Universidade de São Paulo
Bruno Ferreira de Souza

Análise exploratória de genes compostos predominantemente por microexons
arranjados em tandem em uma coleção de anotações de genomas

São Paulo
2016

Universidade de São Paulo
Bruno Ferreira de Souza

Análise exploratória de genes compostos predominantemente por microexons
arranjados em tandem em uma coleção de anotações de genomas

Orientador: Prof. Dr. Sergio Verjovski-Almeida
Coorientador: Prof. Dr. João Carlos Setubal

Dissertação apresentada ao Instituto de Matemática e Estatística da Universidade de São Paulo, para a obtenção de Título de Mestre em Ciências, na Área de Concentração em Bioinformática.

São Paulo
2016

Resumo

Genes microexônicos (MEGs) apresentam uma arquitetura bastante peculiar, sendo compostos predominantemente por quatro ou mais exons simétricos bem pequenos dispostos em tandem (microexons ≤ 36 bp, com tamanhos múltiplos de 3). Eles foram descritos pela primeira vez em 2009 no platelminto parasita *Schistosoma mansoni* (agente etiológico da esquistossomose). Alguns desses genes apresentam evidência (de transcritômica e proteômica) de gerar diferentes isoformas de proteínas através do mecanismo de *splicing* alternativo. Os MEGs não apresentam homólogos fora do gênero *Schistosoma*. Microexons individuais já foram reportados em genes de organismos modelo e do ser humano, normalmente atuando como um *hotspot* de *splicing* alternativo, porém nestes casos apenas um único microexon por gene fora reportado. Com este pano de fundo, fizemos o seguinte questionamento: existem genes com arquitetura similar (ou seja, múltiplos microexons em tandem) em outros organismos? Desenvolvemos uma heurística capaz de detectar genes com a arquitetura dos MEGs e a aplicamos no transcrito de *S. mansoni* e, além de detectar os MEGs originais, detectamos 31 novos genes. Este pipeline foi aplicado a uma coleção de anotações de genomas e detectou 494 genes distribuídos entre 125 organismos (incluindo animais, plantas, fungos e alguns eucariotos unicelulares).

Palavras-chave: microexon, microexon gene, MEG.

Abstract

Microexon genes (MEGs) have an unusual architecture, composed predominantly by four or more very small tandemly disposed symmetric exons (microexons ≤ 36 bp, with exon sizes multiple of 3). They were first described in 2009 in the parasitic platyhelminth *Schistosoma mansoni* (etiologic agent of schistosomiasis). Some of those genes display evidence (by transcriptomics and proteomics) of generating variable protein isoforms through alternative splicing. MEGs have no homologs outside the *Schistosoma* genus. The presence of individual microexons has also been reported in genes of model organisms and humans, usually as a hotspot of alternative splicing, but in those cases only a single microexon per gene was observed. Within this background, we asked the following question: are there genes with similar architecture (i.e., with multiple internal microexons in tandem) in other organisms? We developed a pipeline to detect genes with the architecture of *S. mansoni* MEGs, applied it on a updated transcriptome of *S. mansoni* and successfully detected the original MEGs and 31 new genes. This pipeline was applied it on a collection of genome annotations and detected 594 genes distributed among 125 organisms (including animals, plants, fungi, and some unicellular eukaryotes).

Keywords: microexon, microexon gene, MEG.

Sumário

1	Fundamentação teórica.....	2
1.1	Fluxo da informação genética.....	2
1.2	Complexidade e composição dos genomas eucarióticos.....	2
1.3	O <i>splicing</i> do mRNA.....	3
1.4	Influência do tamanho do exon no <i>splicing</i>	5
1.5	Genes microexônicos.....	6
1.6	Microexons.....	7
1.7	Motivação e justificativa.....	7
1.8	Objetivos alcançados.....	8
2	Materiais e Métodos.....	9
2.1	Materiais.....	9
2.2	Bancos de anotações públicos.....	9
2.3	Dados de transcrição de <i>Schistosoma mansoni</i>	9
2.4	Metodologia.....	10
2.4.1	Mapeamento de transcriptoma.....	10
2.4.2	Escolha de melhor anotação para um transcrito.....	11
2.4.3	Busca de TaMEGs.....	11
2.4.4	Agrupamento em famílias.....	12
2.4.5	Busca de assinaturas funcionais.....	13
2.4.6	Construção de árvores filogenéticas.....	13
3	Resultados e Discussão.....	14
3.1	Resultados metodológicos.....	14
3.1.1	Breve comparação do desempenho de alinhadores.....	14
3.2	Microexons e sua abundância em <i>S. mansoni</i> e organismos modelo.....	16
3.2.1	Breves considerações sobre as anotações de <i>S. mansoni</i>	16
3.3	Definições: TaMEG e parâmetros para sua busca.....	18
3.4	Observações sobre os MEGs previamente descritos.....	19
3.5	Busca de TaMEGs no transcrito de <i>S. mansoni</i>	21
3.6	TaMEGs detectados na coleção de anotações de genomas.....	22
3.6.1	Genes e famílias das espécies com maior abundância de TaMEGs.....	22

3.6.2	Famílias TaMEG de Cordados	24
3.7	Contribuição de μ Ex à ORF	25
4	Conclusões	25
5	Bibliografia	25
6	Figuras e Tabelas	29
6.1	Figuras	29
6.2	Tabelas	34

Índice de Figuras

Figura 1 - Splicing.....	4
Figura 2 – Distribuição de tamanho de exons internos.....	5
Figura 3 – Arquitetura dos exons dos MEGs de <i>S. mansoni</i>	6
Figura 4 – Pipeline de identificação de TaMEGs.	10
Figura 5 – Modelos do MEG-19.....	16
Figura 6 – Distribuição de exons e introns em <i>S. mansoni</i>	17
Figura 7 – MEGs com anotação “piorada” na nova versão do genoma	29
Figura 8 – Gene Models dos MEGs sem SMP.....	29
Figura 9 – MEG-1.2	30
Figura 10 – MEG-21.....	30
Figura 11 – Cluster de MEG-18 no cromossomo 2	30
Figura 12 – Destaque do alinhamento múltiplo do cluster SmFAM_02.	31
Figura 13 – Comparação entre os MEGs originais e os TaMEGs detectados.....	32
Figura 14 – Comparação entre os MEGs reportados em cada publicação.	33

