %
L2/CR1/Rex	179552	47816093 bp	4.51 %
R1/LOA/Jockey	147	48027 bp	0.00 %
R2/R4/NeSL	5914	2467568 bp	0.23 %
RTE/Bov-B	20007	5639774 bp	0.53 %
L1/CIN4	16833	5620236 bp	0.53 %
LTR elements:	8448	4934692 bp	0.47 %
BEL/Pao	156	71877 bp	0.01 %
Ty1/Copia	273	123990 bp	0.01 %
Gypsy/DIRS1	7605	4557739 bp	0.43 %
Retroviral	373	178431 bp	0.02 %
DNA transposons	58590	6353319 bp	0.60 %
hobo-Activator	27664	2536726 bp	0.24 %
Tc1-IS630-Pogo	3478	1199006 bp	0.11 %
En-Spm	0	0 bp	0.00 %
MuDR-IS905	0	0 bp	0.00 %
PiggyBac	0	0 bp	0.00 %
Tourist/Harbinger	20719	2176599 bp	0.21 %
Other (Mirage,P-element,Transib)	0	0 bp	0.00 %
Rolling-circles	0	0 bp	0.00 %
Não classificados:	1124	106319 bp	0.01 %
Total:		78417709 bp	7.40 %
Outros			
Small RNA:	1033	72093 bp	0.01 %

Satellites:	65	25345 bp	0.00 %
Simple repeats:	404018	20102200 bp	1.90 %
Low complexity:	51649	3060937 bp	0.29 %

5.4.2 AVALIAÇÃO DA INTEGRIDADE DOS GENES DO GENOMA MONTADO

O CEGMA é composto por um conjunto de 248 famílias de proteínas conservadas, as quais se espera estarem presentes em todos os genomas de eucariotos. Portanto, quanto maior o número de genes identificados, significa que mais representativo está o conjunto de genes de determinado organismo, ou seja, é mais um parâmetro que indica a qualidade de uma montagem genômica.

Com o CEGMA aplicado à montagem L1, identificamos 164 (66%) dos genes completos e 209 (84.27%) se considerarmos alinhamentos com genes parciais, restando apenas 15% de genes não identificados.

5.4.3 MAPEAMENTO DAS LEITURAS TRANSCRIPTÔMICAS NA VERSÃO 1.0 DO GENOMA

Foram sequenciados os mRNAs de diferentes tecidos (glândula, coração, cérebro, fígado, pulmão, estômago, rim, intestino, pâncreas e músculos intercostais) da jararaca, na plataforma Illumina HiSeq1500. Após o pré-processamento e controle de qualidade das leituras para cada tecido, foi realizado o mapeamento dos *reads* contra os subconjuntos de contigs que constituem a versão 1.0 do genoma da jararaca, permitindo uma estimativa da expressão gênica, variantes e isoformas. Esse mapeamento é importante para a própria predição de genes com o software PASA2, conforme apresentado no próximo item.

A taxa média de mapeamento foi de 50% das leituras, conforme ilustrado na Tabela 8. É necessário notar que a alta estrigência no alinhamento das leituras dos transcriptomas nos *scaffolds* do genoma, excluindo pares não concordantes ou alinhamentos em múltiplas regiões, também contribui para o não mapeamento do total de *reads*.

Tabela 8 - Sumarização da taxa de pares de leituras alinhadas na montagem do genoma.

Tecidos	# leituras	%leituras mapeadas (%pares concordantes)
Cérebro	16361353	42.2% (42.1%)
Glândula	14818931	40.3% (40.1%)
Intestino	17386734	46.3% (45.9%)
Coração	6484363	46.0% (45.9%)
Rim	5739965	53.4% (53.3%)
Fígado	7596450	21.2% (21.1%)
Pulmão	17222368	44.2% (44.2%)
Estômago	6424073	53.6% (53.4%)
Músculo	13257900	48.9% (48.7%)
Pâncreas	37174360	59.4% (59.2%)

5.4.4 PREDIÇÃO DE GENES E ANOTAÇÃO DO GENOMA

Foi utilizado inicialmente o programa PASA2 (*Program to Assemble Spliced Alignments*) para prever transcritos derivados do genoma. Esta é uma ferramenta de anotação de genoma de eucariotos que explora os alinhamentos de sequências transcritas expressas para modelar estruturas de genes e preservar a anotação da estrutura gênica de forma consistente com dados experimentais. Permite identificar e classificar todas as variantes de *splicing* e isoformas, suportadas por alinhamentos contra transcritos obtidos de dados em larga escala, como RNA-seq. A utilização dos dados de RNA-seq foi feita a partir da montagem *De novo* de transcritos e também da montagem guiada pelo genoma, ambas realizadas pelo software Trinity (GRABHERR et al., 2011).

Assim foram gerados pelo PASA2 298.500 transcritos e respectivos variantes e/ou isoformas, dos quais foram preditas 59.117 proteínas com o software TransDecoder (GRABHERR et al., 2011), sendo que cerca de 33% dos transcritos (100 mil) foram anotados baseando-se nos genomas de *Protobothrops mucrosquamatus*, *Python bivittatus* e *Thamnophis sirtalis*, hoje disponíveis no GenBank. Já as montagens transcriptômicas *de novo* a partir do Trinity geraram 572.305 contigs. Após a análise destes com PASA, 39.741 transcritos foram classificados como completos e válidos, por apresentarem múltiplas leituras cobrindo e confirmando a estrutura do gene, e por terem proteína predita com códon de iniciação e parada.

Para a predição *ab initio* de proteínas codificantes de genes, foi utilizado o software Augustus (STANKE et al., 2006a), baseado em modelos de cadeias de markov e integrando um número de métodos e sub-modelos conhecidos. O Augustus emprega um novo caminho para modelar o comprimento dos íntrons e um modelo para sítios de *splicing*. Como principal característica está a possibilidade de melhorar a predição de genes, já que evidências extrínsecas sobre a estrutura de genes podem ser coletadas de várias fontes, como a comparação entre genomas, os alinhamentos com ESTs, RNAseq e proteínas, etc., as quais são combinadas com evidências intrínsecas do próprio genoma (STANKE et al., 2006b), levando a um treinamento do software.

Deste modo, para treinamento e definição dos parâmetros para a predição *ab initio* e estruturas de íntron, éxon, UTR no genoma de *B. jararaca*, foram utilizados os alinhamentos dos transcriptomas gerados por RNA-seq dos tecidos de jararaca contra o genoma, incluindo 528 genes preditos no PASA2 e classificados como completos e possuidores de dois ou mais éxons.

A combinação das abordagens para treinamento do Augustus e predição de genes que codificam para proteínas, permitiu a predição de 24.614 genes no genoma de *B. jararaca*, sendo a maior parte com suporte por sinais de RNA-seq. Essas proteínas preditas foram mapeadas por similaridade com BLASTP, com critério de e-value < 1e-6, contra proteínas disponíveis no GenBank das espécies: *Thamnophis sirtalis*, *Protobothrops mucrosquamatus* e *Python bivittatus*, além do banco de dados do Uniprot (Tabela 9), sendo identificadas 78% das proteínas preditas pelo Augustus.

Tabela 9 - Sumarização do total de proteínas preditas pelo Augustus com similaridade em bancos específicos e com Uniprot.

Espécie	# de proteínas preditas com similaridade com proteínas de cada espécie/banco
<i>P. mucrosquamatus</i>	17520 (72%)
<i>P. bivittatus</i>	14969 (60%)
<i>T. sirtalis</i>	16401 (67%)
Uniprot	15885 (65%)

5.5 PREPARO DE BIBLIOTECAS EM BAC

Foi utilizado o protocolo de clonagem de LUO e WING (2003), com pequenas modificações. Neste protocolo utiliza-se o vetor pIndigoBAC536, que deve ser preparado, digerido com enzima de restrição e desfosforilado para construção da

biblioteca de BACs. O vetor pIndigoBAC536 é de cópia única, o que torna o rendimento de sua preparação bastante baixo. Devido a isto, o protocolo utiliza o vetor pCUGIBAC1, que nada mais é que o vetor pIndigoBAC536 clonado no vetor pGEM-4Z (que é de alta cópia). Por ser um vetor de alta cópia, o rendimento da preparação do pCUGIBAC1 é bem maior. Este pode ser posteriormente digerido com uma enzima de restrição (HindIII nesse caso) e então submetido a uma corrida de PFGE, separando-se o vetor pGEM-4Z do pIndigoBAC536 (Figura 14). Finalmente, o vetor pIndigoBAC536 foi purificado do gel para clonagem dos fragmentos de alto peso.

Figura 14 - Vetor pCUGIBAC1 digerido com a enzima HindIII. Foi feita uma corrida em PFGE, seguida de purificação do vetor pIndigoBAC536 do gel por eletroeluição.

Para a confecção dos plugs foi feita a contagem do número de células sanguíneas totais, através do contador de partículas *Beckman Coulter* ($6,6 \times 10^8$ células/mL) e por contagem em câmara de Neubauer ($4,9 \times 10^8$ células/mL).

Os tempos da digestão parcial do gDNA nos plugs com a enzima HindIII foram avaliados por gel de eletroforese em campo pulsado (Figura 15). O tempo 3 h foi escolhido e então 6 *plugs* foram submetidos à digestão parcial, com recorte do material do gel para duas purificações sucessivas no tamanho desejado. Na primeira purificação o DNA foi fracionado em duas faixas de tamanho (Fração A – 150-250 kb e Fração B – 250-350 kb), já a segunda purificação foi feita para eliminar possíveis contaminantes de

baixo peso que ainda estivessem presentes, assim o DNA em torno de 150 kb ou maior ficou concentrado em uma só banda (figura 16).

Figura 15 - Digestão parcial do DNA de *B. jararaca* com a enzima HindIII.

Canaletas M: Marcador Lambda PFG. Canaleta 1: DNA não digerido. Canaleta 2 a 5: DNA digerido (meio plug) por 1, 2, 3 e 16 horas, respectivamente.

A)

B)

Figura 16 - Seleção de tamanho do DNA digerido com HindIII.

A) Primeira seleção de tamanho (Plugs contendo gDNA digerido com HindIII) e B) segunda seleção de tamanho do gDNA digerido. Fração A – 150-250kb; Fração B – 250-350kb. Observa-se as 2 faixas de tamanho recortadas. M - Marcador Lambda PFG.

O volume eletroeluído foi: A – 94 μ L; B - 91 μ L. Deste volume foram utilizados 5 μ L para determinação da concentração do DNA. Para tanto, foi feita uma corrida de eletroforese convencional utilizando DNA de lambda (íntegro) como padrão (Figura 17).

Foi utilizado o software ImageJ2 (Schindelin, J. et al., 2015) para quantificação das bandas. Comparando-se com o padrão, a banda da fração A apresentou uma intensidade bem maior, que foi estimada em 95 ng; a fração B apresenta um pouco menos, aproximadamente 74 ng. Tais medições indicaram uma concentração de 19 ng/ μ L para a fração A e 14,8 ng/ μ L para a fração B.

Figura 17 - Determinação da concentração de DNA eletroeluído do gDNA de *B. jararaca*. O DNA de lambda íntegro foi usado como padrão variando-se a concentração entre 5ng e 50ng de acordo com a figura acima. Fração A – 150-250kb; Fração B – 250-350kb.

A transformação desse material resultou num número de transformantes em torno de 60 a 70 clones para cada 100 μ L plaqueados. Embora não fosse um número elevado, era um resultado razoável considerando que foi possível fazer várias ligações, uma vez que o DNA eletroeluído estava bem concentrado. Alguns BACs foram preparados individualmente e a maioria demonstrou possuir tamanho elevado (>100 kb), conforme mostra a figura 18. Temos 9216 clones em placas de 96 poços e aproximadamente 40 ml de material transformado armazenado a -80 $^{\circ}$ C, que pode resultar em aproximadamente 26.000 clones.

Figura 18 - Digestão analítica com a enzima Not I de 12 clones (canaletas 1 a 12) da biblioteca de BACs.
Mλ – lambda PFG ladder (NEB). M1kb – Gene Ruler 1kb DNA ladder.

5.6 TESTE PILOTO DE SEQUENCIAMENTO ALEATÓRIO DE BACs EM POOLS

A biblioteca de BACs produzida neste trabalho pode ser utilizada de duas formas:

1) Para futuras buscas por sequências de interesse a partir de sondas ou ampliações por PCR. Para isso, clones dessa biblioteca foram armazenados em glicerol a -80 °C.

2) Para sequenciamento aleatório de BACs em *pools*, conforme descrito na metodologia.

Realizamos um teste piloto com 10 amostras de BACs (individuais e em *pools*) que foram sequenciados num equipamento MiSeq. As sequências das 10 bibliotecas apresentaram variações quanto à qualidade, número dos *reads* e sequências relacionadas ao vetor e ao DNA genômico de *E. coli*. A tabela 10 mostra os filtros aplicados e o número de *reads* e de bases resultantes após processamento.

Tabela 10 - Estatísticas gerais de processamento das sequências de BACs.

a)

Amostra	Biblioteca 1 <i>Pool de BACs</i>		Biblioteca 2 <i>Pool de BACs</i>		Biblioteca 3 <i>Pool de BACs</i>		Biblioteca 4 BAC individual		Biblioteca 5 BAC individual	
	#reads	%	#reads	%	#reads	%	#reads	%	#reads	%
Reads totais	153954		391600		2571944		2037176		1863644	
Reads relacionados ao vetor	5032	3,27	19074	4,87	156719	6,09	93874	4,61	87233	4,68
Reads relacionados a <i>E. coli</i>	41898	27,21	26824	6,85	220957	8,59	95132	4,67	259135	13,9
Reads após filtragem de sequências contaminantes	107024	69,51	345702	88,2	2194268	85,31	1848170	90,72	1517276	81,41
Reads muito curtos	528	0,34	5478	1,39	21998	0,85	18797	0,92	10640	0,57
Filtrados com relação à qualidade	13234	8,59	26144	6,67	185472	7,21	153690	7,54	137278	7,36
Filtrados com relação à presença de homopolímeros	60	0,038	1149	0,29	2709	0,105	1269	0,06	627	0,033
Reads de baixa complexidade	2	0,0012	29	0,007	343	0,013	53	0,002	101	0,005
Total de reads filtrados	13824	8,97	32800	8,37	210522	8,18	173809	8,53	148646	7,97
Reads trimados	87098	56,57	184375	47,08	1533307	59,61	1510520	74,14	1389695	74,56
Número total de reads com boa qualidade	93200	60,53	312902	79,9	1983746	77,13	1674361	82,19	1368630	73,43
Total de bases	1,4E+07		3,9E+07		2,9E+08		2,7E+08		2,3E+08	

b)

Amostra	Biblioteca 6 BAC individual		Biblioteca 7 BAC individual		Biblioteca 8 BAC individual		Biblioteca 9 BAC individual		Biblioteca 10 BAC individual	
	#reads	%	#reads	%	#reads	%	#reads	%	#reads	%
Reads totais	1345270		7817428		5591250		210148		1461468	
Reads relacionados ao vetor	136935	10,18	226525	2,9	52402	0,94	10378	4,94	51790	3,54
Reads relacionados a <i>E. coli</i>	177929	13,23	3037648	38,86	3736980	66,84	41814	19,9	400049	27,37
Reads após filtragem de sequências contaminantes	1030406	76,59	4553255	58,2	1801868	32,22	157956	75,16	1009629	69,08
Reads muito curtos	14251	1,05	88077	1,12	32402	0,57	841	0,4	9483	0,64
Filtrados com relação à qualidade	82151	6,1	361615	4,62	147637	2,64	16544	7,8	71506	4,89
Filtrados com relação à presença de homopolímeros	1306	0,097	27095	0,34	11773	0,21	103	0,04	2077	0,14
Reads de baixa complexidade	58	0,004	577	0,007	174	0,003	12	0,005	54	0,003
Total de reads filtrados	97766	7,26	477364	6,1	191986	3,43	17500	8,32	83120	5,68
Reads trimados	826281	61,42	2313405	29,59	1296658	23,19	117774	56,04	867924	59,38
Número total de reads com boa qualidade	932640	69,32	4075891	52,13	1609882	28,79	140456	66,83	926509	63,39
Total de bases	1,5E+08		4,7E+08		2,4E+08		2E+07		1,5E+08	

Tais variações se devem tanto ao preparo das bibliotecas, quanto ao preparo dos BACs em si e também à qualidade do sequenciamento no equipamento MiSeq®. Por

exemplo, as amostras que passaram pelo filtro MAGVN2250 (amostras 2 a 6) apresentaram uma quantidade significativamente menor de contaminação por DNA genômico de *E. coli*. Além disso, houve um problema não identificado no sequenciamento, e a qualidade da primeira leitura caiu fortemente a partir da base 70. Já com relação à segunda leitura, embora a qualidade média tenha caído por volta da base 150, se manteve alta praticamente até o final da leitura.