Índice de Tabelas

Tabela 1 – Bancos de anotações utilizados.....	9
Tabela 2 – Desempenho dos alinhadores.	14
Tabela 3 – Estatística descritiva de exons e introns da anotação do genoma de <i>S. mansoni</i>	17
Tabela 4 – MEGs com código Smp.	34
Tabela 5 – Composição dos exons internos de MEGs representativos	35
Tabela 6 – Sobreposição do MEG-17 e MEG-25.	35
Tabela 7 – Contagem de genes detectados por grupos taxonômicos.....	36
Tabela 8 – Contagem de genes detectados por espécie.....	36
Tabela 9 – Loci do MEG-18	39
Tabela 10 – Famílias de TaMEGs de <i>S. mansoni</i>	40
Tabela 11 – Famílias da coleção de genomas.....	41

1 Fundamentação teórica

1.1 Fluxo da informação genética

O **Dogma Central da Biologia Molecular** (Crick 1970), estabelece o fluxo das informações genéticas nos seres vivos: as informações contidas no **DNA** são decodificadas como **RNA (transcrição)**; os **RNAs mensageiros (mRNA)**, são decodificados como **proteínas (tradução)**.

A **transcrição** é catalisada pelas **RNA polimerases**, grandes complexos enzimáticos capazes de produzir uma sequência de RNA a partir de um molde de DNA. A RNA polimerase *per se* não é capaz de reconhecer onde se iniciam os genes no genoma, sendo dependente dos **fatores de transcrição**, proteínas que reconhecem sinais de início da transcrição.

Ao ser recrutada por um fator de transcrição, a RNA polimerase avança da região 3' a 5' do molde de DNA enquanto sintetiza uma molécula de RNA complementar ao molde. Ao encontrar uma sequência de terminação, o complexo da RNA polimerase é desacoplado e a transcrição é concluída. No caso do mRNA de eucariotos, o transcrito ainda sofre três processamentos durante sua maturação: adição do *cap* de 7-metilguanossina à região 5', adição da cauda poli-A à região 3' e remoção de introns, sequências não codificadoras localizadas no interior de genes. Este último processo é denominado *splicing* e é o tema da sessão 1.3.

Na **tradução** o mRNA é utilizado como molde para a síntese de uma cadeia polipeptídica. A síntese proteica é catalisada pelos ribossomos, grandes complexos de RNA (**rRNA**) e proteínas. Um código de três bases no mRNA (o códon) especifica qual o aminoácido equivalente de acordo com o código genético (quase) universal. A região do mRNA que contém a sequência de códons que codificam o polipeptídeo é denominada fase aberta de leitura (**ORF**, do inglês *open reading frame*). A transferência da informação de nucleotídeos para aminoácidos é mediada por um tipo especial de RNA, o RNA transportador (**tRNA**).

1.2 Complexidade e composição dos genomas eucarióticos

Um genoma carrega toda a informação hereditária de um organismo, e essa informação é codificada em moléculas de DNA. No interior das células, o DNA é armazenado de forma compacta: enovelado e associado a histonas, formando as fibras de cromatina. Durante a divisão celular, fibras de cromatina formam estruturas ainda mais condensadas, os cromossomos. Embora não sejam sinônimos, cromatina e cromossomo são usualmente tratados como tal nas ciências genômicas.

Os *loci* cromossômicos que carregam informações (expressas como RNAs) são denominados **genes** e podem ser divididos em dois grandes grupos: genes codificadores, que codificam proteínas, e genes não-codificadores.

De acordo com Koonin (2009), genomas podem ser classificados em dois grandes grupos, de acordo com sua composição e complexidade:

- 1) **Genomas compactos e pequenos** de vírus, arqueias, bactérias (< 10 Mb) e eucariotos unicelulares (< 20 Mb), compostos principalmente por genes codificadores.
- 2) **Genomas extensos e complexos** de eucariotos multicelulares e alguns unicelulares (> 100 Mb), composto majoritariamente por genes não codificadores, elementos de transposição, sequências de baixa complexidade e sequências espaçadoras.

A título de comparação, o genoma da bactéria *Escherichia coli* consiste em um único cromossomo circular de 4,6 Mb que contém cerca de 5 mil genes, enquanto o genoma do homem consiste em 23 pares de cromossomos, que somam 3,2 Gb e contém cerca de 20 mil genes. Os tamanhos e conteúdo dos genomas mencionados foram verificados nos bancos de dados Ensembl e Ensembl Genomes (Hubbard et al. 2002; Kersey et al. 2012).

Embora o genoma de humanos seja cerca de mil vezes maior que o da *E. coli*, o número de genes é apenas quatro vezes maior. Essa discrepância é explicada pela quantidade de sequências não codificadoras, que compreendem mais de 98% do genoma humano e apenas 12% em *E. coli* (Rogozin et al. 2002).

A arquitetura dos genes também é diferente entre os dois grandes grupos de genomas. Conforme mencionado anteriormente, eucariotos apresentam genes cuja arquitetura é dividida em regiões que de fato estarão no mRNA maduro, os **exons** (do inglês *expressed region*), e regiões que serão removidas, os **introns** (do inglês *intragenic region*). Em genomas complexos, introns são maiores que os exons e contribuem significativamente no tamanho do genoma. Já em genomas compactos, introns (quando presentes) são raros e curtos (Koonin 2009).

1.3 O *splicing* do mRNA

O *splicing*, um passo fundamental na expressão gênica em eucariotos, consiste na remoção de introns do pré-mRNA bem como na união dos exons que serão mantidos no mRNA maduro, o molde usado para a tradução em proteína.

Introns apresentam sinais cruciais para que o *splicing* ocorra de maneira adequada. Esses sinais são os **sítios de *splicing* (SS) 5' e 3'**, o **branch point (BP)**, proximal ao 3' SS, e o **trato de polipirimidina (PPT)**, uma região rica em nucleotídeos pirimídicos, C e U), posicionado entre o 3' SS e BP (Figura 1a).

O *splicing* é catalisado pelo spliceossomo, um grande complexo formado por diversas proteínas e cinco **RNAs nucleares pequenos (snRNA)**, U1, U2, U4, U5 e U6 (Figura 1b). Os snRNAs são responsáveis pelo reconhecimento dos SS pela atividade catalítica do spliceossomo. Associados a proteínas, os snRNAs formam as **pequenas**

partículas nucleares de ribonucleoproteínas (snRNP). Durante a montagem do spliceossomo os snRNAs sofrem rearranjos responsáveis por ativar o complexo (Staley and Guthrie 1998).