Após a etapa de processamento inicial, foram feitos os testes de montagem com os softwares SGA, SOAPdenovo, ABySS e CLC Genomics com diversos k-mers (com exceção do CLC que detecta automaticamente o k-mer mais adequado). O melhor k-mer para cada um dos softwares foi avaliado quanto ao número de contigs, N50 e contig mais longo. Em geral k-mers maiores (entre 80 e 90) deram melhores resultados, com algumas exceções para o software SGA nas bibliotecas 7, 8 e 10 (que tiveram melhores resultados com k-mers entre 45 e 65).

Ao final das montagens, mesmo com problemas na qualidade dos *reads*, foi possível obter contigs com até 168.588 pb para um dos BACs isoladamente (com o software CLC) e com 84.315 pb para um dos *pools* de BACs (com o software SGA). Tendo o N50 variado bastante de acordo com a biblioteca e com o montador utilizado.

As sequências de alguns BACs foram analisadas para verificar se correspondiam a segmentos genômicos de serpentes e se continham genes completos. Foram identificados vários segmentos de genes de serpentes e outros répteis, inclusive alguns genes completos, como por exemplo, um longo gene de quase 60 kb (AFAP1) que codifica uma proteína associada ao filamento de actina (Figura 19).

Figura 19 - Esquema representando o BAC1 (biblioteca 4 da tabela 10) de 160 kb. Anotação do CDS (em amarelo) e mRNA (em verde) do gene completo AFAP1 que codifica a proteína associada ao filamento de actina, anotado por similaridade com o gene de *Python bivittatus*.

Embora o problema ocorrido durante o sequenciamento do read 1 tenha comprometido uma análise mais fidedigna do teste piloto de sequenciamento aleatório de BACs, este nos indicou que mesmo com problemas de sequenciamento é possível identificar genes completos no genoma.

Com o sucesso do teste piloto, nós aumentamos a escala do sequenciamento para 4608 BACs divididos em *pools* de 12, como descrito a seguir.

5.7 ESTRATÉGIA DE IDENTIFICAÇÃO IN SILICO DE GENES DE INTERESSE CLONADOS EM BACs

5.7.1 SEQUENCIAMENTO DOS POOLS DE BACs E SCREENING DE GENES

A biblioteca de BACs construída teve como objetivo permitir a prospecção de genes de interesse, sendo uma metodologia auxiliar para resolução de genes de toxinas. Usualmente a prospecção pode ser feita de duas maneiras: através de hibridização com sondas marcadas em filtros de alta densidade contendo os clones (macroarray) ou através de amplificação por PCR de pools de DNA dos clones (GREEN; OLSON, 1990). Essas técnicas exigem certo conhecimento prévio das estruturas gênicas em questão, além de equipamentos específicos (no caso do macroarray), ainda possuem um alto custo, demandam bastante tempo e geram muitos resultados falsos positivos (ainda mais se considerarmos uma lista extensa de transcritos a serem prospectados). Optamos nesse caso por desenvolver uma estratégia de *screening in silico* que pudesse substituir as formas tradicionais de seleção física de clones de BACs. Dessa forma, a estratégia que desenvolvemos envolve o sequenciamento rápido e aleatório de milhares de BACs em *pools* para então identificar aqueles com fragmentos de interesse. Isso é possível devido à capacidade de combinar múltiplas amostras e de sequenciá-las com alta cobertura, permitidas hoje pelos protocolos de preparo e utilização dos sequenciadores de nova geração, como o Illumina HiSeq. Pools de clones de BACs já foram sequenciados antes em plataforma 454 e Illumina (HAIMINEN; FELTUS; PARIDA, 2011; SATO et al., 2011; TAUDIEN et al., 2011) embora a escala, o desenho experimental e os objetivos destes tenham sido diferentes.

Para isso, realizamos as seguintes etapas de preparação:

- 1) Agrupamos 12 clones de BACs correspondendo a 12 poços (uma linha) de uma placa de estoque em um *pool*.
- 2) Repetimos dessa forma para 384 *pools* (correspondendo a 4.608 clones).
- 3) Inoculamos esses *pools* em meio líquido para crescimento e extração do DNA plasmidial (BACs).

- 4) Construímos bibliotecas individuais de sequenciamento para cada *pool* de BACs, sendo cada *pool* indexado com um único Illumina Index (uma pequena sequência nucleotídica única adicionada ao adaptador de sequenciamento).
- 5) Cada lote contendo 192 *pools* de BACs foi reunido em um único *pool* de sequenciamento e cada *pool* de sequenciamento foi aplicado em uma canaleta da célula de sequenciamento do equipamento HiSeq 1500, no modo Rapid Run, na configuração 2 X 150 bp.

Dessa forma, foi possível sequenciar, com cobertura alta, cerca de 4.608 clones (em *pools* de 12). Posteriormente, repetimos o processo para outro conjunto de 4.608 clones. Em um cálculo simples, considerando que 9.216 clones de 150 kb em média tenham sido sequenciados, temos 1.4 Gbp do genoma coberto, ou seja, cerca de 2/3 do genoma. Embora possa haver redundância entre os segmentos de DNA clonados nos BACs, conseguimos amostrar uma parte considerável do genoma utilizando essa estratégia. De maneira mais fidedigna podemos utilizar a seguinte fórmula para calcular a probabilidade que qualquer sequência de DNA única possa estar representada em uma biblioteca genômica (CLARKE; CARBON, 1976):

$$P = 1 - e^{-(N * \ln(1 - i/GS))}$$

Onde: P = Probabilidade, N = número de clones = 9216 ; i = média do tamanho dos insertos = 120 Kb (valor aproximado); GS = tamanho haplóide do genoma = 2.200 Mb = 2.200.000 kb. Seguindo esse raciocínio, a probabilidade aproximada de se encontrar determinada sequência nesse conjunto é de 40%. Obviamente, quanto maior o número de cópias de um determinado gene, maior será essa probabilidade. Essa estratégia pode ser ainda repetida e refinada em trabalhos futuros, se necessário.

As sequências de cada *pool* original foram então montadas automaticamente (software SOAP De Novo e Gapcloser) para se obter contigs de cada BAC. Cálculos como N50 não são adequados para avaliar as montagens nesse caso porque apenas uma parte do genoma está representada em cada *pool* de BAC e os fragmentos clonados podem ter tamanhos muito variáveis. Assim uma forma de se avaliar a qualidade da montagem é verificar o número de *scaffolds* longos por *pool*. Em média, cada *pool* de BAC gerou 23 +- 10 *scaffolds* maiores que 23 kb ou 51 +- 20 *scaffolds* maiores que 5Kb. Numa situação ideal na qual as montagens fossem perfeitas (cada um dos BACs resolvido em um único scaffold), cada *pool* deveria ter 12 *scaffolds* (um para cada BAC). Na prática, as repetições

fazem com que muito mais contigs sejam gerados por *pool*. Entretanto, pelo fato dos BACs conterem segmentos restritos do genoma, o software de montagem tem que lidar com uma complexidade menor e com menos repetições. Assim, esses dados foram bastante satisfatórios e importantes para as análises secundárias, conforme descreveremos.

A etapa seguinte então foi prospectar por genes de interesse *in silico*. Para isso, sequências de cDNAs e proteínas de *B. jararaca* e de outras espécies foram usadas para busca por contigs similares (usando a ferramenta Blastn no CLC Genomics Workbench). Dessa forma, foi possível identificar *pools* de BACs que contendo os genes de interesse. Em alguns casos, os contigs identificados já eram longos o suficiente para permitir o entendimento da estrutura gênica alvo. Em outros casos, onde a montagem não era boa o suficiente ou nos quais se desejava a confirmação do BAC que continha o gene de interesse, procedemos ao isolamento e ressequenciamento de BACs específicos, conforme descrito a seguir.

5.7.2 SELEÇÃO E RESSEQUENCIAMENTO DE BACs ESPECÍFICOS

Para o isolamento de BACs específicos dentro do *pool*, *primers* específicos foram desenhados baseados nas sequências dos contigs obtidos previamente, para: 1) confirmar por PCR a presença do gene de interesse no *pool*; 2) verificar qual dos 12 BACs do *pool* continha aquele gene. A figura 20 ilustra a estratégia:

Figura 20 - Esquema para identificação do clone de interesse no *pool* de BACs.

Dessa forma, procedemos ao isolamento de BACs para três tipos de toxinas conforme mostra a tabela 11.

Tabela 11 - Lista de BACs de interesse identificados por PCR, seus *pools* de origem e o gene de interesse.

Gene	Código do <i>pool</i> (#poço_#placa)	BAC identificado (#placa-#poço)
SVSP	G04_2	20-G10
	H09_2	25-H10
VEGF-F	H01_4	(2+4+6)-H08
SVMP	B06_1	9-B06
	D07_2	23-D05
	E02_3	30-E11
	E04_2	20-E09
	G07_2	23-G12
	G09_3	37-G01
	H11_2	27-H01

A Figura 21 mostra a eletroforese das reações de PCR de quatro dos 12 BACs que compõem o *pool* H09_2, positivo para um gene de SVSP. Este gel exemplifica um resultado positivo no *screening* de BACs, caracterizado pela amplificação de uma banda intensa do tamanho esperado em uma das amostras. Foi utilizado o pool como controle positivo.

Figura 21 - Exemplo de eletroforese de reações de PCR para *screening* de BACs contendo sequências de interesse. Os BACs que compõem o *pool* H09_2 foram analisados em busca do BAC contendo um gene de SVSP (Snake Venom Serine Protease). (+): controle positivo; (-): controle negativo; M: marcador de peso molecular.

Após a identificação do clone de interesse, o mesmo foi plaqueado em meio sólido para obtenção de colônias isoladas. A partir de colônias únicas, o DNA do BAC foi isolado e a presença do gene de interesse novamente confirmada por PCR (exemplo na Figura 22).

Figura 22 - Exemplo de confirmação da presença do *locus* genômico de interesse por PCR em colônias isoladas de um mesmo estoque de BAC. Os fragmentos foram amplificados por PCR com os *primers* específicos utilizados na identificação dos BACs 20G10 e 25H10. M: Marcador de tamanho.

Esse DNA foi utilizado então para o ressequenciamento de cada BAC. Inicialmente tentamos sequenciar o DNA por pirosequenciamento em sistema 454-Junior da Roche, porém o mesmo não apresentou bons resultados (dados não mostrados), provavelmente devido ao vencimento dos kits de preparação que dispúnhamos, uma vez que o sistema foi descontinuado pelo fabricante. Em seguida, realizamos sequenciamentos individuais em sistema Illumina MiSeq, na configuração 2 x 300bp, visando obter sequências mais longas que facilitassem as montagens. Embora a qualidade das leituras longas tenha ficado abaixo do esperado a partir da base 150 (dados não mostrados), o conjunto de sequências pôde ser usado para remontar os fragmentos contendo genes de interesse. O detalhamento da estrutura e demais observações a respeito dos genes de toxinas indentificados por essa abordagem serão mostrados conjuntamente com os demais resultados advindos da estratégia *shotgun*, no item 5.11, visto serem abordagens complementares para a mesma finalidade.

5.8 SEQUENCIAMENTO EM SISTEMA PACBIO

A fim de auxiliar na montagem do genoma e também para resolver *scaffolds* específicos com genes de interesse, optamos por ampliar consideravelmente a cobertura do genoma com *reads* longos gerados em sistema PacBio (Pacific Biosciences). Contratamos então o serviço de sequenciamento da *facility* “Genome Sequencing Shared Resource” da Duke University, em Durham, Carolina do Norte, EUA. Para isso, o DNA foi remetido para os Estados Unidos, após autorização de exportação junto ao IBAMA e ao CEGEN. O sequenciamento de 60 SMRT cells permitiu a obtenção de 4.685.281 *reads* com tamanho médio em torno de 15kb, conforme ilustrado na figura 23. Isso corresponde a cerca de 30X de cobertura do genoma.

Figura 23 - Perfil do sequenciamento de um lote de 23 células SMRT (de um total de 60) em sistema PacBio, mostrando a distribuição de tamanhos e qualidade dos reads.

Especificamente solicitamos à *facility* a construção de bibliotecas de fragmentos mais longos que o usual e também que o sequenciamento fosse realizado com a última química disponível do fabricante (P6-C4), a qual apresenta melhor qualidade que as anteriores. Por gerar *reads* mais longos, em teoria, essa metodologia tem potencial para resolver regiões repetitivas, uma vez que a longa leitura poderia avançar além da maioria dos *repeats* permitindo unir duas regiões não repetitivas. Porém, pelo fato dos fragmentos serem mais longos, o sequenciamento circular não chega a reler cada fragmento muitas vezes, gerando poucas leituras (*subreads*) do mesmo inserto. Isso difere do sequenciamento de fragmentos médios, que podem ser lidos várias vezes, e geram um consenso circular (CCS) que aumenta exponencialmente a certeza de cada base para até 99% de precisão. Considerando que a qualidade nativa do sequenciamento

PacBio é baixa, esse sequenciamento deve então ser combinado com *reads* Illumina para sua correção, ou ele deve ser feito com alta cobertura para que possa ser tirado um consenso dos *reads* independentes.

Para a correção das leituras-longas do PacBio baseado em leituras *paired-end*, está sendo utilizado o software PacBioToCA (KOREN et al., 2012), porém os trabalhos ainda estão em andamento para definição dos melhores parâmetros e tem exigido grande poder computacional e longo tempo de processamento. As leituras do PacBio corrigidas poderão ser utilizadas pelo software SSPACE-Long para criação de novos *scaffolds* a partir das montagens com o software AllPaths-LG.

Considerando que, em meio a esse processo, progredimos com a versão 1.0 do genoma descrita acima, optamos por lidar com essas sequências em paralelo e estamos avaliando a possibilidade de estabelecer uma colaboração para uso desses dados na próxima versão do genoma. Para a presente tese, somente alguns *reads* foram utilizados para complementar a montagem de *scaffolds* de interesse, conforme descrito nos próximos itens.

5.9 MONTAGEM DO GENOMA MITOCONDRIAL

Um dos resultados preliminares do projeto foi a montagem do genoma mitocondrial de *B. jararaca*, já que o sequenciamento *shotgun* gera naturalmente uma alta cobertura para o mesmo. A sequência foi depositada no GenBank sob número de acesso KU194299. A Figura 24 mostra um desenho esquemático com as anotações de genes e regiões importantes. A sequência completa possui 17.526 pb, tamanho semelhante ao de outras serpentes. Vale notar que, assim como todas as espécies de serpentes amostradas até o momento (com exceção dos scolecófidios *Leptotyphlops dulcis*, *Ramphotyphlops australis* e *Typhlops murius*) o genoma mitocondrial de *B. jararaca* possui uma região controladora duplicada (JIANG et al., 2007). Outra consideração interessante é a presença de uma possível duplicação do tRNA da Fenilalanina. Uma das sequências fica à jusante da Região controladora 1 e a outra fica à montante da sequência do RNA ribossomal 12S. Elas possuem uma identidade de 88% entre si.

Figura 24 - Desenho esquemático mostrando as anotações na sequência de DNA mitocondrial de *B. jararaca*.

5.10 RNA RIBOSSOMAL

Além da mitocôndria, investigamos especificamente a montagem do precursor dos RNAs ribossomais de *B. jararaca* (rRNA45S). Utilizamos sequências genômicas parciais (disponíveis no GenBank) de rRNA das serpentes *Ophiophagus hannah* e *Crotalus mitchelli* como referência para montagem. Sabe-se que o precursor genômico rRNA45S é processado, dando origem aos rRNA18S, 5-8S e 28S, os quais desempenham um importante papel na estrutura e função do ribossomo. Essas sequências são bem conservadas entre as diversas espécies de eucariotos, entretanto, ainda possuem certas variações que podem ser importantes. Surpreendentemente, pouquíssima informação sobre o rRNA 45S completo de animais não-mamíferos está disponível na literatura. Dessa forma, para anotação das sequências funcionais, nós utilizamos o rRNA45S humano como referência, já que está bem anotado e com as sequências genômicas bem estabelecidas. A figura 25 mostra um alinhamento entre o contig de *B. jararaca* gerado (9.079 pb) e a sequência 45S humana, a conservação entre elas é mostrada em forma de barras na cor cinza. Percebe-se que as regiões 18S e 5-8S são bastante conservadas, já a

região 28S apresenta-se mais variável, embora ainda mostrando certo grau de conservação (73% de identidade). Essa variação se deve principalmente à presença de gaps ou deleções no rRNA de *B. jararaca* em relação à sequência de rRNA45S humana. As regiões ITS (*Internal Transcribed Sequences*) que se situam entre as sequências 18S e 5-8S, e entre 5-8S e 28S, apresentam baixo grau de conservação como esperado.