Figura 1 - Splicing

- a) arquitetura típica do intron;
- b) montagem do spliceossomo. Fonte: (Keren et al. 2010)

A montagem do spliceossomo tem início com o reconhecimento do 5' SS e BP pelos snRNPs U1 e U2, respectivamente. Em seguida, um complexo pré-formado contendo U4/U6 e U5 é recrutado. U5 liga-se ao 5' SS e U4 e U1 são liberados. O próximo passo é uma segunda reação de transesterificação entre o BP e 5' SS que resulta na clivagem do 5' SS e formação de um intermediário em laço. Finalmente U5 liga-se ao sítio 3' SS e uma segunda reação de transesterificação resulta na junção dos exons e liberação do intron contendo o laço (Staley and Guthrie 1998).

Sequências degeneradas intrônicas e exônicas influenciam o *splicing*, atuando como *enhancers* ou *silencers* que recrutam proteínas específicas e regulam a inclusão de um exon no transcrito final, como as proteínas SR (serina/arginina) (Berget 1995; Keren et al. 2010).

O modelo descrito acima explica o splicing de exons constitutivamente expressos. No entanto, muitas variações podem ocorrer devido à interferência de fatores que regulam o *splicing*. Essa regulação pode levar ao reconhecimento de sítios 5' SS

ou 3' SS alternativos, retenção de introns curtos e exclusão/inclusão alternativa de exons (Keren et al. 2010), processo denominado *splicing* alternativo.

1.4 Influência do tamanho do exon no *splicing*

Exons constitutivamente expressos apresentam uma distribuição simétrica em torno de um tamanho ótimo, no caso de mamíferos ~ 140 bp (Berget 1995), enquanto que exons alternativos tendem a ser mais curtos (Figura 2). Essa limitação de tamanho sugere que a maquinaria de *splicing* tem preferência para reconhecer exons de certo tamanho e que há uma pressão seletiva atuando sobre o tamanho de exons, garantindo a manutenção desses tamanhos (Berget 1995).

Figura 2 – Distribuição de tamanho de exons internos.

Painel superior: exons constitutivos; Painel inferior: exons diferencialmente expressos. Fonte: (Berget 1995)

Exons curtos (≤ 51 bp) são menos comuns ($< 6\%$ em mamíferos) (Li et al. 2015), sua inclusão é mais dependente de *enhancers* intrônicos (Carlo et al. 1996) e normalmente é limitada a tecidos específicos (Holter et al. 2005; Frank-Hansen et al. 2008; Ovadia and Shifman 2011; Irimia et al. 2014; Li et al. 2015). Por esses motivos, microexons são considerados *hotspots* para a ocorrência do *splicing* alternativo (McAllister et al. 1992; Carlo et al. 1996; Holter et al. 2005; Li et al. 2015).

1.5 Genes microexônicos

Cerca de 40 genes ricos em exons internos ≤ 36 bp *in tandem* (ou seja, dispostos lado a lado) foram identificados no genoma do platelminto trematódeo *Schistosoma mansoni*. A detecção desses genes somente foi possível a partir do mapeamento de transcritos no genoma, uma vez que exons tão curtos não tem um sinal significativo para preditores *ab initio*. Dada sua rica composição de exons curtos estes genes foram denominados Genes microexônicos (MEGs, *microexon genes* em inglês) (Berriman et al. 2009).

MEGs são compostos predominantemente por exons simétricos (múltiplos de três), curtos (6 a 36 bp, com 'outliers' de até 61 bp) e dispostos em tandem (Figura 3). A maioria deles também apresenta peptídeo sinal e/ou domínios transmembrana, mas nenhum domínio funcional já descrito. Atualmente conhecem-se em *S. mansoni* 32 famílias de MEGs (DeMarco et al. 2010; Almeida et al. 2012; Wilson et al. 2015) e nenhuma delas apresenta similaridade com quaisquer genes descritos fora do gênero *Schistosoma*.

Figura 3 – Arquitetura dos exons dos MEGs de *S. mansoni*.

Polígonos indicam exons que compõem peptídeo sinal (triângulo) e domínio transmembrana (losango). O tamanho dos exons está indicado. Código de cores: μ Ex de comprimento divisível por 3 (vermelho) ou não (amarelo); μ Ex de 37 a 61 nt de comprimento divisível por 3 (azul) ou não (verde); regiões terminais da ORF (cinza) e UTR (preto). Introns estão fora de escala.

Fonte: (Berriman, et al., 2009)

Recentemente foram detectados 27 transcritos (pertencentes a 22 famílias dos MEGs) expressos predominantemente na cabeça do parasita (Wilson et al. 2015).

Posteriormente MEGs foram identificados em outros platelmintos, mas da classe cestoda: *Echinococcus multilocularis* (9 genes) e *E. granulosus* (5 genes) (Tsai et al. 2013). Nenhum deles apresenta similaridade de sequência com os MEGs de *S. mansoni*.

1.6 Microexons

Ainda não existe uma definição de microexon consolidada na literatura. O primeiro uso do termo que pudemos detectar na literatura foi usado para se referir a dois exons de 51 nt presentes no gene *Ubx*, da mosca da fruta (Beachy et al. 1985). Uma definição mais conservadora (≤ 25 nt) foi usada na descrição da primeira metodologia desenvolvida para detecção de microexons (Volfovsky et al. 2003). No contexto dos genes microexônicos de *S. mansoni*, exons ≤ 36 nt são considerados “verdadeiros” microexons, enquanto que os de 37 a 61 nt são considerados “outliers” (Berriman et al. 2009).

Do ponto de vista da **biologia molecular**, podemos dizer que microexons são exons tão curtos que escapam da faixa de tamanho reconhecida com eficiência pelo spliceossomo (50 a 300 nt), sendo, então, mais propensos a sofrer *exon skipping*. Daí viria uma definição com limite superior mais nebuloso, que deveria ser analisada caso a caso (qual o tamanho ótimo para cada organismo?). Caso a distribuição de tamanho de exons seja similar entre duas espécies, podemos presumir que a amplitude de tamanhos reconhecida pelo spliceossomo também o seja.

Provavelmente a definição mais conservadora, do artigo do grupo de Steve Salzberg (Volfovsky et al. 2003), apenas levava em consideração apenas a dificuldade em se alinhar exons ≤ 25 nt. Seria, então, uma definição do **ponto de vista computacional**.