Figura 25 - Alinhamento entre o precursor rRNA45S de *B. jararaca* e o humano. As setas vermelhas (*B. jararaca*) e azuis (*Homo sapiens*) são as anotações dos rRNA18S, 5-8S e 28S. A conservação é mostrada em forma de barras na cor cinza.

5.11 BUSCA E ANÁLISE DOS GENES DE TOXINAS

5.11.1. GENES DE TOXINAS RECUPERADOS

De posse do conjunto de *scaffolds* que compõem a versão 1.0 do genoma e do conjunto de contigs derivados dos BACs (em *pools* e individuais) iniciamos as buscas por genes de toxinas. Dessa forma, utilizamos como sondas *in silico* o conjunto de transcritos referenciais que foram definidos anteriormente como representativos das toxinas da espécie (JUNQUEIRA-DE-AZEVEDO et al., 2015). Por meio de buscas via Blast local, foram então identificados contigs genômicos com identidade significativa e extensa com segmentos dos transcritos de cada tipo de toxina. Cada *output* de alinhamento foi interpretado individualmente, para considerar o real encontro de um provável gene. Posteriormente, com a anotação do genoma, outras informações foram adicionadas e permitiram refinar a identificação e o entendimento das estruturas gênicas, como a adição da cobertura por *reads* transcriptômicos dos vários tecidos, a predição gênica pelos diferentes métodos (Augustus, PASA) e a similaridade dos produtos derivados com proteínas depositadas nos bancos de dados. Dessa forma, no presente momento reconhecemos a identificação de genes (parciais ou inteiros) para a maioria das classes de toxinas, conforme apresentado na Tabela 12. Para a maioria dessas classes há ao

menos um gene completamente resolvido, portanto permitindo a dedução do arquétipo de organização gênica de praticamente todas as toxinas conhecidas do gênero *Bothrops*.

Tabela 12 - Lista de potenciais componentes de venenos de serpentes que tiveram ao menos uma estrutura gênica recuperada no genoma de *B. jararaca* por meio das diferentes estratégias desse trabalho.

Família protéica	BACs	Shotgun	PacBio	Completo/ Parcial
5' nucleotidase		x		C
Peptídeo potenciador de bradicinina (BPP)	-	-	x	C
CRISP		x		C
C-type Lectin (CTL)		x	x	P
Complement factor 3 (similar to CVF)	x	x		P
Cystatin	x	x		C
Diesterase		x		C
Dipeptidyl peptidase		x		C
Hyaluronidase	x	x		C
L-aminoácido oxidase (LAAO)	x	x	x	P
NGF	x	x		C
Fosfolipase A2		x	x	P
Fosfolipase B	x	x		C
VEGF-F (svVEGF)	x			C
VEGF-A		x		C
Waprin-like		x		P
Metalloproteinase (SVMP) P-III	x	x		C
Metalloproteinase (SVMP) P-II	x	x	x	C
Metalloproteinase (SVMP) P-I				-
Serino proteinase (SVSP)	x	x	x	C
Inibidores enzimáticos de glândula:				
Gamma-PLA2	x	x		C
Alpha-PLA2		x		C
BJ46 (inibidor de SVMP)		x		P

As análises transcricionais das glândulas de veneno de várias espécies indicaram que algumas das classes de toxinas deveriam possuir múltiplos parálogos, enquanto que outras deveriam possuir genes únicos. Entretanto, poucas eram as evidências dessa paralogia ao nível genômico, e menos ainda propuseram como estariam distribuídos esses parálogos ao longo do genoma (por exemplo, se *in tandem* ou dispersos). De fato, nossos dados indicaram que algumas classes como SVMPs e SVSPs apresentam ampla paralogia, enquanto que outras parecem ter cópia única (ex: VEGF-F), conforme será detalhado adiante. Entretanto, o conjunto genômico que obtivemos não permite precisar

exatamente quantos parálogos existem para cada classe de toxina com múltiplos parálogos. Isso se deve principalmente ao fato de termos muitos genes apenas parcialmente resolvidos, de forma que fragmentos gênicos em diferentes contigs podem corresponder ao mesmo gene ou a genes diferentes. Não obstante essa limitação, diversas informações relevantes a respeito do arranjo de alguns parálogos e de seu contexto genômico puderam ser deduzidas, como por exemplo, a confirmação de que as duas classes de proteases que estão entre os componentes principais do veneno do gênero *Bothrops* possuem múltiplos parálogos arranjados *in tandem*. Os principais achados são descritos a seguir.

5.11.2 ORGANIZAÇÃO GENÔMICA DE SERINO PROTEINASES DE VENENO

Uma das classes mais importantes de toxinas de *Bothrops jararaca* corresponde a enzimas do tipo serino proteinases (SVSPs). O mecanismo tóxico das SVSPs está relacionado à ativação de fatores da coagulação, podendo atuar também nos sistemas fibrinolíticos, calicreína-cinina e em plaquetas. Essas enzimas são bastante comuns em venenos de serpentes viperídeas, sendo frequentemente chamadas de enzimas tipo calicreína, devido à suposta relação evolutiva com estas enzimas. Entre os *scaffolds* de jararaca, localizamos dois principais contendo múltiplas cópias de genes de SVSPs (figuras 26 e 27). A mesma região de um deles (Figura 27) também foi encontrada em um dos BACs da biblioteca. É interessante notar que a maior parte desses genes se encontra flanqueada por um gene parcial da subunidade 6B1 da enzima citocromo c oxidase, indicando que as duplicações devem ter envolvido um segmento genômico compreendendo os dois genes.

Curiosamente, uma das poucas estruturas gênicas de toxinas ofídicas conhecidas, ainda na década de 1980, foi a da Batroxobina, uma SVSP de *Bothrops atrox* (ITOH et al., 1988). Justamente pela análise da organização intrônica desse gene, é que foi proposta sua origem comum com as calicreínas e tripsinas. O arranjo dos genes ora identificados segue o mesmo padrão, incluindo o longo éxon que codifica a região 3'UTR do mRNA, no entanto, alguns dos genes encontrados parecem possuir um éxon a mais na região 5'UTR.

Figura 26 - Diagrama ilustrando o scaffold L2_002391, com tamanho total de 150 Kbp, contendo oito genes para SVMPS, sendo quatro completos (setas azuis) e flanqueados por pseudogenes *cox6b1*. O código BJAR_SVSPisotigXXXX refere-se ao transcrito referencial cuja sequência foi mapeada nos éxons de cada gene, com identidade maior que 96%. No detalhe, observa-se um dos genes com sua estrutura de íntrons e éxons e, sob este, um gráfico de cobertura de *reads* do RNAseq da glândula de veneno, que mapeiam majoritariamente nos éxons.

5.11.3 CONTEXTO GENÔMICO DAS SVSPs

Apesar de termos identificado que os genes das SVSPs estão dispostos *in tandem* no genoma, não foi possível determinar somente com o nossos dados o contexto genômico no qual tais genes estão inseridos, ou seja, sua localização relativa a outros genes. No entanto, após o fechamento da versão 1.0 do genoma da jararaca, foi disponibilizado no NCBI o genoma do viperídeo *Protobothrops mucrosquamatus*. Em um dos scaffolds dessa montagem foi possível observar genes de SVSPs que ocorrem entre o gene “SCAN domain containing protein 1-like” (o qual está próximo do gene PSMC4) e o gene “RBM42”. Como não havia na nossa montagem scaffolds correspondentes a essa região, recorremos aos reads longos do PACBIO e foi possível identificar que existe ao menos um gene de SVSP adjacente ao gene “RBM42” e também ao menos um outro gene de SVSP adjacente ao “SCAN domain containing protein 1-like”. A baixa qualidade dos reads PACBIO dificulta a identificação precisa de genes com múltiplos parálogos, como é o caso das SVSPs, porém podemos dizer que provavelmente na *B. jararaca*, assim como em *P. mucrosquamatus*, o cluster de SVSPs ocorre entre os genes “RBM42” e “SCAN domain containing protein 1-like” (Figura 28). Não sabemos ainda se existem outros genes de SVSPs em outros *loci*. Essa organização genômica também está presente em outras serpentes, como *Thamnophis sirtalis* (que possui apenas um gene de SVSP na região), além de outros répteis como o lagarto *Anolis carolinensis* (o qual pertence ao clado Toxicofera) e o *Gekko japonicus* (não pertencente ao clado Toxicofera). Surpreendentemente, encontramos uma organização similar no genoma da ave *Falco peregrinus*, com um cluster de genes kallikrein-like e trypsin-like (tais genes também codificam serino-proteases com a mesma estrutura gênica das SVSPs) próximo aos genes RBM42 e PSMC4 (Figura 28). Porém, diferentemente dos genes de répteis escamados, as sequências nesse contexto não parecem ser monofiléticas em relação aos genes de toxinas (dados não mostrados), portanto sua possível relação evolutiva com genes de toxinas ainda não é clara.

PACBIO. Abaixo os *scaffolds*: NW_0015386730 (*P. mucrosquamatus*), NW_013658697 (*T. sirtalis*), NW_003339061 (*A. carolinensis*), NW_015173163 (*G. japonicus*), NW_004930070 (*F. peregrinus*).

5.11.4 A ESTRUTURA GÊNICA DE UMA SVMP DO TIPO P-III

Sanz et al., em 2012, publicaram a sequência de um gene de SVMP do tipo P-III de *Echis ocellatus*. Os autores observaram que a organização de íntrons desse gene diferia daquela observada para o gene da ADAM28, na medida em que o gene sequenciado não apresentava 5 dos 16 íntrons presentes no gene da ADAM28, trazendo a dúvida se haveria outro gene ancestral mais próximo. Prospectando nossas sequências genômicas, pudemos identificar alguns contigs contendo genes de SVMPs do tipo P-III. Particularmente os *scaffolds* L6_005584 e L8_00079 revelaram genes parciais de SVMPs que puderam ser associados a diferentes transcritos referenciais (BJARREF_SVMP03 e BJARREF_SVMP04, respectivamente) que identificamos anteriormente para a espécie (JUNQUEIRA-DE-AZEVEDO et al., 2015). Embora parciais, duas porções desses genes mostraram a existência de íntrons em posições conservadas com íntrons da ADAM28 de vertebrados e que estavam ausentes no gene de P-III de *Echis ocellatus*.

Procurando entre os BACs, identificamos nos BACs 23-G12 e 23-D05 um gene completo codificando para uma PIII muito semelhante à berythtractivase (SILVA et al., 2003) e expresso nas glândulas de veneno (figura 29). Esse gene também apresentou o número e a posição de introns semelhante ao gene de ADAM28. Outras evidências da relação desse gene com as SVMPs serão descritas adiante e reforçam a ideia de uma origem comum entre eles.

Figura 29 - Diagrama ilustrando a posição dos íntrons de um gene de SVMP do tipo P-III encontrado em um dos BACs, em comparação com o gene previamente publicado de uma SVMP de *E. ocellatus* (Sanz et al., 2012). As setas pretas indicam íntrons em posições conservadas entre as SVMPs de *B. jararaca*, de *E. ocellatus* e no gene da ADAM28, enquanto que setas vermelhas indicam íntrons presentes nos genes das SVMPs de *B. jararaca* e na ADAM28, mas ausentes do gene de *E. ocellatus*.

5.11.5 A ESTRUTURA GÊNICA DE UMA SVMP DO TIPO P-II

Prospectamos também genes de SVMPs do tipo P-II, sendo que pudemos identificar alguns especialmente nos *pools* de BACs. O *pool* de BACs G7_2 e o *pool* D7_2 mostraram conter um gene para a bothrostatina, uma conhecida metaloproteínase do tipo P-II (FERNANDEZ et al., 2005). A partir do *screening* por PCR dos clones que compunham esses dois *pools*, pudemos isolar os clones BAC_23G12 e BAC_23D5, respectivamente do *pool* G7_2 e D7_2. O ressequenciamento dos mesmos confirmou a presença do gene completo da SVMP do tipo P-II, idêntica à bothrostatina, mas também revelou a presença, nos mesmos BACs, de um gene completo de uma SVMP do tipo P-III, semelhante à berythactivase (mostrado no item anterior). Isso indicou, portanto, que ambos os tipos, P-II e P-III, se avizinham *in tandem* no mesmo locus. O gene da bothrostatina foi completamente resolvido e está apresentado na figura 30A. Este é o segundo gene descrito de uma metaloproteínase do tipo P-II e ele revela uma organização de íntrons e éxons semelhante ao gene da ADAM28 e também ao gene do tipo P-III que identificamos (Figura 31).

É também interessante notar que há uma alta cobertura de *reads* de RNAseq da glândula de veneno mapeando nos éxons do gene da bothrostatina, conforme esperado para um gene altamente expresso e específico do tecido e em menor proporção no gene

B)

Figura 30 - Diagrama ilustrando as sequências parciais do BAC 23G12 contendo dois genes de SVMPs, sendo um deles codificante para a enzima bothrostatina, uma P-II (A) e o outro a berythactivase-like, uma P-III (B). Como ambos os genes estão localizados no mesmo BAC, supomos que ocorrem *in tandem*. No destaque, os genes têm sua estrutura esquematizada, mostrando sobre ele os diferentes elementos repetitivos preditos (pelo software *repeat masker*) e, sob o mesmo, a cobertura de leituras do RNAseq da glândula de veneno e de todos os outros tecidos.

Figura 31 - Comparação entre os genes da ADAM28, Berythactivase-like (PIII) e Bothrostatina (PII) de *B. jararaca*, enfatizando a conservação de éxons. A estrutura gênica sugere que a ADAM28 perdeu os últimos 8 éxons, dando origem às SVMPs do tipo PIII, as quais posteriormente sofreram uma deleção interna que deu origem às SVMPs do tipo PII. As linhas verdes representam a conservação da estrutura éxon-íntron entre a ADAM28 e SVMPs. As linhas vermelhas indicam a diferença entre as estruturas dos genes de P-III e P-II.

Figura 32 - Mapeamento dos reads do transcriptoma das glândulas de veneno obtidos nesse trabalho em sequências de cDNA de referência de metaloproteínas do tipo P-I de *B. jararaca* (Junqueira-de-Azevedo, 2015). A figura mostra o final da região codificante (sob a seta amarela) para os transcritos BJAR_SVMP_17ed (A) e BJAR_SVMP_19ed (B) com destaque para o códon de parada. Os reads mapeados nessa região não apresentam um códon de parada, além de outras diferenças observadas, portanto devem ser originados de mRNAs codificantes para SVMPs de outros tipos, porém com sequência similar o suficiente para permitir o mapeamento.