Feitas essas considerações, adotamos a seguinte definição neste trabalho: ao usar o termo **μEx** (microexon), caso o tamanho não seja especificado, estaremos nos referindo a exons com **36 nt ou menos**. Caso contrário, especificaremos o tamanho.

1.7 Motivação e justificativa

Embora genes com arquitetura similar à dos MEGs de *S. mansoni* tenham sido descritos em outras espécies relativamente próximas (Cestoda) (Tsai et al. 2013), não existe semelhança entre a sequência deles. Assim, a fim de se investigar a prevalência real de genes compostos por múltiplos microexons em tandem, buscas por similaridade de sequência não são suficientes. É preciso buscar pela própria arquitetura desse tipo de gene.

No entanto, não existem ferramentas para esse tipo de busca. A identificação dos MEGs de *S. mansoni* foi feita manualmente, por inspeção dos alinhamentos de

transcritos ao genoma. O presente trabalho preenche essa lacuna com um script que busca genes de arquitetura similar em anotações de genomas (como outputs de programas de mapeamento de rna ao genoma ou bancos de dados de anotações, como o ensembl).

Recentemente foi feita a caracterização de um grande número de microexons que são expressos exclusivamente no tecido nervoso (Irimia et al. 2014; Li et al. 2015). No entanto, nenhum dos exemplos destacados nestes trabalhos apresenta mais de um microexon. Assim, esta questão (existem genes com arquitetura de exons internos semelhante aos MEGs em outros organismos além de platelmintos parasitas?) continua em aberto.

1.8 Objetivos alcançados

- Desenvolvemos uma metodologia que reduz a necessidade de refinamentos manuais nas anotações de MEGs. Esta metodologia consiste num pipeline que escolhe a “melhor” anotação de um transcrito no genoma. Ela foi testada somente no contexto dos MEGs de *S. mansoni*, mas poderia ser aplicada em outros casos.
- Desenvolvemos uma biblioteca (em python) para lidar com anotações genômicas. A utilizamos fazer conversões entre formatos de anotações (GFF3, GTF e Bed) e filtragens por quantidade de exons, tamanho de introns. No entanto, a biblioteca não está limitada a estes usos. Ela pode ser importada em qualquer script python para um uso mais avançado de suas funções e classes.
- Desenvolvemos um scripts para detecção e classificação de genes com composição de μ Ex em tandem similar à dos MEGs que reproduzem os resultados previamente descritos na literatura.
- Mapeamos dados de RNA-seq e ESTs para gerar uma nova anotação do genoma de *S. mansoni* e, com as ferramentas desenvolvidas neste trabalho, detectamos novos MEGs em *S. mansoni*.
- Usando nossa ferramenta, detectamos a existência de genes com composição de exons internos similar à dos MEGs em uma coleção de anotações de genomas.

2 Materiais e Métodos

Neste trabalho foram utilizados dados de anotação genômica de bancos públicos (Ensembl e Ensembl Genomes) e anotações de *Schistosoma mansoni* obtidas à partir de dados de RNA-Seq públicos, dados gerados pelo Lab. de Expressão Gênica em Eucariotos do Instituto Butantan e bancos de EST públicos.

2.1 Materiais

2.2 Bancos de anotações públicos

Para compor a coleção de genomas analisada no trabalho, foram usados dados de anotação de genoma (gff3/gtf) e arquivos fasta contendo o proteoma desses organismos. O número de genes presentes em cada coleção, de onde veio e qual sua versão estão sumarizados na Tabela 1.

Tabela 1 – Bancos de anotações utilizados.

O número de genes contabiliza apenas genes anotados como codificadores de proteínas ou, na ausência desta informação, que apresente CDS.

Fonte	Versão	# espécies	# genes
Ensembl (vertebrados)	76	65	1201428
Ensembl Metazoa (invertebrados)	25	55	1016711
Ensembl Plants	25	38	1321745
Ensembl Fungi	25	53	579358
Ensembl Protists	25	32	460577
Total	-	243	4579819

2.3 Dados de transcrição de *Schistosoma mansoni*

Foram usados dois conjuntos de dados de *Schistosoma mansoni* processados no Laboratório de Expressão Gênica em Eucariotos do Instituto Butantan. Um dos conjuntos consiste de dados públicos e o outro foi de dados do laboratório ainda não publicados.

Os dados públicos consistem em bibliotecas de *reads* de RNA-seq obtidas no SRA (*Sequence Read Archive* do NCBI). Esses dados compreendem fases larvais e adulta do ciclo de vida do verme e foram sequenciados em diversas plataformas, incluindo Illumina e Roche 454.

Os dados privados foram obtidos de vermes adultos de *S. mansoni* submetidos a tratamentos com drogas. Os *reads* deste conjunto de dados foram produzidos na plataforma Illumina HiSeq 2000 e são fita específicos. Esses dados fazem parte do projeto A-ParaDDisE (*Anti Parasitic Drug Discovery in Epigenetics*) e, como ainda não

foram publicados, somente os reads das amostras controle, sem droga, foram usados para as análises deste trabalho.

Os reads foram submetidos a um pipeline de montagem de transcriptoma *de novo* com o montador Trinity. Os reads de dados públicos foram montados juntos, num “sopão” de reads do SRA. Já os reads controle do A-ParaDDisE foram montados à parte, uma montagem por tratamento. Estas montagens feitas pelo Dr. Elton Vasconcelos, pós-doutorando do grupo do Dr. Sergio Verjovski-Almeida.

Além destes dados oriundos de RNA-seq, também utilizamos ESTs de *S. mansoni* depositados em bancos de dados públicos. Estes foram obtidos do nucest, do NCBI, através de um *script ad hoc* e os dados no ftp do Instituto Sanger.

2.4 Metodologia

Figura 4 – Pipeline de identificação de TaMEGs.

2.4.1 Mapeamento de transcriptoma

Arquivos fasta dos contigs montados pelo trinity e das bibliotecas de ESTs foram mapeados ao genoma de *S. mansoni* com o gmap.

Transcritos de interesse (por exemplo, aqueles cuja anotação for selecionada pelo **muExTandemFinder.py**) tiveram sua sequencia original mapeada também com Blat e Splign. Em seguida, escolhemos o mapeamento cuja sequencia recuperada do genoma é mais similar à sequencia do transcrito original. Essa metodologia está descrita a seguir.

2.4.2 Escolha de melhor anotação para um transcrito

Chamemos a sequência original de um transcrito usada em um mapeamento de **Seq₀**. Para cada mapeamento *i* feito de **Seq₀** ao genoma, **Seq_i** é a sequência genômica e sem introns deste mapeamento. Cada mapeamento *i* é convertido para bed (com auxílio de scripts em bash para converter o output do Blat, Gmap e Splign para bed) e **Seq_i** é recuperada com auxílio do programa **getfasta**, do pacote bedtools.

Para cada mapeamento *i*, comparamos a similaridade de **Seq_i** com **Seq₀** (por Blastn). Também comparamos **Seq₀** com ela mesma, a fim de saber qual o maior bitscore possível nas comparações com **Seq₀** e usamos este valor para normalizar os bitscores das demais mapeamentos *i* > 0.

O mapeamento *i* com maior bitscore é considerado o melhor e é selecionado.

Esta metodologia foi implementada parte em bash (script **bitscore4eachMapping.sh**, que “chama” o blastp e gera os bitscores para cada mapeamento) e python (script **FilterBestMappings.py**, que tabula os resultados do blastn e escolhe os mapeamentos com maior bitscore).

2.4.3 Busca de TaMEGs

A busca de genes com arquitetura similar à dos MEGs foi feita com o script **muExTandemFinder.py**, desenvolvido neste trabalho. Este script busca em arquivos BED12 (ou qualquer arquivo TSV que tenha um campo equivalente ao “BlockSizes” do BED12) transcritos que apresentem composição de microexons internos em tandem similar à dos MEGs de *S. mansoni*. Essa composição foi definida à partir da análise dos MEGs previamente descritos e foi implementada na função “isTaMEG”, que tem as seguintes exigências:

1. Apresentar no mínimo 3 μ Ex em tandem;
2. Apresentar exons internos com mediana ≤ 36 nt;
3. Satisfazer a seguinte condição:

$$\begin{aligned} \mu ExT_{max} &= iEx, & \text{se } 3 \leq iEx < 5 \\ \frac{\mu ExT_{max}}{iEx} &> 0.75, & \text{se } iEx \geq 5 \end{aligned}$$

Sendo $\mu\text{ExT}_{\max}$ a maior contagem de $\mu\text{Ex} \leq 61$ em tandem e $i\text{Ex}$ o total de exons internos. Como exons terminais tendem a ser grandes, o script contabiliza exons terminais ≤ 61 nt como exons internos.

Por padrão, a contagem $\mu\text{ExT}_{\max}$ tem tolerância a **um exon > 61nt**. O parâmetro **--big_tolerance** (ou **-b**, argumento **N**, um número inteiro não negativo) modifica o comportamento da contagem de μEx , que passa a tolerar até **N** exons > 61 nt. Note que exon > 61 nt não é contabilizado como μEx . Adicionamos essa tolerância para casos em que um microexon apresente sítios de splicing alternativos, fazendo com que o mesmo seja expresso como um exon > 61 nt em certas condições.

Não há exigência de um número mínimo de exons simétricos na definição *hardcoded*, mas é possível exigir isso com os seguintes parâmetros:

- **--x3_ratio** (ou **-x3**, cujo argumento é um *float* de zero a 1): permite exigir uma razão mínima de exons internos simétricos;
- **--symmetric** (ou **-s**, sem argumento): exige que apenas μEx simétricos sejam contabilizados.

Além disso, o script tem um parâmetro **--invert_match** (ou **-v**, sem argumento), que modifica o comportamento do programa, retornando apenas os resultados que não satisfazem os parâmetros da busca.

2.4.4 Agrupamento em famílias

O agrupamento de MEGs em famílias é feito à partir da similaridade da sequência de aminoácidos de suas respectivas ORFs. No artigo original, os MEGs foram submetidos a um blastp e hits com *evaluate* inferior a $1e-4$ foram agrupados em famílias.

Reproduzimos esta metodologia fazendo um blastp com esse cutoff $1e-4$ e submetendo o resultado do blastp (em formato "m8") para o script python `blast2fam.py`, que faz o agrupamento dos hits.

Além do cutoff de *evaluate* usado no blast, o script `blast2fam.py` usa dois cutoffs adicionais, **tamanho do alinhamento ≥ 20** e **bitscore ≥ 30** (estes cutoffs adicionais reproduzem a classificação original dos MEGs). Na classificação dos genes detectados no transcriptoma de *S. mansoni*, usamos estes mesmos cutoff. Já para a coleção de genomas, aumentamos o bitscore mínimo para 40.

Adotando *a* e *b* como *query* e *subject* (ou *subject* e *query*, pois nesse caso a ordem não importa) de um dado *hit* da saída do blastp e definindo famílias como conjuntos de genes, temos que o algoritmo de classificação consiste nos seguintes passos:

- 1) Checar se o *hit* satisfaz os *cutoffs*; em caso negativo, avaliar o próximo hit.
- 2) Agrupar os genes *a* e *b*, de acordo com dos seguintes cenários:

- a. Se $a \in \emptyset$ e $b \in \emptyset$, criar uma família $A = \{a, b\}$.
 - b. Se $a \in A$ e $b \in \emptyset$, incluir b em A , $A = A \cup \{b\}$.
 - c. Se $a \in A$ e $b \in B$, criar uma nova família $C = A \cup B$ (e em seguida descartar as famílias A e B).
- 3) Após processar todos os *hits*, retornar as famílias e seu conteúdo.

O resultado final consiste num arquivo de texto com uma família por linha, em ordem decrescente de seus respectivos tamanhos (quantidade de genes).

2.4.5 Busca de assinaturas funcionais

Para fazer a busca de assinaturas de domínios funcionais nos genes detectados, submetemos sua sequência de aminoácidos à análise no *interproscan*

2.4.6 Construção de árvores filogenéticas

Para famílias de interesse, usamos a sequência codificadora de seus genes para gerar alinhamentos múltiplos (AM) e árvores filogenéticas (AF). Para tanto, usamos o *workflow* “standard phyl” do *etetoolkit* (Huerta-Cepas et al. 2016). Neste *workflow*, o Clustal Omega (Sievers et al. 2011) gera o AM e o Phylml (Guindon et al. 2010) usa este AM para construção da árvore.

Após geração do AF, os AM foram ordenados da mesma forma que o AF.