5.11.7 O CONTEXTO GENÔMICO DAS METALOPROTEINASES

Outra questão de interesse em relação às SVMPs é saber a distribuição dos parálogos no genoma. A análise do genoma de *Ophiophagus hannah*, um elapídeo de veneno predominantemente neurotóxico mas que também contém SVMPs em menor quantidade e menor grau de diversificação, mostrou fragmentos de dois possíveis genes de SVMPs do tipo P-III em um mesmo scaffold (VONK et al., 2013), como mostra a figura 33. O caráter fragmentado daquele genoma, entretanto, não permitiu contextualizar o *locus* de ocorrência desses genes, embora tenha sido observada a presença de um gene semelhante à ADAM28 à montante das SVMPs. Na verdade, a própria anotação deste *locus* contendo a ADAM28 de *O. hannah* não deixava claro se eram genes de reais SVMPs, pois haviam alguns erros de anotação causando a combinação de éxons de genes diferentes. Mais recentemente, encontramos depositado no GenBank um *scaffold* genômico do viperídeo *Protobothrops mucrosquamatus* (acession NW_015389796), ainda não publicado, que indicou a presença de quatro cópias de genes de SVMPs adjacentes a um gene de uma SVMP-like que possivelmente corresponderia a ADAM28 daquele organismo, embora não fosse anotada como tal (figura 33). Mas também nesse caso, o *scaffold* incompleto não permitia definir o *locus* desse conjunto e, portanto confirmar se o gene era o da ADAM28. Nesse contexto, buscamos então por um gene que

pudesse ser a ADAM28 no genoma de *B. jararaca* e o localizamos no scaffold L8_0000079 que apresentava um gene semelhante a ADAM28 seguido por ao menos duas cópias de SVMPs (figura 33). O segundo gene, embora incompleto, pôde inclusive ser associado a um dos nossos transcritos referenciais (BJARREF_SVMP04), indicando, portanto, ser um gene funcional. Mais do que isso, por ser bastante longo, o scaffold de *B. jararaca* permitiu uma melhor contextualização genômica na medida em que revelou a presença dos genes *stc1*, *NKX3-1* e *slc25a37* à montante do gene da ADAM28, além de outros genes relacionados a esse locus (não mostrados na figura 33 para simplificação). Com isso foi possível confirmar se tratar do ortólogo da ADAM28 de outros vertebrados, uma vez que em vários outros organismos este gene é também flanqueado pelo mesmo conjunto gênico (vide figura 33). É interessante notar que os genes ADAMDEC1 e ADAM7, ao contrário do que ocorre em mamíferos, estão ausentes em répteis. Já em *Xenopus tropicalis* (um anfíbio), no mesmo contexto da ADAM28 há um gene anotado como ADAM7, além de outros genes relacionados às ADAMs. Dessa forma, temos agora não apenas evidências de que as SVMPs ocorrem *in tandem* no genoma, como também de que estes parálogos estão localizados à jusante do seu suposto gene ancestral.

Figura 33 - Esquema comparativo do *locus* genômico que contém o gene ADAM28 (caixa vermelha) em diversos organismos. Em *B. jararaca*, o mesmo é antecedido pelos genes *stc1*, *nkx3* e *slc25a37* (caixa verde), tal como em outros vertebrados. Genes de SVMPs (serpentes) e ADAMs de *H. sapiens* e *X. tropicalis* estão representados por uma caixa azul. Genes de ADAMs à jusante da ADAM28 estão ausentes em aves e répteis da ordem Squamata. Os genes não relacionados às ADAMs à jusante da ADAM28 em diversos organismos estão destacados por uma caixa laranja.

5.11.8 POSSÍVEL PSEUDOGENE E GENES TRUNCADOS

Além dos genes de SVMs que estão expressos nas glândulas do indivíduo amostrado, ou seja, com reads do RNAseq mapeando uniformemente nos éxons, há também sequências semelhantes aos genes de SVMs que aparentemente não são expressas neste indivíduo, isto é, em que os *reads* derivados do RNAseq não são mapeados com 100% de identidade, ou poucos mapeiam. Entre estes genes está um

possível pseudogene, o qual se assemelha a um retrogene (não tem íntrons), que é interrompido por uma sequência análoga a um microssatélite de *Agkistrodon contortrix* (CASTOE et al., 2010) (Figura 34) e cujas extremidades se assemelham ao elemento repetitivo OposCharlie3 da família de transposons hAT. Tal possível pseudogene parece também estar presente em outras espécies da família Viperidae, uma vez que a sequência inteira apresenta identidade de 96% com outro possível pseudogene que identificamos no genoma de *Protobothrops mucrosquamatus*, e também identidades de 90% com GP-I (AF378673), com GP-III (AF378675) e com GP-IV (AF378676) e de 80% com GP-II (AF378674), sendo todas estas sequências amplificadas por PCR a partir do DNA genômico de *Crotalus scutulatus scutulatus* (DAGDA et al., 2013), as quais foram anotadas como genes de SVMPs, cuja estrutura não se conhecia até então.

Figura 34 - A) Comparação entre o gene de uma SVMP-PIII e pseudogenes. Os diferentes domínios são identificados por diferentes cores. As interrupções no CDS estão mostradas por riscos vermelhos (íntrons e inserção do elemento repetitivo). A região hachurada representa uma sequência não funcional. B) Desenho esquemático das sequências de pseudogenes (região similar ao CDS em azul) de *B. jararaca*, *P. mucrosquamatus* (NW_015386790) (parte de cima) e de regiões amplificadas de *C. scutulatus scutulatus* GP-I (AF378673), GP-II (AF378674), GP-III (AF378675), GP-IV (AF378676) (parte de baixo).

Outro achado inesperado foi observado em um contig do BAC 23D05, que parece conter um gene de SVMP truncado no domínio metaloprotease. Após o truncamento há a inserção de outro gene de SVMP. Não é possível saber se o gene truncado possui as

regiões do peptídeo sinal e prodomínio completas, pois o segmento genômico clonado inicia-se no fim do prodomínio (Figura 35), mas é clara sua interrupção após o domínio metaloprotease. Também não temos certeza se o gene seguinte está completo, pois o *contig* termina após o segundo éxon da região do domínio metaloprotease. A mesma região foi encontrada em reads PACBIO confirmando o truncamento de um gene de SVMP.

Figura 35 - Extremidade 5' terminal do BAC 23D05. Um gene de SVMP é interrompido no domínio metaloprotease por outro gene de SVMP.

5.11.9 MONTAGEM DO GENE DO PRECURSOR DOS BPPs/CNP A PARTIR DE READS PACBIO

Os peptídeos potenciadores de bradicinina (BBPs) são importantes agentes hipotensores do veneno de jararaca e de outros viperídeos, sendo conhecidos por terem servido de modelo para a síntese de fármacos anti-hipertensivos de amplo uso, como o Captopril (CUSHMAN et al., 1982). Seu precursor, descrito por MURAYAMA (1997), apresenta uma estrutura bem peculiar pois codifica regiões repetitivas que correspondem aos BPPs, mas também possuem um *linker* espaçador que antecede uma região que codifica o peptídeo natriurético do tipo C (CNP). Nossas buscas iniciais pelo gene correspondente a esse precursor nos *scaffolds* genômicos e nos BACs não resultou na identificação de nenhum segmento gênico, foram encontradas apenas regiões do 3'UTR. Verificamos que o cDNA desse precursor possui um alto conteúdo de GC, especialmente pelo fato da proteína ser rica em prolinas que são codificadas por códons CCX. Além disso, a sequência dos genes de CNPs em outros organismos também parece possuir alto conteúdo GC. Esse alto teor de bases GC é conhecido por representar um obstáculo ao sequenciamento na plataforma Illumina e consequentemente ao processo de montagem (CHEN et al., 2013). Deduzimos então que tais sequências estivessem suprimidas de nossos sequenciamentos nessa plataforma e consequentemente das montagens (tanto *shotgun*, quanto dos BACs). Assim, procuramos manualmente por

reads PacBio, menos sensíveis ao alto teor de GC, e de fato encontramos alguns que alinhavam com o cDNA dos BPPs. Um destes *reads* contém um gene inteiro do precursor dos BPPs. Como mencionado anteriormente, *reads* brutos PacBio possuem alta taxa de erros, então realizamos a correção dos mesmos baseada em *reads* advindos do RNAseq de tal forma que conseguimos resolver a sequência genômica do gene dos BPPs (Figura 37).

O gene possui cerca de 11kb e dois íntrons, sendo apenas um internamente ao CDS. Esse é um resultado bastante interessante e relevante, pois mostra que a região codificante para o precursor está dividida em dois éxons: o primeiro contendo o peptídeo sinal e a região dos BPPs e o segundo contendo o espaçador e o CNP. Em outros organismos os genes de CNPs também apresentam dois íntrons na região codificante. Tal achado suporta uma hipótese levantada anteriormente por nosso grupo, de que o precursor dos BBPs/CNPs de viperídeos e o precursor dos CNPs de elapídeos e colubrídeos eram ortólogos, dada a conservação de alguns segmentos. Ambos os precursores seriam evolutivamente derivados de um precursor de CNP de vertebrados, sendo que nos Viperídeos houve a aquisição de uma região codificadora para BPPs (CHING et al., 2006; JACKSON et al., 2013). Quando comparamos os genes de CNP/BPP de viperídeos com os de outros organismos fica claro que o primeiro éxon em Viperídeos é mais longo do que os outros (Figura 37). Este éxon em *Anolis carolinensis*, *Gekko japonicus* e *Homo sapiens* codifica apenas o peptídeo sinal, enquanto que, em *B. jararaca* e *P. mucrosquamatus*, além do peptídeo sinal existem sequências de BPPs. É interessante notar que em *B. jararaca* há ainda mais BPPs do que em *P. mucrosquamatus*.

Para confirmar ainda se o problema em relação à obtenção do gene pelo *shotgun* era por conta mesmo da ausência de *reads* ou por alguma dificuldade dos montadores, remapeamos os *reads* Illumina no gene montado. Como pode ser observado na figura 36, *reads* foram mapeados apenas nas regiões intrônicas e 3'UTR, indicando que o alto teor de GC dos éxons realmente deve dificultar a leitura ou a amplificação dessas regiões.

Figura 36 - Diagrama ilustrando o gene do precursor dos BPPs, abaixo o mapeamento de *reads* genômicos curtos (Illumina) que ocorre principalmente nas regiões intrônicas. A região codificante (CDS) compreende dois éxons, representados pelas setas amarelas.

Figura 37 - Alinhamento dos genes de CNPs e CNPs/BPPs (feito com o programa ClustalO) de *Homo sapiens* (NC_000002), *Gekko japonicus* (NW_015164248), *Anolis carolinensis* (NW_003338851), *Protobothrops mucrosquamatus* (NW_015388670) e *Bothrops jararaca*. Em amarelo estão indicados os éxons da região codificante. No detalhe está a sequência resolvida do gene, com destaque para a tradução do precursor, incluindo o peptídeo sinal (vermelho), os BPPs ativos (azul), a região espaçadora (verde) e o peptídeo natriurético do tipo C (CNP) (roxo).

5.11.10 GENE DO VEGF-F (svVEGF) E UMA NOVA HIPÓTESE DE SUA ORIGEM

Em 2002 foi descrita a existência de uma toxina em venenos de serpente semelhante ao vascular endotelial growth fator (svVEGF) (JUNQUEIRA DE AZEVEDO et al., 2001). Posteriormente, com a identificação de ortólogos em vários gêneros de viperídeos, os svVEGFs foram agrupados na classe F de VEGFs. Por possuir certa similaridade com o VEGF-A, supunha-se que o VEGF-F poderia ser um gene parálogo do VEGF-A. O VEGF-A é uma proteína bem conservada de vertebrados, conhecida por atuar na angiogênese, vasculogênese e crescimento de células endoteliais (MURPHY; FITZGERALD, 2001). Ou seja, as serpentes viperídeas possuem o VEGF-A, conservado entre os vertebrados e que atua endogenamente, e também o VEGF-F, cuja expressão é exclusiva da glândula de veneno e atua no envenenamento. O VEGF-A humano possui seis variantes de splicing, sendo que em *P. flavoviridis* foram encontradas três variantes que são fracamente expressas nas glândulas de veneno TfVEGF122, TfVEGF166 e TfVEGF190 (TAKAHASHI et al., 2004). Ao menos uma variante de splicing sem o éxon seis (VEGF166) parece ser expressa nas glândulas de veneno de *B. jararaca* (JUNQUEIRA-DE-AZEVEDO et al., 2015). A estrutura gênica de VEGFs-F de serpentes asiáticas foi elucidada e os autores hipotetizaram que esse gene tivesse se originado por duplicação do gene do VEGF-A (YAMAZAKI et al., 2009). A evidência para isso foi a suposta conservação de uma pequena região dentro desses íntrons. A figura 38-B, retirada desse trabalho, ilustra a hipótese.

Obtivemos aqui sequências completas dos genes do VEGF-A (28Kb) e do VEGF-F (5 Kb) de *B. jararaca* (Figura 38-A). O gene do VEGF-A é claramente um ortólogo do VEGF-A descrito em outras espécies de vertebrados, incluindo *Homo sapiens*, pois além de ter sequência conservada foi encontrado no mesmo contexto genômico. Investigamos então a relação do gene do VEGF-A com o do VEGF-F. Numa comparação direta entre as sequências intrônicas dos dois genes, nos chamou a atenção que além de terem tamanhos completamente diferentes, não possuíam qualquer conservação entre as sequências, gerando alguma dúvida sobre a relação evolutiva entre eles. Assim, fomos investigar o contexto genômico do VEGF-F para comparar seu locus com o de outras espécies. Usando como referência um segmento genômico de *P. mucrosquamatus* contendo esse gene (NW_015387372) conseguimos resolver parcialmente a sequência de um dos BACs de *B. jararaca* que continha o gene do VEGF-F (Figura 39). O locus de *B.*

jararaca contém além do gene VEGF-F, um conjunto específico de genes flanquadores (MARK4, CKM, KLC3, ERCC2, PPP1R13L).

Recuperamos então o locus que contém este mesmo conjunto de genes em outros animais. Para o viperídeo *P. mucrosquamatus*, há grande sintonia, estando o VEGF-F em meio aos mesmos genes. Para a serpente mais basal *Python molurus*, neste segmento há um gene anotado como VEGF-F-like, mas cuja sequência é mais assemelhada ao PGF-like (Placental Growth Factor-like) de outros organismos. Em diversos outros organismos, é esse gene PGF-like que encontra-se flanqueado pelo conjunto de genes que circundam o VEGF-F de viperídeos. É interessante observar que o produto do gene PGF-like (*placental growth factor-like*) pertence à mesma família dos fatores de crescimento endoteliais VEGFs, PGFs e PDGFs (*platelet derived growth factor*).

Dessa forma, o contexto genômico nos indicou que o VEGF-F na realidade pode ter derivado deste outro gene (PGF-like) da família dos fatores de crescimento vascular e não do VEGF-A, hipotetizado anteriormente. Nesse caso, o processo evolutivo que levou ao surgimento do VEGF-F envolveria a simples especialização gênica por seleção positiva ao invés do clássico modelo de neofuncionalização pós-duplicação gênica, comumente observado para toxinas.

Ainda encontramos em um dos pools de BACs e na montagem do ALLPATHs LG um retrogene do VEGF-F, aparentemente trata-se de um pseudogene, pois há vários códons de parada na região equivalente à codificante, além de não ser um gene expresso nas glândulas de veneno.

Figura 38 - A) Diagrama ilustrando os genes do VEGF-A e VEGF-F de *B. jararaca*. B) Diagrama ilustrando os mesmos genes em *Protobothrops (Trimeresurus) flavoviridis* (Yamazaki et al., 2009).

Figura 39 - Esquema comparativo do *locus* genômico que contém o gene VEGF-F em diversos organismos. Em *B. jararaca*, um BAC isolado contendo o gene VEGF-F foi sequenciado. A primeira figura mostra um contig correspondente a esse locus e reads do BAC cobrindo o mesmo. As caixas correspondem aos genes flanqueadores conservados à montante (verdes) e à jusante (marrons) dos genes VEGF-F e PGF-like (circulados em vermelho). Os números de acesso dos scaffolds utilizados para comparação são: *Protobothrops mucrosquamatus* (NW_015387372), *Python molurus* (NW_006533531), *Chrysemys picta* (NW_007281958) e *Latimeria chalumnae* (NW_005819112).

5.11.11 GENE DE FOSFOLIPASE B

As fosfolipases B são expressas em baixos níveis nas glândulas de veneno de *B. jararaca*. Na montagem do genoma com o ALLPATHs LG e nos *pools* de BACs foram encontrados contigs e scaffolds relacionados à Fosfolipase B de *B. jararaca*. A região codificante encontrada é idêntica àquela encontrada no transcriptoma. O gene da fosfolipase B de veneno ocorre entre os genes “GUCY2C” e “ATF7IP” (scaffold l1scaffold_1202, Figura 40), e essa posição parece ser conservada em mamíferos, aves e répteis escamados para a PLBD1(Phospholipase B domain containing 1), o que sugere que a mesma foi recrutada para o veneno em serpentes.

Figura 40 - Organização genômica do gene da Fosfolipase B de *B. jararaca*. No detalhe está mostrada a estrutura éxon-ínton deste gene. Abaixo está o locus equivalente em *Anolis carolinensis* (NW_003339261).