3 Resultados e Discussão

3.1 Resultados metodológicos

3.1.1 Breve comparação do desempenho de alinhadores

Comparamos o desempenho dos alinhadores blat, gmap e splign no alinhamento da sequência dos 50 MEGs conhecidos de *S. mansoni* à versão v5.2 de seu genoma. Para mensurar o quão “corretos” estão esses alinhamentos, comparamos a da sequência genômica extraída à partir dos alinhamentos produzidos com sua sequência original por blast, resultado sumarizado na Tabela 2.

Tabela 2 – Desempenho dos alinhadores.

O bitscore corresponde ao bitscore da sequência original comparada com ela mesma (por blast). As demais colunas correspondem ao alinhador usado. Para facilitar a visualização e comparação, as demais colunas foram normalizadas pelo valor da coluna “bitscore”.

ID	bitscore	splign	gmap	blat
MEG-1.2	1024	0,956	0,958	0,926
MEG-1_Smp_122630	1075	0,99	0,99	0,985
MEG-2_Smp_159800	715	0	0,792	0,31
MEG-2_Smp_159810	787	0	0,82	0,545
MEG-2_Smp_159830	715	0	0,319	0,29
MEG-2_Smp_179860	636	0	0,335	0,286
MEG-2_Smp_180310	444	0,986	0,986	0,322
MEG-2_Smp_180320.1	1116	0,935	0,935	0,935
MEG-2_Smp_180340	588	1	0,759	1
MEG-2_Smp_181510.1	477	0,99	0,99	0,99
MEG-2_gb EX495163.1	634	0	0,336	0,287
MEG-3_Smp_138060	1134	0,99	0,99	0,99
MEG-3_Smp_138070	1236	0,795	0,891	0,707
MEG-3_Smp_138080	1142	0,89	0,496	0,652
MEG-3_Smp_138090	1225	0,984	0,984	0,984
MEG-4_Smp_085840.1	795	1	1	1
MEG-4_Smp_163630	1288	0,494	0,577	0,494
MEG-5_Smp_152580	856	0,965	0,641	0,959
MEG-6_Smp_163710	743	0	0,692	0,354
MEG-7_Smp_165050	977	0,97	0,97	0,97
MEG-8.3	1332	0	0,268	0,268
MEG-8.4	760	1	1	0,975
MEG-8_Smp_171190	1142	0,987	0,987	0,877
MEG-8_Smp_172180	992	0,971	0,971	0,971
MEG-9_Smp_125320	628	0,975	0,975	0,954
MEG-10.2	361	1	1	1
MEG-10_Smp_152590	418	1	1	1

ID	bitscore	splign	gmap	blat
MEG-11_Smp_176020	793	0,731	0,758	0,441
MEG-12_Smp_152630	514	0,965	0,965	0,965
MEG-13_Smp_127990	909	1	1	1
MEG-14_Smp_124000	819	0,974	0,957	0,974
MEG-15_Smp_010550	1400	0,771	0,771	0,771
MEG-16_Smp_158890	800	0,994	0,994	0,994
MEG-17_Smp_180620	577	0,695	0,932	0,546
MEG-18_Smp_032560	632	0	0,34	0,407
MEG-19	713	0,985	0,985	0,993
MEG-20	381	0,811	0,811	0,811
MEG-21	756	0,971	0,971	0,963
MEG-22	935	0,706	0,736	0,62
MEG-23	510	1	0,99	1
MEG-24	732	0,985	0,985	0,985
MEG-25	586	0,99	0,971	0,99
MEG-26	411	1	1	1
MEG-27	311	0,994	0,994	0,868
MEG-28	372	1	1	1
MEG-29	388	1	1	1
MEG-30	327	1	1	0,878
MEG-31	355	1	1	0,476
MEG-32.1	438	1	1	0,735
MEG-32.2	499	1	1	0,846

Apenas em 7 casos os três alinhadores tiveram bitscore máximo. Em outros 11 casos os três empatam com o mesmo valor, porém abaixo do máximo.

Comparando em pares, gmap e splign apresentaram melhor desempenho, superando o blat em 17 casos. Splign e blat superam o gmap em 6 hits e gmap + blat superam o splign em 3 hits.

Individualmente, o blat supera os outros alinhadores em apenas dois casos, MEG-18 e MEG-19. O gmap tem o melhor desempenho, com maior bitscore em 12 dos MEGs. O splign, finalmente, tem desempenho melhor em apenas 3 casos. O splign também tem uma taxa de identificação menor: ele falha em alinhar 8 dos MEGs (bitscore 0), enquanto que blat e gmap conseguiram identificar alinhar todos.

Nos dois casos em que o blat tem melhor desempenho, MEG-18 e MEG-19, avaliamos que o gmap produziu um modelo melhor, apesar do bitscore menor. No caso do MEG-18, o gmap produz um modelo com cobertura maior, porém com mais substituições (e, conseqüentemente, reduzindo o bitscore do alinhamento). Já no caso do MEG-19, os modelos produzidos pelo gmap e splign concordam em remover 2

nucleotídeos do início do último exon pois ele corresponde ao **SS 3'** do último intron (Figura 5, painel inferior). Os modelos do A-ParaDDise concordam com este sítio de splicing, nos levando a concluir que estes 2 nt extras sejam um possível erro da sequência original do MEG-19 (já que este veio de ESTs, que tem uma qualidade inferior aos reads produzidos nas plataformas de sequenciamento atuais). Note que o também não há concordância entre a localização do exon inicial entre os modelos produzidos por blat e splign/gmap. Ambos os loci tem exatamente a mesma sequência do exon 1.

Figura 5 – Modelos do MEG-19.

Visualização no Genome Browser dos modelos obtidos para o MEG-19. Também estão incluídas as tracks do A-ParaDDise e SRA. O painel inferior corresponde ao último exon, destacado (em ciano) no painel superior.

Estes casos em que o blat apresenta melhor desempenho revelam uma limitação de nossa abordagem, que não está levando em conta cobertura e se os sítios de *splicing* são canônicos ou não. Isto poderia ser melhorado usando uma pontuação que leve em conta essas características, atribuindo um peso para cada um desses itens. Os pesos poderiam ser modelados empiricamente ou, melhor ainda, treinando um classificador por reconhecimento de padrões. Pretendemos melhorar essa técnica no futuro usando uma das abordagens propostas acima.

3.2 Microexons e sua abundancia em *S. mansoni* e organismos modelo

3.2.1 Breves considerações sobre as anotações de *S. mansoni*

Na anotação recente do genoma de *S. mansoni* (Protasio et al. 2012), 90% dos exons internos apresentam tamanho entre 54 e 514 nt, com média de 211 nt e mediana 162 nt. Em contraste, exons terminais apresentam entre 39 e 1454 nt, com média e mediana de, respectivamente, 659 e 208 (Tabela 3).

Exons terminais curtos podem ser, na verdade, exons internos de genes com modelo ainda incompleto. No caso de *gene models* que só foram obtidos por mapeamento de transcritos ao genoma (como é o caso dos MEGs), estes exons terminais

curtos representariam transcritos cujo mRNA completo ainda não foi detectado. O fato da sequencia do genoma de *S. mansoni* ainda estar fragmentada (Protasio et al. 2012) reforça ainda mais essa possibilidade.

Tabela 3 – Estatística descritiva de exons e introns da anotação do genoma de *S. mansoni*.

As contagens foram feitas considerando cada componente (exon/intron/etc) único, ignorando suas repetições (por exemplo, um exon presente em mais de um gene/isoforma é contabilizado apenas uma vez).

feature	count	mean	std	min	5%	25%	50%	75%	95%	max
iEx	49244	211	196	1	54	102	162	258	514	6114
introns	58390	2467	3571	0	34	643	1580	2921	8017	177535
single exon	4381	451	2704	0	66	68	72	336	1575	160008
5' exon	9636	713	4328	1	15	72	156	355	2728	220723
3' exon	9702	605	2891	2	39	132	272	537	1454	152465
5' & 3'	19338	659	3678	1	23	96	208	462	1953	220723

Figura 6 – Distribuição de exons e introns em *S. mansoni*.

Em ambos os painéis superiores, distribuição de exons internos, destacando se são simétricos (azul) ou não (verde). Nos painéis inferiores, histogramas da distribuição de comprimento de introns. Introns < 100 nt são mostrados somente no painel inferior esquerdo pois estão em uma escala diferente. O tamanho máximo estabelecido nos painéis da direita é aprox. o 95º percentil dos dados.

Na distribuição de introns (Figura 6, quadrante inferior esquerdo), é possível notar os microintrons, previamente reportados em outros platelmintos (Tsai et al. 2013). Sua abundância põe o resto da distribuição de introns fora de escala, por isso os separamos dos demais. Microintrons já haviam sido reportados no *S. mansoni* pre-

viamente, embora não tenham usado essa nomenclatura: na publicação do primeiro rascunho do genoma, os autores reportaram que o menor intron encontrado tinha 26 nt (Berriman et al. 2009).

Exons de tamanho inferior a 51 bp equivalem a menos de 5% do total de exons internos presentes no genoma de *S. mansoni* (Tabela 3). Em mamíferos, exons com esse tamanho equivalem a $\cong 6\%$ do total (Li et al. 2015).

3.3 Definições: TaMEG e parâmetros para sua busca

A fim de desenvolver um script que busque anotações de transcritos com as características dos MEGs, analisamos características de suas respectivas anotações para definir os parâmetros que devem ser verificados.

Nem todos os MEGs conhecidos de *S. mansoni* tem um código Smp (*Schistosoma mansoni* prediction) e, portanto, não dispõem de anotações (Tabela 4). Dos 26 MEGs que receberam código Smp, dois foram removidos na anotação recente (v5.2) (Protasio et al. 2012) (Tabela 4): MEG-3.4 (GenBank: GU258185.1) e MEG-18 (GenBank: EX507192.1). Além disso, alguns modelos que apresentavam refinamentos manuais feitos pelo Dr. Ricardo DeMarco foram substituídos por anotações automáticas [REF: comunicação pessoal] nas anotações v5.2, resultando em modelos com anotação piorada (Tabela 4 e Figura 7). Portanto, MEGs com código Smp estão melhor representados na anotação da versão antiga (v3.1) do genoma (Berriman et al. 2009) e, por isso, optamos por analisar inicialmente esta versão.

A análise de membros representativos de cada família MEG (de 1 a 18) (Tabela 5) nos permitiu chegar a uma **definição** que foi implementada no script python muEx-TandemFinder.py. Depois disso, testamos e ajustamos essa definição com os melhores alinhamentos obtidos para os MEGs não anotados (Figura 8).

Por conveniência, relembremos aqui a definição apresentada na sessão de metodologia: um transcrito pertencente à quaisquer família de MEG atende aos seguintes requisitos:

1. Apresentar no mínimo 3 μ Ex em tandem;
2. Apresentar exons internos com mediana ≤ 36 nt;
3. Satisfazer a seguinte condição:

$$\begin{aligned} \mu ExT_{max} &= iEx, & \text{se } 3 \leq iEx < 5 \\ \frac{\mu ExT_{max}}{iEx} &> 0.75, & \text{se } iEx \geq 5 \end{aligned}$$

Sendo $\mu\text{ExT}_{\max}$ a maior contagem de $\mu\text{Ex} \leq 61$ em tandem e $i\text{Ex}$ o total de exons internos. A contagem $\mu\text{ExT}_{\max}$ tolera um exon > 61 nt.

O modelo que obtivemos para o MEG-1.2 não seria incluído sem a tolerância ao exon grande pois sua razão $\mu\text{ExT}_{\max}/i\text{Ex}$ é 7/12 (Figura 8). Notamos que o exon de 138 nt que interrompe a sequência de microexons em tandem é, na verdade, um variante de um μEx de 21 nt com início alternativo (Figura 9), o que o enquadraria na definição que adotamos. Dando tolerância a um exon grande, a razão $\mu\text{ExT}/i\text{Ex}$ aumenta para 10/12. Note que isso não conta o exon grande como um μEx , apenas não interrompe a contagem.

Embora os MEGs analisados tenham no mínimo 75% de exons internos simétricos (Tabela 5), exigir esse valor excluiria outros MEGs conhecidos. Em nossas análises notamos que 50% seria um valor suficiente para incluir todos, mas optamos por não exigir uma quantidade mínima ou razão de exons simétricos na definição de MEGs, para não excluir potenciais candidatos interessantes com a busca.

No entanto, essa definição tem dois “problemas”:

- Não considera o quanto os μEx contribuem para a sequência codificadora do transcrito;
- Não captura o MEG-21.