5.11.12 GENE DE HIALURONIDASE

O gene completo da hialuronidase de *B. jararaca* foi encontrado em um dos *pools* de BACs e também na montagem do ALLPATHs LG (Figura 41). Este gene está flanqueado pelos genes “WASL” e “Hialuronidase 4-like”, tal configuração também foi encontrada em *Protobothrops mucrosquamatus*, *Anolis carolinensis* e até em alguns mamíferos, como *Mus musculus* (Figura 41). Dessa forma, a Hialuronidase de veneno parece ter uma origem comum com a Hyal6 ou Hyalp1. Em *Anolis carolinensis*, existe uma duplicação das hialuronidasas venom-like, além da Hialuronidase 4-like.

Bothrops jararaca

Figura 41 - Organização genômica do gene da Hialuronidase de *B. jararaca*. No detalhe está mostrada a estrutura éxon-ítron deste gene. Abaixo estão scaffolds de *P. mucrosquamatus* (NW_015386248), *A. carolinensis* (NC_014780) e *Mus musculus* (NC_000072) para comparação. As caixas vermelhas mostram genes Hyalp1 e Hyal4 em outros organismos.

5.11.13 GENE DE CRISP

A sequência completa do gene de uma CRISP foi encontrada em um *scaffold* gerado na montagem feita com o software ALLPATHs LG. O gene contém 8 éxons e 7 íntrons, no entanto, somente 7 éxons compõem a sequência codificante. Este se situa entre os genes “fam161a” e “Serotriflin-like”. É interessante notar que em *Anolis carolinensis*, a configuração é semelhante, havendo um gene “venom CRISP-like” próximo aos genes “fam161a” e “Serotriflin-like”, no entanto o gene “venom CRISP-like” está entre dois Serotriflin-like (Figura 42). A Serotriflina também é uma proteína da família das CRISPs, tendo sido encontrada no soro de *Trimeresurus flavoviridis*. A mesma se liga a outras proteínas séricas chamadas de SSPs (Small Serum Proteins)(AOKI et al., 2008). Aparentemente, em vertebrados “não-squamata” não há genes CRISP-like na região do gene “fam161a”(Figura 42). Entretanto, no ganso *Anser cygnoides domesticus* e na tartaruga *Pelodiscus sinensis* existe um locus com genes svCRISP-like e serotriflin-like. Tal locus está no mesmo contexto genômico que as CRISPs-1, -2 e -3 de mamíferos, as quais estão envolvidas com o funcionamento do sistema reprodutivo. As CRISPs de serpentes, aves, mamíferos e répteis escamados, além de serem homólogas, possuem uma estrutura gênica semelhante, sugerindo uma origem evolutiva comum.

Figura 42 - Organização e contexto genômico do gene da CRISP de *B. jararaca*. Abaixo estão mostradas regiões genômicas de *Anolis carolinensis* (NC_014776), *Anser cygnoides domesticus* (NW_013185717), *Pelodiscus sinensis* (NW_005859216) e *Homo sapiens* (NC_000006) contendo genes da família CRISP.

5.11.14 OUTROS GENES DE TOXINAS E COMPONENTES DO VENENO

Outros genes de toxinas importantes como os das Fosfolipases A2 (PLA2) e Lectinas foram encontrados apenas parcialmente na versão 1.0 do ALLPATHs-LG. No entanto, encontramos sequências semelhantes às PLA2 e Lectinas nos reads PACBIO. As

PLA2 possuem uma estrutura gênica com 4 éxons e 3 íntrons, confirmando o que já se sabia sobre tais estruturas (IKEDA et al., 2010; YAMAGUCHI et al., 2014). As Lectinas são um grupo diverso de toxinas de *B. jararaca*, sendo possível que apresentem estruturas gênicas diferentes entre si. Entretanto, foi possível identificar 5 éxons e 4 íntrons em várias das sequências de referência (JUNQUEIRA-DE-AZEVEDO et al., 2015). Ainda dentro das sequências de *pools* de BACs encontramos o gene do componente de veneno *Cobra Venom Factor* (CVF). Trata-se de um gene muito extenso e embora esteja bastante fragmentado foi possível identificar todos os éxons (40 no total). O CVF é um componente do veneno típico de elapídeos que não é expresso nas glândulas de veneno de *B. jararaca*, mas é expresso no fígado (JUNQUEIRA-DE-AZEVEDO et al., 2015). Outro gene encontrado e que é pouco expresso nas glândulas de veneno de viperídeos foi o do NGF (Nerve Growth Factor). Este se situa ao lado do gene TSPAN2 e “mitochondrial glutamate carrier-1”. Uma peculiaridade desse gene é que o mesmo não possui íntrons na sequência codificante, embora haja dois íntrons na região 5'UTR (dados não mostrados). Ainda encontramos em alguns dos *pools* de BACs pelo menos três genes semelhantes a L-aminoácido oxidases com 7 éxons e 6 íntrons, embora apenas dois pareçam estar expressos no transcriptoma das glândulas de veneno. É provável que estes dois genes encontrados expressos nas glândulas de veneno, a exemplo de outros genes de toxinas, estejam *in tandem* no genoma, uma vez que foram encontrados mais de uma cópia em diferentes contigs do mesmo pool de BACs. O isolamento e posterior sequenciamento destes BACs individuais pode esclarecer melhor a estrutura destes *loci* gênicos.

6 DISCUSSÃO

Répteis divergiram dos primeiros tetrápodes no final do período Carbonífero há cerca de 310-320 milhões de anos (WANG; KUMAR; HEDGES, 1999). A evolução do ovo amniótico foi uma das grandes inovações evolucionárias na história de vida, libertando vertebrados de uma conexão obrigatória com a água e permitindo assim a conquista de ambiente terrestre (VITT; CALDWELL, 2014). Desde então, estes amniotas antigos se espalharam por todo o mundo e adotaram diversas morfologias e habitats (aquáticos e terrestres). Os répteis pertencentes à ordem Squamata são os vertebrados terrestres mais diversificados. O sequenciamento dos genomas de espécies pertencentes a este grande grupo foi importante para o estudo de diversas adaptações evolutivas, dentre as quais podem ser citadas: a habilidade de regeneração de caudas, presença de dedos adesivos, visão noturna, diferentes mecanismos de determinação sexual (ZW, XY e dependente de temperatura) e capacidade de mergulho por longos períodos (ALFOLDI et al., 2011; LIU et al., 2015; WAN et al., 2013). As serpentes são o segundo grupo mais diverso dentro da ordem, apresentando muitas características interessantes do ponto de vista evolutivo, como: perda de membros, alongamento do corpo, reprodução variada, presença de órgãos sensoriais especializados e o desenvolvimento de glândulas de veneno. Tais características foram pouco exploradas em análises moleculares comparativas com outros animais. O sequenciamento genômico fornece subsídios para estudos mais aprofundados em relação às características peculiares das serpentes, em especial do sistema venenífero. Neste trabalho, a partir da construção e sequenciamento de uma biblioteca *paired-end* e cinco bibliotecas *mate-pair*, pudemos obter uma diversidade de fragmentos de diferentes tamanhos que auxiliaram nas etapas de montagem. Os *reads* de bibliotecas *mate-pair* são fundamentais nas etapas de scaffolding do genoma, sendo a utilização de várias abordagens de sequenciamento comuns em projetos genoma (GREEN et al., 2014; ZHANG et al., 2012; ZIMIN et al., 2014).

Comparativamente, a cobertura de 225 X obtida é maior do que em outros exemplos de genomas complexos sequenciados recentemente, havendo quatro genomas publicados de maior relevância para este projeto: *Ophiophagus hannah* (VONK et al., 2013), *Python molurus bibittatus* (CASTOE et al., 2013), *Alligator sinensis* (WAN et al., 2013) e *Anolis carolinensis* (ALFOLDI et al., 2011). Porém, ficou claro, conforme

discutido mais adiante, que apenas uma alta cobertura do genoma não é suficiente para se garantir uma boa montagem do mesmo.

O presente trabalho insere-se dentro do contexto de um projeto amplo que objetiva o sequenciamento genômico da serpente *Bothrops jararaca* para descoberta e análise dos genes de toxinas. Dessa forma, as responsabilidades diretamente alocadas a este trabalho de doutorado, envolveram as seguintes etapas: 1) obtenção e preparo de material biológico; 2) construção, sequenciamento e análise de bibliotecas de BACs; 3) construção de bibliotecas de sequenciamento em sistema Illumina; 4) acompanhamento dos sequenciamentos em sistema Illumina; 5) geração e análise de dados transcriptômicos; 6) identificação e análise preliminar de genes de interesse. Todas estas etapas foram cumpridas com êxito, apesar das dificuldades inerentes ao projeto.

A etapa de montagem bioinformática do genoma, ficou a cargo da equipe de bioinformática do projeto, coordenada pelo Prof. João Carlos Setubal, do Instituto de Química da USP, com participação indireta do doutorando. Os resultados dessa etapa foram críticos e necessários para a identificação e análise de genes de interesse e o consequente descerramento das conclusões esperadas do estudo. Uma análise prévia das montagens iniciais obtidas pela equipe indicou que o tamanho médio dos contigs obtidos era inferior ao desejado (melhor montagem com um N50 de 3134 pb), a despeito da alta cobertura obtida. Felizmente, identificamos o software ALLPATHS-LG em meio a descrição metodológica informada no BioProject PRJNA170536 do GenBank que disponibilizava contigs genômicos da serpente *Vipera berus berus*, porém sem artigo publicado e sem informações a respeito de parâmetros utilizados. A descrição do arquivo informava uma métrica bastante eficiente usando tal software. O mesmo software também foi utilizado na montagem do genoma de *Thamnophis sirtalis*, com resultados semelhantes, como consta no Bioproject PRJNA294278. A montagem realizada pelo bioinformata Milton Nishiyama com um subconjunto dos reads e utilizando o ALLPATHS-LG resultou em métricas muito superiores às obtidas até então e um pouco melhor do que as de *Ophiophagus hannah* (Tabela 13).

Tabela 13 - Comparação entre montagens de diferentes genomas de répteis.

	Trabalhos publicados				Não publicados		Este trabalho
Organismo	<i>Ophiophagus hannah</i>	<i>Python molurus</i>	<i>Anolis carolinensis</i>	<i>Alligator sinensis</i>	<i>Thamnophis sirtalis</i>	<i>Protobothrops mucrosquamatus</i>	<i>Bothrops jararaca</i>
Tamanho total (Gpb)	1,59	1,43	1,80	2,27	1,42	1,67	1,4
Número de scaffolds	296.399	39.113	6.646	9.317	7.930	52.281	15.748
Scaffold N50 (pb)	241.519	213.970	4.033.265	2.188.296	647.592	424.052	526.000
Número de contigs	816.633	274.244	41.987	177.282	175.977	167.852	281.308
Contig N50 (pb)	5.201	10.658	79.867	23.412	10.447	21.948	7.200

O maior problema de montagem de genomas de eucariotos superiores ocorre por causa de repetições longas e muito frequentes e pelo alto conteúdo GC de determinadas regiões. Uma análise feita com o software RepeatMasker utilizando parâmetros *default* e a espécie *Anolis carolinensis* como referência revelou que 10% do total de bases dos scaffolds produzidos na montagem 1.0 com o software ALLPATHs LG são repetições, sendo os *Long Interspersed Elements* (LINEs) do tipo L2/CR1/Rex os mais abundantes. No entanto, esse número deve ser bem maior, uma vez que uma das etapas da montagem *De novo* com o AllPaths-LG é a remoção de leituras com alta representatividade e cobertura, como dito anteriormente. Castoe et al., (2011) fizeram uma comparação entre os genomas da *Agkistrodon contortrix* (família Viperidae) e *Python molurus bivittatus* (família Pythonidae) com relação ao conteúdo repetitivo, tendo mostrado que o número de sequências repetitivas em geral é maior na primeira. O elemento L2/CR1/Rex também foi o mais frequentemente encontrado. A atividade deste elemento na superfamília Colubroidea pode ter facilitado, ao menos em parte, a expansão das famílias gênicas de veneno e/ou a reorganização das redes regulatórias gênicas (CASTOE et al., 2011).

Um fator que pode contribuir para a elongação dos contigs é o uso de sequenciamentos de leituras longas. No projeto original havíamos previsto o uso da tecnologia 454-Roche para obtenção de reads de até 800pb, mas com baixa cobertura. Porém, verificamos que essa tecnologia vem sendo cada vez menos utilizada, inclusive foi descontinuada pelo fabricante. Em contrapartida, o sistema da Pacific Biosciences (PacBio), que ainda era incipiente e com poucos prestadores de serviço quando da proposição do projeto se popularizou e hoje pode ser usado *per se* para a geração de

genomas complexos *de novo* (vide <http://www.pacb.com/smrt-science/smrt-resources/pacbio-literature/>). Dessa forma, requisitamos também a execução de sequenciamento PacBio por meio de uma *facility* prestadora de serviços da Duke University que possui a plataforma PacBio RS II. Apesar de gerar reads mais longos, o sequenciamento em sistema PACBIO perde em qualidade dos *reads*, requerendo uma alta cobertura para serem utilizados por si só. Dessa forma, os dados do PacBio foram usados em casos isolados nesse trabalho, porém estes ainda serão agregados ao conjunto de sequências curtas gerado, para uma nova tentativa de montagem futuramente.

Considerando as dificuldades citadas para a montagem *de novo*, além do sequenciamento do genoma total, é comum em projetos desse tipo a construção de bibliotecas de insertos longos como as de fosmídeos e BACs (ALFOLDI et al., 2011; ZHANG et al., 2012). Particularmente, a obtenção de bibliotecas de BACs foi uma tarefa assumida integralmente como parte desse trabalho de doutorado, para a qual a maior parte do esforço foi dispendido, uma vez que prevíamos dificuldades relacionadas à montagem. Bibliotecas de grandes insertos de DNA têm sido ferramentas importantes no mapeamento físico, clonagem de genes e análise da estrutura e função gênica de vários organismos (ALFOLDI et al., 2011; KIRSCH et al., 2008; SHIRASU et al., 1999). Sendo os vetores de BACs frequentemente utilizados para a construção destas bibliotecas devido à facilidade do manuseio e isolamento, propagação dos clones e relativa alta estabilidade (YUKSEL; PATERSON, 2005). Bibliotecas de BACs já foram feitas para diferentes organismos e tem sido empregadas em diversos campos da genômica, por exemplo: sequenciamento de genomas (SATO et al., 2011; VENTER; SMITH; HOOD, 1996), isolamento de genes (COYNE et al., 2007; PAIVA et al., 2011), análises comparativas de estrutura gênica e sintenia (ZHAO et al., 2014), integração de mapas de ligação e cromossômico (FUJI et al., 2014), análise do genoma funcional em larga escala (JOHNSON; WADE-MARTINS, 2011), prospecção de SNPs (OLLITRAULT et al., 2012), entre outras.

Apesar das vantagens, a construção desse tipo de biblioteca não é trivial. É necessário bastante tempo para a padronização da técnica, além disso normalmente esse tipo de biblioteca requer a utilização de equipamentos robóticos para coleta e análise dos clones. O manuseamento de DNA linear de alta massa molecular exige cuidados especiais devido à extrema instabilidade. Alguns trabalhos e kits comerciais

sugerem o uso da enzima agarase para purificação do DNA de alta massa necessário para clonagem (SHIZUYA et al., 1992). No entanto, essa abordagem se mostrou pouco útil, tendo sido gasto um longo tempo com tentativas de preparação do vetor e de clonagens que resultaram em baixa eficiência. Finalmente, após recebermos auxílio do Laboratório de Análise Genética e Molecular (UNICAMP), na pessoa da professora Anete Pereira de Souza e seu aluno Danilo Augusto, adotamos o método da eletroeluição para purificação do DNA de alta massa molecular (LUO; WING, 2003). Tal método se mostrou fundamental para a obtenção de DNA íntegro e tem sido bastante utilizado para esse propósito (JANG et al., 2010; LUO et al., 2001; OSOEGAWA et al., 1998). Dessa forma, obtivemos uma biblioteca com fragmentos no tamanho desejado (>100 kb) e com um rendimento razoável. Alguns destes clones foram selecionados aleatoriamente e sequenciados conforme descrito anteriormente, para confirmar a qualidade da biblioteca. Apesar do sucesso na obtenção de clones com insertos genômicos longos de *B. jararaca*, não dispúnhamos do equipamento necessário para coleta de todos os clones. Além disso, o *screening* desses clones demandaria o uso de membranas de nylon de alta densidade para hibridização ou seria necessário uma estratégia de triagem usando PCR em pools de clones. Para a seleção por PCR de pools de BACs é necessário um agrupamento da biblioteca, envolvendo combinações de placas, linhas e colunas. Essa combinação é feita de forma que um mínimo de reações de amplificação seja realizado para a identificação de um clone específico dentro da biblioteca (GREEN; OLSEN, 1990). Apesar do PCR em pools ser uma estratégia viável, requer: tempo, a confecção de diversos primers em quantidades razoáveis e conhecimento prévio das sequências gênicas.

Considerando que o resultado do sequenciamento piloto mostrou que é possível obter contigs longos mesmo utilizando *pools* de BACs resolvemos elaborar uma nova estratégia para prospecção e análise dos BACs, o que envolveu o sequenciamento em larga escala de *pools* de 12 BACs em plataforma Illumina. Os dados gerados tem a vantagem de serem menos complexos do que o genoma total, o que resulta em contigs maiores e ainda permitem realizar uma prospecção rápida de genes de interesse *in silico*, a qualquer tempo, conforme surjam novos interesses de pesquisa. Os clones físicos permanecem como um banco relacionado às sequências dos pools, facilitando a identificação e isolamento de clones individuais. Uma vez isolados, os clones podem ser sequenciados individualmente ou utilizados como sondas para mapeamento físico de

cromossomos através da técnica FISH (*Fluorescence In situ Hybridization*). Utilizando essa estratégia foi possível, com apenas uma parte da biblioteca de BACs, estabelecer a estrutura gênica de diversas classes de toxinas. Essa nova abordagem foi muito bem sucedida e ainda pode ser utilizada em projetos futuros do nosso grupo envolvendo genômica de serpentes.

Paralelamente ao sequenciamento genômico, realizamos também o sequenciamento de mRNAs de diferentes tecidos de *B. jararaca*. Essa análise transcriptômica possui três finalidades: 1) prover subsídios para a identificação de elementos expressos no genoma; 2) realizar, uma investigação comparativa em larga escala entre transcritos da glândula de veneno e de outros tecidos; 3) Facilitar a montagem de scaffolds para alguns genes. Além disso, identificar genes relacionados às toxinas que possam prover um *background* para análises evolutivas. Esse aspecto é bastante relevante para o entendimento da evolução das toxinas e vem sendo discutido na literatura (CASEWELL et al., 2013).

Com os dados obtidos nas etapas discutidas acima, nós focamos na determinação das estruturas gênicas e contexto genômico de toxinas. Os resultados obtidos trazem implicações importantes.

Até o momento, foi possível determinar a estrutura gênica completa ou quase completa de vários genes de toxinas importantes, entre eles: Metaloproteases, Serinoproteases, BPP/CNP, Fosfolipase B, VEGF, Hialuronidase e CRISP. Além disso, depreendemos o contexto genômico de vários destes genes. A partir dessas novas informações conseguimos entender um pouco melhor a variabilidade das toxinas presentes em venenos de serpentes e levantar novas questões acerca da origem evolutiva das toxinas. Como dito anteriormente, uma das perguntas levantadas é: como surgem as diversas isoformas de toxinas? Nossos resultados mostraram que muitas das classes de toxinas evoluíram a partir de duplicações *in tandem*, mas não todas elas. Algumas das duplicações podem ter ocorrido antes da diversificação das serpentes (uma vez que podem ser observadas em genomas de lagartos), outras parecem ter ocorrido depois. Por exemplo, é possível que um gene CRISP ancestral tenha dado origem aos genes CRISP1, CRISP2 e CRISP3 de mamíferos e às sequências Serotriflin-like e svCRISP-like de aves e répteis. Parece ter havido um rearranjo gênico em squamata que deu origem a um contexto genômico diferente daquele de mamíferos, aves e tartarugas. No clado toxicofera as svCRISPs seriam então recrutadas para o veneno, conforme as

análises filogenéticas indicavam (FRY et al., 2006). Com relação às hialuronidases, possivelmente o ortólogo do gene da hialuronidase expresso nas glândulas de veneno é o Hyal6(Hyalp1) de mamíferos, sendo que este gene pode ter sido simplesmente recrutado para a glândula de veneno em serpentes, embora possivelmente tenha havido duplicações em um organismo ancestral para dar origem aos genes Hyal6, Spam1, Hyal4 e Hyal5 (que possuem estrutura gênica semelhante). Aparentemente em *Anolis carolinensis* também podem ter ocorrido duplicações mais recentes do gene da Hyalp1.

Outro gene que aparentemente foi recrutado para ser expresso nas glândulas de veneno é o que codifica a Fosfolipase B, que corresponde ao gene PLBD1 em mamíferos e outros organismos. Fosfolipases B atuam clivando ligações éster nas posições sn-1 e sn-2 de glicerofosfolipídios de membranas (CHAPEAUROUGE et al., 2015) e já foram identificadas em venenos de abelhas e serpentes (DOERY; PEARSON, 1964). A PLB de *Pseudechis colletti* pode degradar a fosfatidilcolina e fosfatidiletanolamina, além disso, apresenta forte atividade hemolítica *in vitro* sobre eritrócitos de coelho e humanos (BERNHEIMER et al., 1987). Porém, seu papel no envenenamento ainda é pouco compreendido. O seu provável ortólogo PLBD1 exibe uma atividade fosfolipásica fraca, agindo em diversos fosfolipídios, incluindo fosfatidilcolina, fosfatidilinositol, fosfatidiletanolamina e lisofosfolipídios. Foi proposto que o mesmo pode atuar como uma amidase ou uma peptidase tendo em vista o pequeno tamanho do seu provável sítio ativo (REPO et al., 2014).

Já o VEGF-F, surpreendentemente, parece possuir ortólogos em diversos organismos, incluindo *Latimeria chalumnae*, que é considerado um “fóssil vivo”, cuja morfologia mudou pouco quando comparada a fósseis datados de pelo menos 300 milhões de anos (AMEMIYA et al., 2013). Yamazaki et al., (2009) levantaram a hipótese de que o VEGF-F seria derivado do VEGF-A, baseando-se em uma pequena conservação de regiões específicas dos íntrons e aproximadamente 50% de identidade na sequência nucleotídica da região C-terminal. Nossos dados, por outro lado, sugerem que o VEGF-F evoluiu a partir de um gene anotado como “PGF-like” em outros organismos, não tendo sido uma duplicação recente do VEGF-A. Curiosamente, proteínas do tipo VEGF-F foram identificadas apenas em glândulas de veneno de viperídeos, embora possa haver expressão em baixos níveis em outros tecidos (JUNQUEIRA-DE-AZEVEDO et al., 2015). Reyes-velasco et al., (2015) encontraram uma expressão considerável de transcritos homólogos ao VEGF de serpentes no ovário de *Python molurus bivittatus*, embora não

tenha ficado claro qual gene está sendo considerado. É provável que os ortólogos do VEGF-F desempenhem algum papel fisiológico importante ainda desconhecido em outros organismos.

Dentre os componentes de veneno de *B. jararaca*, as principais classes de toxinas são as SVMPs, SVSPs, Lectinas, BPPs e Fosfolipases A2. Sendo as SVMPs e SVSPs as mais representativas e mais variáveis. Já foi mostrado anteriormente que os genes de Fosfolipases A2 ocorrem *in tandem* no genoma de *P. flavoviridis* (IKEDA et al., 2010). Nossos dados demonstram que o mesmo ocorre para as SVMPs e SVSPs. As SVSPs de veneno já foram encontradas em lagartos do clado toxicofera em glândulas da mandíbula inferior, além disso, genes de SVSPs estão presentes tanto no genoma de *Anolis carolinensis* quanto de *Gekko japonicus* no mesmo contexto genômico dos genes de SVSPs de *P. mucrosquamatus* e muito provavelmente de *B. jararaca*. Isso demonstra que tais genes já existiam nesse locus antes mesmo da diversificação do clado toxicofera, apesar de existirem poucas cópias destes (apenas duas cópias). Nós encontramos pelo menos 6 genes completos e 4 incompletos de SVSPs de *B. jararaca*, sendo que 5 deles puderam ser identificados com transcritos já conhecidos no transcriptoma (JUNQUEIRA-DE-AZEVEDO et al., 2015). Isso sugere que a diversificação dos genes de SVSPs deve ter ocorrido nas serpentes avançadas (Caenophidia), contribuindo para a complexidade e variabilidade do veneno de serpentes.

Em relação aos genes de SVMPs, pudemos observar que pelo menos dois genes de SVMPs do tipo PIII ocorrem no mesmo locus genômico do gene da ADAM28, à jusante deste. Tal disposição não é encontrada em *Anolis carolinensis*, *Gekko japonicus* e *Gallus gallus*, onde há somente o gene da ADAM28 seguido dos genes NEFM e NEFL. Além disso, outros genes de SVMPs do tipo PII e PIII foram encontrados em BACs e outros *scaffolds*, vários também aparecem *in tandem*. Isso leva a crer que muitos outros genes de SVMPs podem estar dispostos em sequência após o gene da ADAM28, formando um *cluster* enorme. Já se sabia através de análises filogenéticas que a ADAM28 é o gene mais próximo das SVMPs do ponto de vista evolutivo. Além disso, Sanz e Calvette (2012) mostraram que um dos genes de SVMP do tipo PIII de *Echis ocellatus* conserva uma estrutura gênica semelhante à ADAM28, mas não idêntica. Nossos dados mostram que as SVMPs do tipo PIII possuem uma estrutura éxon-íntron idêntica à do gene da ADAM28. As SVMPs do tipo PII seguem o mesmo padrão, mas parecem ter perdido uma região que compreende o final do éxon 14 até o início do éxon 16, incluindo o éxon 15 (ver figura

31). Dessa forma, o domínio rico em cisteína foi praticamente deletado, assim como as SVMPs-PIII perderam a região C-terminal da ADAM28. Tal mecanismo evolutivo, com perda de domínios funcionais foi proposto para as metaloproteases de veneno, incluindo a origem dos tipos P-II e PI (CASEWELL et al., 2011).

Nesse contexto, é interessante que não tenhamos encontrado nenhum gene correspondente a uma SVMP do tipo P-I no genoma e nem no transcriptoma das glândulas de veneno desse indivíduo. A ausência de SVMPs do tipo P-I foi uma surpresa para nós e, juntamente com os outros achados, nos faz refletir sobre a origem da variabilidade de composição e atividade do veneno de *Bothrops jararaca*. Os eventos de duplicação no locus da ADAM28 parecem ter ocorrido nas serpentes avançadas assim como no locus das SVSPs. Tais eventos ocorrem por algum tipo de recombinação genômica, provavelmente *crossing over* desigual. Estes eventos contituem-se numa explicação bastante plausível para a ocorrência recorrente de perdas de domínios, podem explicar ainda a presença de genes truncados (provavelmente não funcionais) e a ocorrência de múltiplas P-Is com diferentes estruturas evoluindo independentemente em Viperidae. A pergunta que fica é: esses eventos de recombinação ainda ocorrem ao nível individual de forma recorrente? Em caso positivo, isso também explicaria a ausência de um gene de P-I nesse indivíduo, bem como a observação de *exon shuffling* em metaloproteases de veneno (MOURA-DA-SILVA et al., 2011). Se não, outra explicação seria a oscilação alélica nas populações de *B. jararaca*. Dessa forma, indivíduos diferentes podem apresentar um perfil diferente de SVMPs tanto quantitativamente como qualitativamente, o que pode refletir diretamente na atividade do veneno.

A disposição *in tandem* também pode estar presente em genes de outras toxinas, como L-amino ácido oxidases, embora ainda não tenhamos alcançado uma boa resolução para este *locus*.

Nossos dados apontam uma história evolutiva diferente para os BPPs. Estes parecem derivar de um precursor do gene do Peptídeo Natriurético do tipo C, cujo éxon que codifica o peptídeo sinal foi ampliado ganhando a região codificante dos BPPs. Apesar disso, não está claro ainda se há mais de uma cópia deste gene em *B. jararaca*.

O trabalho desenvolvido representa um dos primeiros esforços para estabelecer as bases genômicas de serpentes, em especial para a família Viperidae. As serpentes constituem uma infraordem dos répteis escamados (Squamata), apresentando uma morfologia alongada e sem membros. Apesar desse padrão, apresentam um amplo

espectro de formas, tamanho e aparência relacionado à diversidade comportamental, ecológica, fisiológica e alimentar (VITT; CALDWELL, 2014). No caso das serpentes avançadas, há uma gama de compostos produzidos em glândulas orais e de veneno (JUNQUEIRA-DE-AZEVEDO et al., 2016) que são úteis para predação e defesa. Esses compostos variam enormemente nas diversas famílias e subfamílias de serpentes, sendo um modelo interessante para estudos de evolução molecular. Foi possível observar ainda que arquétipos e o contexto genômico de alguns genes de toxinas (como SVSPs e CRISPs) são semelhantes para serpentes e lagartos, o que suporta a origem comum das toxinas ao menos para certas classes, apesar de cada grupo possuir venenos peculiares, com tipos de toxinas exclusivos. Porém, apenas com base nisso não é possível afirmar em que momento os mesmos assumiram o papel de toxinas para cada grupo, já que tais genes foram encontrados inclusive fora do clado Toxicofera. As nossas análises indicam que a natureza do veneno reflete mecanismos de duplicação, deleção, inserção, recrutamento ou restrição/especialização de genes envolvidos com diferentes funções fisiológicas em outros organismos. Esses fenômenos evolutivos repercutem na variabilidade e nos diversos mecanismos de ação conhecidos para toxinas de serpentes. Os resultados alcançados são importantes para o entendimento de tais fenômenos, bem como servem de base para outros estudos sobre venenos de serpentes e biologia em geral.

7 CONCLUSÃO

Neste projeto, foi proposto o sequenciamento com alta cobertura do genoma de *B. jararaca* para obter um conjunto de segmentos genômicos que permitisse identificar genes de toxinas e descrever suas estruturas. Como conclusões do sequenciamento, montagem e anotação desse genoma temos que:

- 1) O genoma foi sequenciado com mais de 200 X de cobertura bruta em *reads* curtos, utilizando leituras tipo *paired-end* e *mate-pair*, geradas na plataforma Illumina em nosso laboratório.
- 2) Para a montagem desses dados, o pipeline baseado no software ALLPATHS-LGv2 e aplicado sobre subconjuntos limitados de *reads* foi a estratégia que apresentou melhor resultado, gerando 281.000 contigs com N50 = 7.2 kb, arranjados em 15.748 *scaffolds* com N50 = 526 kb. Estas métricas de montagem são superiores à do genoma de *Ophiophagus hannah* publicado e próximas a alguns recentemente disponibilizados em bancos públicos.
- 3) Diversas outras tentativas de montagem desse conjunto de dados mostraram ser pouco viável a obtenção de um rascunho genômico baseando-se apenas nessa estratégia.
- 4) Tal dificuldade foi atribuída ao alto teor de elementos repetitivos no genoma e a eventual necessidade de maiores recursos computacionais para a tarefa, incompatíveis com as dimensões do projeto.
- 5) O genoma mitocondrial pôde ser completamente resolvido e disponibilizado.
- 6) Os transcriptomas de 11 tecidos foram sequenciados com alta cobertura na plataforma Illumina, facilitando a anotação gênica e permitindo a comparação entre os níveis de expressão de genes referenciais de componentes do veneno.
- 7) O conjunto de *scaffolds* foi anotado por meio da predição *Ab initio* de genes, pelo mapeamento de leituras transcriptômicas e pelo alinhamento das regiões codificantes com proteínas ofídicas disponíveis no Uniprot. Tais informações foram inseridas num visualizador genômico (gBrowse) vinculado ao servidor público CeTICs. Esta está sendo considerada a versão 1.0 do genoma e o acesso público aos dados deverá ser liberado em breve no GenBank dentro do Bioproject PRJNA343256.

- 8) O genoma foi também sequenciado com leituras longas produzidas em plataforma PacBio com cerca de 30 X de cobertura. A alta taxa de erros desse tipo de sequenciamento e a grande demanda de recursos computacionais para correção de bases não permitiu ainda a combinação desses dados com o restante das sequências geradas no projeto. Entretanto, *reads* individuais mapeados contra genes específicos mostraram-se importantes para a junção de *scaffolds* selecionados e preenchimento de gaps dos mesmos.
- 9) Como estratégia paralela para um sequenciamento mais orientado, construímos bibliotecas de grandes fragmentos genômicos (100-200 kb) clonados em BACs. Clones individuais de BACs foram estocados e sequenciados em *pools* na plataforma Illumina. Após *screening in silico*, clones específicos contendo genes de interesse foram ressequenciados.
- 10) Essa estratégia mostrou-se muito produtiva, pois permitiu a identificação da maior parte dos genes de interesse. Mais do que isso, sendo essas sequências derivadas de trechos definidos do genoma, suas análises permitiram maior clareza na determinação das estruturas gênicas e, em alguns casos, certeza na identificação do contexto genômico, incluindo a co-localização de genes parálogos.

A partir dos dados genômicos produzidos e organizados conforme acima, pudemos prospectar e analisar os genes de toxinas de *B. jararaca*. Como conclusões em relação à identificação e caracterização da organização genômica dos genes de toxinas temos que:

- 1) Foram recuperados genes da maioria das famílias de toxinas conhecidas do gênero *Bothrops*.
- 2) Para a maioria das famílias de toxinas, ao menos um gene completo com fronteiras intro-éxon foi determinado, permitindo o delineamento da estrutura gênica típica de cada família de toxina.
- 3) Particularmente, genes completos de SVMPs tipo P-II e P-III indicaram um arranjo de íntrons semelhante ao observado para o suposto gene ancestral (ADAM28), diferentemente do previamente descrito na literatura.

- 4) O gene do precursor dos BPPs, rico em CG, só pôde ser obtido a partir de *reads* do PacBio e sua organização estrutural mostrou a presença de um íntron separando a região codificante dos BPPs da região espaçadora que antecede o CNP. Isso reforça a hipótese da origem do gene dos BPPs a partir de uma inserção no gene dos CNPs.
- 5) Para algumas famílias de toxinas, como SVMPS e SVSPs, há claras evidências de organização *in tandem* dos parálogos. Para outras, o contexto genômico indica a presença de cópias únicas. Algumas classes de toxinas como Fosfolipases A2 e L-aminoácido oxidases ainda precisam ser melhor investigadas, mas há indícios de que também ocorram *in tandem*.
- 6) O contexto genômico também forneceu algumas indicações a respeito da origem de algumas famílias de toxinas. Por exemplo, SVMPS flanqueiam o gene da ADAM28, seu suposto ancestral, indicando que de fato devem ter se originado a partir da duplicação desse gene. Por outro lado, o gene do VEGF-F ocupa a posição de um gene relacionado ao PGF em outros animais, indicando que deva ter surgido por derivação desse gene e não por duplicação a partir do gene do VEGF-A.

REFERÊNCIAS*

- ALFOLDI, J. et al. The genome of the green anole lizard and a comparative analysis with birds and mammals. **Nature.**, v. 477, n. 7366, p. 587-591, Sep 29 2011.
- ALVES, R. M. et al. Evidence of caspase-mediated apoptosis induced by l-amino acid oxidase isolated from *Bothrops atrox* snake venom. **Comp. Biochem. Physiol. A. Mol. Integr. Physiol.**, v. 151, n. 4, p. 542-550, Dec 2008.
- AMEMIYA, C. T. et al. The African coelacanth genome provides insights into tetrapod evolution. **Nature.**, v. 496, n. 7445, p. 311-316, Apr 18 2013.
- AOKI, N. et al. Serotriflin, a CRISP family protein with binding affinity for small serum protein-2 in snake serum. **Biochim. Biophys. Acta.**, v. 1784, n. 4, p. 621-628, Apr 2008.
- ATKIN, N. B. et al. The comparative DNA content of 19 species of placental mammals, reptiles and birds. **Chromosoma.**, v. 17, n. 1, p. 1-10, 1965.
- AUGUSTO-DE-OLIVEIRA, C. et al. Dynamic Rearrangement in Snake Venom Gland Proteome: Insights into *Bothrops jararaca* Intraspecific Venom Variation. **J. Proteome. Res.**, Sep 12 2016.
- BAO, W.; KOJIMA, K. K.; KOHANY, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. **Mob. DNA.**, v. 6, p. 11, 2015.
- BARRETT, A. J.; RAWLINGS, N. D. Families and clans of serine peptidases. **Arch. Biochem. Biophys.**, v. 318, n. 2, p. 247-250, Apr 20 1995.
- BECAK, W.; BECAK, M. L.; NAZARETH, H. R. Karyotypic studies of two species of South American snakes (*Boa constrictor amarali* and *Bothrops jararaca*). **Cytogenetics.**, v. 1, p. 305-313, 1962.
- BEJDER, L.; HALL, B. K. Limbs in whales and limblessness in other vertebrates: mechanisms of evolutionary and developmental transformation and loss. **Evol. Dev.**, v. 4, n. 6, p. 445-458, Nov-Dec 2002.
- BERNARDONI, J. L. et al. Functional variability of snake venom metalloproteinases: adaptive advantages in targeting different prey and implications for human envenomation. **PLoS. One.**, v. 9, n. 10, p. e109651, 2014.
- BERNHEIMER, A. W. et al. Isolation and characterization of a phospholipase B from venom of Collett's snake, *Pseudechis colletti*. **Toxicon.**, v. 25, n. 5, p. 547-554, 1987.
- BJARNASON, J. B.; FOX, J. W. Snake venom metalloendopeptidases: reprotinsins. **Methods. Enzymol.**, v. 248, p. 345-368, 1995.
- BODE, W.; GOMIS-RUTH, F. X.; STOCKLER, W. Astacins, serralsins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the 'metzincins'. **FEBS. Lett.**, v. 331, n. 1-2, p. 134-140, Sep 27 1993.
- BOETZER, M. et al. Scaffolding pre-assembled contigs using SSPACE. **Bioinformatics.**, v. 27, n. 4, p. 578-579, Feb 15 2011.
- BRAHMA, R. K. et al. Venom gland transcriptomics for identifying, cataloging, and characterizing venom proteins in snakes. **Toxicon.**, v. 93, p. 1-10, Jan 2015.
- CAMARGO, A. C. et al. Bradykinin-potentiating peptides: beyond captopril. **Toxicon.**, v. 59, n. 4, p. 516-523, Mar 15 2012.
- CAMPBELL, J. A.; LAMAR, W. W. **The venomous reptiles of Latin America.** Comstock Pub. Associates, 1989.

*De acordo com: ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS.
NBR6023: informação e documentação: referências: elaboração. Rio de Janeiro, 2002.

- CASEWELL, N. R. On the ancestral recruitment of metalloproteinases into the venom of snakes. **Toxicon.**, v. 60, n. 4, p. 449-454, Sep 15 2012.
- CASEWELL, N. R.; HUTTLEY, G. A.; WUSTER, W. Dynamic evolution of venom proteins in squamate reptiles. **Nat. Commun.**, v. 3, p. 1066, 2012.
- CASEWELL, N. R. et al. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. **Mol. Biol. Evol.**, v. 28, n. 9, p. 2637-2649, Sep 2011.
- CASEWELL, N. R. et al. Complex cocktails: the evolutionary novelty of venoms. **Trends. Ecol. Evol.**, v. 28, n. 4, p. 219-229, Apr 2013.
- CASTOE, T. A. et al. The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. **Proc. Natl. Acad. Sci. U. S. A.**, v. 110, n. 51, p. 20645-20650, Dec 17 2013.
- CASTOE, T. A. et al. Discovery of highly divergent repeat landscapes in snake genomes using high-throughput sequencing. **Genome. Biol. Evol.**, v. 3, p. 641-653, 2011.
- CASTOE, T. A. et al. Rapid identification of thousands of copperhead snake (*Agkistrodon contortrix*) microsatellite loci from modest amounts of 454 shotgun genome sequence. **Mol. Ecol. Resour.**, v. 10, n. 2, p. 341-347, Mar 2010.
- CHAPEAUROUGE, A. et al. Interrogating the Venom of the Viperid Snake *Sistrurus catenatus edwardsii* by a Combined Approach of Electrospray and MALDI Mass Spectrometry. **PLoS. One.**, v. 10, n. 5, p. e0092091, 2015.
- CHEN, Y. C. et al. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. **PLoS. One.**, v. 8, n. 4, p. e62856, 2013.
- CHING, A. T. et al. Venomics profiling of *Thamnodynastes strigatus* unveils matrix metalloproteinases and other novel proteins recruited to the toxin arsenal of rear-fanged snakes. **J. Proteome. Res.**, v. 11, n. 2, p. 1152-1162, Feb 3 2012.
- CHING, A. T. et al. Some aspects of the venom proteome of the Colubridae snake *Philodryas olfersii* revealed from a Duvernoy's (venom) gland transcriptome. **FEBS. Lett.**, v. 580, n. 18, p. 4417-4422, Aug 7 2006.
- CHOMCZYNSKI, P.; SACCHI, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. **Anal. Biochem.**, v. 162, n. 1, p. 156-159, Apr 1987.
- CIDADE, D. A. et al. Bothrops jararaca venom gland transcriptome: analysis of the gene expression pattern. **Toxicon.**, v. 48, n. 4, p. 437-461, Sep 15 2006.
- CINTRA-FRANCISCHINELLI, M. et al. The C-terminal region of a Lys49 myotoxin mediates Ca²⁺ influx in C2C12 myotubes. **Toxicon.**, v. 55, n. 2-3, p. 590-596, Feb-Mar 2010.
- CLARKE, L.; CARBON, J. A colony bank containing synthetic Col EI hybrid plasmids representative of the entire E. coli genome. **Cell.**, v. 9, n. 1, p. 91-99, Sep 1976.
- COYNE, C. J. et al. Construction and characterization of two bacterial artificial chromosome libraries of pea (*Pisum sativum* L.) for the isolation of economically important genes. **Genome.**, v. 50, n. 9, p. 871-875, Sep 2007.
- CUSHMAN, D. W. et al. Development and design of specific inhibitors of angiotensin-converting enzyme. **Am. J. Cardiol.**, v. 49, n. 6, p. 1390-1394, Apr 21 1982.
- DAGDA, R. K. et al. Genetic Basis for Variation of Metalloproteinase-Associated Biochemical Activity in Venom of the Mojave Rattlesnake (*Crotalus scutulatus scutulatus*). **Biochem. Res. Int.**, v. 2013, p. 251474, 2013.

- DE CARVALHO, D. D.; MARANGONI, S.; NOVELLO, J. C. Primary structure characterization of Bothrops jararacussu snake venom lectin. **J. Protein. Chem.**, v. 21, n. 1, p. 43-50, Jan 2002.
- DESHIMARU, M. et al. Accelerated evolution of crotalinae snake venom gland serine proteases. **FEBS. Lett.**, v. 397, n. 1, p. 83-88, Nov 11 1996.
- DIAS, G. S. et al. Individual variability in the venom proteome of juvenile Bothrops jararaca specimens. **J. Proteome. Res.**, v. 12, n. 10, p. 4585-4598, Oct 4 2013.
- DOERY, H. M.; PEARSON, J. E. Phospholipase B in snake venoms and bee venom. **Biochem. J.**, v. 92, n. 3, p. 599-602, Sep 1964.
- DOLEY, R.; MACKESSY, S. P.; KINI, R. M. Role of accelerated segment switch in exons to alter targeting (ASSET) in the molecular evolution of snake venom proteins. **BMC. Evol. Biol.**, v. 9, p. 146, 2009.
- DOS SANTOS, J. I. et al. The intriguing phospholipases A2 homologues: relevant structural features on myotoxicity and catalytic inactivity. **Protein. Pept. Lett.**, v. 16, n. 8, p. 887-893, 2009.
- ELSIK, C. G. et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. **Science.**, v. 324, n. 5926, p. 522-528, Apr 24 2009.
- FERNANDEZ, J. H. et al. Molecular cloning, functional expression, and molecular modeling of bothrostatin, a new highly active disintegrin from Bothrops jararaca venom. **Biochem. Biophys. Res. Commun.**, v. 329, n. 2, p. 457-464, Apr 8 2005.
- FOX, J. W.; SERRANO, S. M. Structural considerations of the snake venom metalloproteinases, key members of the M12 repolysin family of metalloproteinases. **Toxicon.**, v. 45, n. 8, p. 969-985, Jun 15 2005.
- FOX, J. W.; SERRANO, S. M. Insights into and speculations about snake venom metalloproteinase (SVMP) synthesis, folding and disulfide bond formation and their contribution to venom complexity. **FEBS. J.**, v. 275, n. 12, p. 3016-3030, Jun 2008.
- FRY, B. G. et al. Early evolution of the venom system in lizards and snakes. **Nature.**, v. 439, n. 7076, p. 584-588, Feb 2 2006.
- FRY, B. G. et al. Evolution and diversification of the Toxicofera reptile venom system. **J. Proteomics.**, v. 72, n. 2, p. 127-136, Mar 6 2009.
- FUJI, K. et al. Construction of a high-coverage bacterial artificial chromosome library and comprehensive genetic linkage map of yellowtail Seriola quinqueradiata. **BMC. Res. Notes.**, v. 7, p. 200, 2014.
- FUNASA. **Manual de diagnóstico e tratamento de acidentes por animais peçonhentos**. SAÚDE, M. D. Brasília: Fundação Nacional de Saúde 2001.
- GIRISH, K. S. et al. Snake venom hyaluronidase: an evidence for isoforms and extracellular matrix degradation. **Mol. Cell. Biochem.**, v. 240, n. 1-2, p. 105-110, Nov 2002.
- GOPALAKRISHNAKONE, P.; KOCHVA, E. Histological features of the venom apparatus of sea snake Lapemis curtus. **Snake.**, v. 25, p. 27-37, 1993.
- GRABHERR, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. **Nat. Biotechnol.**, v. 29, n. 7, p. 644-652, Jul 2011.
- GREEN, E. D.; OLSON, M. V. Systematic screening of yeast artificial-chromosome libraries by use of the polymerase chain reaction. **Proc. Natl. Acad. Sci. U. S. A.**, v. 87, n. 3, p. 1213-1217, Feb 1990.
- GREEN, R. E. et al. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. **Science.**, v. 346, n. 6215, p. 1254449, Dec 12 2014.

GREGORY, T. R. Animal Genome Database. Disponível em: < <http://www.genomesize.com> >. Acesso em: 23 set. 2016

GUERREIRO, J. R. et al. Argininosuccinate synthetase is a functional target for a snake venom anti-hypertensive peptide: role in arginine and nitric oxide production. **J. Biol. Chem.**, v. 284, n. 30, p. 20022-20033, Jul 24 2009.

GUTIERREZ, J. M.; LOMONTE, B. Phospholipase A2 myotoxins from Bothrops snake venoms. **Toxicon.**, v. 33, n. 11, p. 1405-1424, Nov 1995.

HAIMINEN, N.; FELTUS, F. A.; PARIDA, L. Assessing pooled BAC and whole genome shotgun strategies for assembly of complex genomes. **BMC. Genomics.**, v. 12, p. 194, 2011.

HARGREAVES, A. D. et al. Testing the Toxicofera: comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system. **Toxicon.**, v. 92, p. 140-156, Dec 15 2014.

HILDEBRAND, M. **Análise da estrutura dos vertebrados**. Atheneu, 2006.

HUXLEY-JONES, J. et al. The evolution of the vertebrate metzincins; insights from *Ciona intestinalis* and *Danio rerio*. **BMC. Evol. Biol.**, v. 7, p. 63, 2007.

IANZER, D. et al. Do the cardiovascular effects of angiotensin-converting enzyme (ACE) I involve ACE-independent mechanisms? new insights from proline-rich peptides of *Bothrops jararaca*. **J. Pharmacol. Exp. Ther.**, v. 322, n. 2, p. 795-805, Aug 2007.

IKEDA, N. et al. Unique structural characteristics and evolution of a cluster of venom phospholipase A2 isozyme genes of *Protobothrops flavoviridis* snake. **Gene.**, v. 461, n. 1-2, p. 15-25, Aug 1 2010.

ITOH, N. et al. Organization of the gene for batroxobin, a thrombin-like snake venom enzyme. Homology with the trypsin/kallikrein gene family. **J. Biol. Chem.**, v. 263, n. 16, p. 7628-7631, Jun 5 1988.

JACKSON, T. N. et al. Venom down under: dynamic evolution of Australian elapid snake toxins. **Toxins. (Basel)**. v. 5, n. 12, p. 2621-2655, Dec 2013.

JANG, S. et al. Construction and characterization of two bacterial artificial chromosome libraries of grass carp. **Mar. Biotechnol. (NY)**. v. 12, n. 3, p. 261-266, Jun 2010.

JIANG, Z. J. et al. Comparative mitochondrial genomics of snakes: extraordinary substitution rate dynamics and functionality of the duplicate control region. **BMC. Evol. Biol.**, v. 7, p. 123, 2007.

JOHNSON, S. J.; WADE-MARTINS, R. A BACwards glance at neurodegeneration: molecular insights into disease from LRRK2, SNCA and MAPT BAC-transgenic mice. **Biochem. Soc. Trans.**, v. 39, n. 4, p. 862-867, Aug 2011.

JUAREZ, P. et al. Evolution of snake venom disintegrins by positive Darwinian selection. **Mol. Biol. Evol.**, v. 25, n. 11, p. 2391-2407, Nov 2008.

JUNQUEIRA-DE-AZEVEDO, I. L. et al. Venom-related transcripts from *Bothrops jararaca* tissues provide novel molecular insights into the production and evolution of snake venom. **Mol. Biol. Evol.**, v. 32, n. 3, p. 754-766, Mar 2015.

JUNQUEIRA-DE-AZEVEDO, I. L. et al. Colubrid Venom Composition: An -Omics Perspective. **Toxins. (Basel)**. v. 8, n. 8, 2016.

JUNQUEIRA-DE-AZEVEDO, I. L. et al. *Lachesis muta* (Viperidae) cDNAs reveal diverging pit viper molecules and scaffolds typical of cobra (Elapidae) venoms: implications for snake toxin repertoire evolution. **Genetics.**, v. 173, n. 2, p. 877-889, Jun 2006.

- JUNQUEIRA DE AZEVEDO, I. L. et al. Molecular cloning and expression of a functional snake venom vascular endothelium growth factor (VEGF) from the *Bothrops insularis* pit viper. A new member of the VEGF family of proteins. **J. Biol. Chem.**, v. 276, n. 43, p. 39836-39842, Oct 26 2001.
- KAMIGUTI, A. S. et al. Coagulopathy and haemorrhage in human victims of *Bothrops jararaca* envenoming in Brazil. **Toxicon.**, v. 29, n. 8, p. 961-972, 1991.
- KIM, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. **Genome. Biol.**, v. 14, n. 4, p. R36, 2013.
- KINI, R. M. Excitement ahead: structure, function and mechanism of snake venom phospholipase A2 enzymes. **Toxicon.**, v. 42, n. 8, p. 827-840, Dec 15 2003.
- KIRSCH, S. et al. Chromosomal evolution of the PKD1 gene family in primates. **BMC. Evol. Biol.**, v. 8, p. 263, 2008.
- KOREN, S. et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. **Nat. Biotechnol.**, v. 30, n. 7, p. 693-700, Jul 2012.
- LANGMEAD, B.; SALZBERG, S. L. Fast gapped-read alignment with Bowtie 2. **Nat. Methods.**, v. 9, n. 4, p. 357-359, Apr 2012.
- LECHT, S. et al. Anti-angiogenic activities of snake venom CRISP isolated from *Echis carinatus* sochureki. **Biochim. Biophys. Acta.**, v. 1850, n. 6, p. 1169-1179, Jun 2015.
- LIU, Y. et al. Gekko japonicus genome reveals evolution of adhesive toe pads and tail regeneration. **Nat. Commun.**, v. 6, p. 10033, 2015.
- LOMONTE, B.; GUTIERREZ, J. M. Phospholipases A2 from viperidae snake venoms: how do they induce skeletal muscle damage? **Acta. Chim. Slov.**, v. 58, n. 4, p. 647-659, Dec 2011.
- LUO, M. et al. Melon bacterial artificial chromosome (BAC) library construction using improved methods and identification of clones linked to the locus conferring resistance to melon Fusarium wilt (Fom-2). **Genome.**, v. 44, n. 2, p. 154-162, Apr 2001.
- LUO, M.; WING, R. A. An improved method for plant BAC library construction. **Methods. Mol. Biol.**, v. 236, p. 3-20, 2003.
- LYNCH, V. J. Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. **BMC. Evol. Biol.**, v. 7, p. 2, 2007.
- MADHANI, M. et al. Vascular natriuretic peptide receptor-linked particulate guanylate cyclases are modulated by nitric oxide-cyclic GMP signalling. **Br. J. Pharmacol.**, v. 139, n. 7, p. 1289-1296, Aug 2003.
- MARAGANORE, J. M. et al. A new class of phospholipases A2 with lysine in place of aspartate 49. Functional consequences for calcium and substrate binding. **J. Biol. Chem.**, v. 259, n. 22, p. 13839-13843, Nov 25 1984.
- MATSUI, T.; FUJIMURA, Y.; TITANI, K. Snake venom proteases affecting hemostasis and thrombosis. **Biochim. Biophys. Acta.**, v. 1477, n. 1-2, p. 146-156, Mar 7 2000.
- MENEZES, M. C. et al. Sex-based individual variation of snake venom proteome among eighteen *Bothrops jararaca* siblings. **Toxicon.**, v. 47, n. 3, p. 304-312, Mar 2006.
- MONTECUCCO, C.; GUTIERREZ, J. M.; LOMONTE, B. Cellular pathology induced by snake venom phospholipase A2 myotoxins and neurotoxins: common aspects of their mechanisms of action. **Cell. Mol. Life. Sci.**, v. 65, n. 18, p. 2897-2912, Sep 2008.
- MORITA, T. Structures and functions of snake venom CLPs (C-type lectin-like proteins) with anticoagulant-, procoagulant-, and platelet-modulating activities. **Toxicon.**, v. 45, n. 8, p. 1099-1114, Jun 15 2005.

- MOURA-DA-SILVA, A. M. et al. Diversity of metalloproteinases in *Bothrops neuwiedi* snake venom transcripts: evidences for recombination between different classes of SVMPs. **BMC. Genet.**, v. 12, p. 94, 2011.
- MOURA-DA-SILVA, A. M.; THEAKSTON, R. D.; CRAMPTON, J. M. Evolution of disintegrin cysteine-rich and mammalian matrix-degrading metalloproteinases: gene duplication and divergence of a common ancestor rather than convergent evolution. **J. Mol. Evol.**, v. 43, n. 3, p. 263-269, Sep 1996.
- MURAYAMA, N. et al. Cloning and sequence analysis of a *Bothrops jararaca* cDNA encoding a precursor of seven bradykinin-potentiating peptides and a C-type natriuretic peptide. **Proc. Natl. Acad. Sci. U. S. A.**, v. 94, n. 4, p. 1189-1193, Feb 18 1997.
- MURPHY, J. F.; FITZGERALD, D. J. Vascular endothelial growth factor induces cyclooxygenase-dependent proliferation of endothelial cells via the VEGF-2 receptor. **FASEB. J.**, v. 15, n. 9, p. 1667-1669, Jul 2001.
- NAKASHIMA, K. et al. Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A2 isozyme genes. **Proc. Natl. Acad. Sci. U. S. A.**, v. 92, n. 12, p. 5605-5609, Jun 6 1995.
- OGAWA, T. et al. Molecular diversity and accelerated evolution of C-type lectin-like proteins from snake venom. **Toxicon.**, v. 45, n. 1, p. 1-14, Jan 2005.
- OHNO, S. **Evolution by gene duplication**. Berlin, New York,: Springer-Verlag, 1970. xv, 160 p.
- OLLITRAULT, P. et al. SNP mining in *C. clementina* BAC end sequences; transferability in the Citrus genus (Rutaceae), phylogenetic inferences and perspectives for genetic mapping. **BMC. Genomics.**, v. 13, p. 13, 2012.
- OSIPOV, A. V. et al. Cobra venom contains a pool of cysteine-rich secretory proteins. **Biochem. Biophys. Res. Commun.**, v. 328, n. 1, p. 177-182, Mar 4 2005.
- OSOEGAWA, K. et al. An improved approach for construction of bacterial artificial chromosome libraries. **Genomics.**, v. 52, n. 1, p. 1-8, Aug 15 1998.
- PAIVA, J. A. et al. Advancing Eucalyptus genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage BAC libraries. **BMC. Genomics.**, v. 12, p. 137, 2011.
- PARRA, G.; BRADNAM, K.; KORF, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. **Bioinformatics.**, v. 23, n. 9, p. 1061-1067, May 1 2007.
- REN, L.; REN, X. Meta-analyses of four polymorphisms of lipoprotein lipase associated with the risk of Alzheimer's disease. **Neurosci. Lett.**, v. 619, p. 73-78, Apr 21 2016.
- REPO, H. et al. Is the bovine lysosomal phospholipase B-like protein an amidase? **Proteins.**, v. 82, n. 2, p. 300-311, Feb 2014.
- REYES-VELASCO, J. et al. Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. **Mol. Biol. Evol.**, v. 32, n. 1, p. 173-183, Jan 2015.
- RHOADS, A.; AU, K. F. PacBio Sequencing and Its Applications. **Genomics. Proteomics. Bioinformatics.**, v. 13, n. 5, p. 278-289, Oct 2015.
- RIESENFELD, C. S.; SCHLOSS, P. D.; HANDELSMAN, J. Metagenomics: genomic analysis of microbial communities. **Annu. Rev. Genet.**, v. 38, p. 525-552, 2004.
- RODRIGUES, R. S. et al. Structural and functional properties of Bp-LAAO, a new L-amino acid oxidase isolated from *Bothrops pauloensis* snake venom. **Biochimie.**, v. 91, n. 4, p. 490-501, Apr 2009.
- SANTORO, M. L.; SANO-MARTINS, I. S. Platelet dysfunction during *Bothrops jararaca* snake envenomation in rabbits. **Thromb. Haemost.**, v. 92, n. 2, p. 369-383, Aug 2004.

- SANZ, L.; CALVETE, J. J. Insights into the Evolution of a Snake Venom Multi-Gene Family from the Genomic Organization of *Echis ocellatus* SVMP Genes. **Toxins. (Basel)**, v. 8, n. 7, 2016.
- SANZ, L.; HARRISON, R. A.; CALVETE, J. J. First draft of the genomic organization of a PIII-SVMP gene. **Toxicon**, v. 60, n. 4, p. 455-469, Sep 15 2012.
- SATO, K. et al. 454 sequencing of pooled BAC clones on chromosome 3H of barley. **BMC. Genomics**, v. 12, p. 246, 2011.
- SCHINDELIN, J. et al. The ImageJ ecosystem: An open platform for biomedical image analysis. **Mol. Reprod. Dev.**, v. 82, n. 7-8, p. 518-529, Jul-Aug 2015.
- SERRANO, S. M.; MAROUN, R. C. Snake venom serine proteinases: sequence homology vs. substrate specificity, a paradox to be solved. **Toxicon**, v. 45, n. 8, p. 1115-1132, Jun 15 2005.
- SERRANO, S. M. T. et al. The proteinase-rich proteome of *Bothrops jararaca* venom. **Toxin. Reviews**, v. 33, n. 4, p. 169-184, 2014/12/01 2014.
- SHIRASU, K. et al. A novel class of eukaryotic zinc-binding proteins is required for disease resistance signaling in barley and development in *C. elegans*. **Cell**, v. 99, n. 4, p. 355-366, Nov 12 1999.
- SHIZUYA, H. et al. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. **Proc. Natl. Acad. Sci. U. S. A.**, v. 89, n. 18, p. 8794-8797, Sep 15 1992.
- SILVA, M. B. et al. A prothrombin activator from *Bothrops erythromelas* (jararaca-da-seca) snake venom: characterization and molecular cloning. **Biochem. J.**, v. 369, n. Pt 1, p. 129-139, Jan 1 2003.
- SIMPSON, J. T.; DURBIN, R. Efficient de novo assembly of large genomes using compressed data structures. **Genome. Res.**, v. 22, n. 3, p. 549-556, Mar 2012.
- SIMPSON, J. T. et al. ABySS: a parallel assembler for short read sequence data. **Genome. Res.**, v. 19, n. 6, p. 1117-1123, Jun 2009.
- STANKE, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. **Nucleic. Acids. Res.**, v. 34, n. Web Server issue, p. W435-439, Jul 1 2006a.
- STANKE, M. et al. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. **BMC. Bioinformatics**, v. 7, p. 62, 2006b.
- TAKAHASHI, H. et al. A novel snake venom vascular endothelial growth factor (VEGF) predominantly induces vascular permeability through preferential signaling via VEGF receptor-1. **J. Biol. Chem.**, v. 279, n. 44, p. 46304-46314, Oct 29 2004.
- TARAILO-GRAOVAC, M.; CHEN, N. Using RepeatMasker to identify repetitive elements in genomic sequences. **Curr. Protoc. Bioinformatics**, v. Chapter 4, p. Unit 4 10, Mar 2009.
- TAUDIEN, S. et al. Sequencing of BAC pools by different next generation sequencing platforms and strategies. **BMC. Res. Notes**, v. 4, p. 411, 2011.
- VENTER, J. C.; SMITH, H. O.; HOOD, L. A new strategy for genome sequencing. **Nature**, v. 381, n. 6581, p. 364-366, May 30 1996.
- VIDAL, N. et al. Snakes (Serpentes). **The timetree of life**, 2009.
- VITT, L. J.; CALDWELL, J. P. **Herpetology : an introductory biology of amphibians and reptiles**. Fourth edition. 2014. xiv, 757 pages.
- VONK, F. J. et al. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. **Proc. Natl. Acad. Sci. U. S. A.**, v. 110, n. 51, p. 20651-20656, Dec 17 2013.

- WAN, Q. H. et al. Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. **Cell. Res.**, v. 23, n. 9, p. 1091-1105, Sep 2013.
- WANG, D. Y.; KUMAR, S.; HEDGES, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. **Proc. Biol. Sci.**, v. 266, n. 1415, p. 163-171, Jan 22 1999.
- XU, Q. et al. Cloning of a galactose-binding lectin from the venom of *Trimeresurus stejnegeri*. **Biochem. J.**, v. 341 (Pt 3), p. 733-737, Aug 1 1999.
- YAMAGUCHI, K. et al. The finding of a group IIE phospholipase A2 gene in a specified segment of *Protobothrops flavoviridis* genome and its possible evolutionary relationship to group IIA phospholipase A2 genes. **Toxins. (Basel)**, v. 6, n. 12, p. 3471-3487, 2014.
- YAMAMOTO, T. et al. Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. **BMC. Genomics.**, v. 11, p. 267, 2010.
- YAMANOUYE, N. et al. Control of venom production and secretion by sympathetic outflow in the snake *Bothrops jararaca*. **J. Exp. Biol.**, v. 200, n. Pt 19, p. 2547-2556, Oct 1997.
- YAMAZAKI, Y. et al. Snake venom Vascular Endothelial Growth Factors (VEGF-Fs) exclusively vary their structures and functions among species. **J. Biol. Chem.**, v. 284, n. 15, p. 9885-9891, Apr 10 2009.
- YAMAZAKI, Y.; MORITA, T. Structure and function of snake venom cysteine-rich secretory proteins. **Toxicon.**, v. 44, n. 3, p. 227-231, Sep 1 2004.
- YUKSEL, B.; PATERSON, A. H. Construction and characterization of a peanut HindIII BAC library. **Theor. Appl. Genet.**, v. 111, n. 4, p. 630-639, Aug 2005.
- ZELANIS, A. et al. *Bothrops jararaca* venom proteome rearrangement upon neonate to adult transition. **Proteomics.**, v. 11, n. 21, p. 4218-4228, Nov 2011.
- ZELANIS, A. et al. Analysis of the ontogenetic variation in the venom proteome/peptidome of *Bothrops jararaca* reveals different strategies to deal with prey. **J. Proteome. Res.**, v. 9, n. 5, p. 2278-2291, May 7 2010.
- ZHANG, G. et al. The oyster genome reveals stress adaptation and complexity of shell formation. **Nature.**, v. 490, n. 7418, p. 49-54, Oct 4 2012.
- ZHANG, X.; LIU, F. Increased risk of pneumonia associated with angiotensin-converting enzyme (CD143) rs4340 polymorphism. **Clin. Exp. Med.**, v. 16, n. 3, p. 423-428, Aug 2016.
- ZHAO, Z. X. et al. Duplication and differentiation of common carp (*Cyprinus carpio*) myoglobin genes revealed by BAC analysis. **Gene.**, v. 548, n. 2, p. 210-216, Sep 15 2014.
- ZIMIN, A. et al. Sequencing and assembly of the 22-gb loblolly pine genome. **Genetics.**, v. 196, n. 3, p. 875-890, Mar 2014.