Preferimos não exigir uma contribuição mínima das sequências de μEx em tandem pois perderíamos possíveis resultados interessantes. Por esta razão, decidimos usar a nomenclatura **TaMEG** (*Tandemly arranged Microexon Gene*) para nos referirmos ao nosso objeto de estudo. Dentro dessa definição, **MEGs** seriam um subgrupo de **TaMEGs** cujos μEx contribuiriam para **aproximadamente 75%** da ORF (Berriman et al. 2009) (discutiremos isso na parte 3.7) e, de acordo com nossas observações, apresentam **50% de exons internos simétricos**.

Verificamos a localização genômica do MEG-21 a fim de investigar se há modelos mais completos que o aproximem da definição de TaMEG e encontramos transcritos maiores tanto no subconjunto de contigs do projeto A-ParaDDise quanto do SRA (Figura 10B). O contig **DN13403_c0_g1_i1** (Figura 10A), apresenta apenas 3 μEx para um total de 7 exons internos, sendo que eles não são simétricos e não encontramos evidência de que os exons grandes possam ser expressos alternativamente como μEx . Por estas razões, consideramos o MEG-21 um *outlier* cuja detecção não é de interesse neste trabalho.

3.4 Observações sobre os MEGs previamente descritos

Conforme mencionado anteriormente, dois MEGs com código Smp foram removidos na anotação recente (v5.2): MEG-3.4 e MEG-18 (Smp_138090 e

Smp_032560) (Tabela 4). Alinhamos suas sequências depositadas no GenBank (respectivamente, GU258185.1 e EX507192.1) à versão v5.2 do genoma. Pudemos verificar que GU258185.1 (MEG-3.4) mapeia completamente no local esperado, o cluster de MEG-3 presente no scaffold SC_0163, aprox. 4 knt *downstream* dos demais MEG-3 presentes na anotação.

Já para a sequência EX507192.1 (MEG-18) obtivemos apenas alinhamentos parciais no cromossomo 2. Curiosamente, esta mesma sequência é mapeada completamente à versão v3.1 do genoma (usando o mesmo mapeador - gmap - e mesmos parâmetros).

Tentando entender por que o mapeador falha em alinhar o gene em uma versão mas não na outra, decidimos investigar onde cada exon do MEG-18 (de acordo com o mapeamento da versão v3.1) pode ser mapeado tanto na versão v5.2 quanto v3.1 (Tabela 9). Ao fazer isso, notamos que os exons 2, 3 e 4 são exatamente idênticos.

Na montagem v5.2 foi possível detectar 5 loci do MEG-18 no cromossomo 2, conforme previamente reportado (Philippsen et al. 2015) (Tabela 9). No caso do locus 5, apenas detectamos o exon 1. Já na montagem v3.1, os mesmos loci foram detectados, com início e fim, e também um possível 6º locus que apresenta somente o último exon do MEG-18.

O locus 2, onde o MEG-18 mapeia completamente na v3.1, tem 6.2 mil nt. Esse mesmo locus na v5.2 tem aprox 2 mil nt. Apesar da diferença de tamanho, acreditamos que estes sejam o mesmo loci com base na identidade dos alinhamentos obtidos e posição relativa aos demais loci (Tabela 9).

A explicação mais simples para esse desfalque de quase 4 mil nt no locus 2 do MEG-18 é que esta região teria sido removida por um erro de montagem.

Outra possibilidade é que a região não teria sido resolvida na montagem e, portanto, teria um dos *gaps* de poli-N, bastante comuns no rascunho do genoma de *S. mansoni*. (Na montagem do genoma, scaffolds foram unidos com base em evidência molecular - FISH; quando não há sobreposição de sequência entre os scaffolds unidos desta forma, sequências poli-N foram inseridas). No entanto, verificamos que não há gaps entre as coordenadas do exon 1 e 6 do locus 2 (Figura 11).

Uma hipótese alternativa é que o locus 2 seja diferente entre os clones *S. mansoni* sequenciados. Este locus teria sofrido modificações mediadas por TE (*transposable elements*) ativos e, no clone do sequenciamento mais recente, um TE teria “saltado”. Esse hipotético salto seria recente pois o EST do MEG-18 foi sequenciado de RNA extraído de uma cepa diferente de *S. mansoni* (cepa LE <http://www.ncbi.nlm.nih.gov/nucest/166340298/> -- aparentemente o artigo não foi publicado). O fato dos exons 2, 3 e 4 terem a mesma sequência sugere que sua inser-

ção no MEG-18 tenha sido mediada por elementos transponíveis, como ocorreu nos loci de genes da família MEG-3 (DeMarco et al. 2010).

Uma última observação sobre os MEGs previamente descritos diz respeito ao **MEG-17**. Enquanto o MEG-17 anotado (Smp_180620) está localizado no *scaffold C-hr_2.unplaced.SC_0108* (coordenadas 1,141,820 1,146,941), a sequência disponibilizada pelo Prof. Ricardo DeMarco mapeia melhor em um locus do cromossomo 1 (coordenadas 65,414,642 a 65,419,874). Ambos os loci tem evidência de transcritos, e, por esse motivo, anotamos este locus do cromossomo 1 como MEG-17.2.

Usando o resultado de mapeamentos dos MEGs conhecidos ao genoma, geramos uma track com todos a isoforma que apresenta a maior CDS de cada um dos MEGs previamente descritos (incluindo o MEG-17.2, MEG-18 e MEG-21), denominada “knownMEGs.bed”. Nesta track também foram incluídas as anotações da v5.2 para os MEGs que não estavam presentes no mapeamento (MEG 2.1, 2.3 e 17.1). Esta track está disponível para consulta no genome browser de *Schistosoma mansoni* (<http://schistosoma.usp.br>).

3.5 Busca de TaMEGs no transcrito de *S. mansoni*

A tabela Tabela 10 mostra a classificação dos genes detectados pelo pipeline de busca de TaMEGs no transcriptoma de *S. mansoni*. Os 50 MEGs previamente descritos foram detectados, o que não é uma surpresa dado que usamos praticamente todos os dados de expressão de RNA públicos disponíveis de *S. mansoni*. A surpresa são os 31 novos genes detectados, sendo 13 deles pertencentes à famílias MEG previamente descritas. Dos demais 17 genes, 15 formam famílias monogênicas e dois deles formam a família que denominamos **SmFAM_10** (Tabela 10).

Ambos os genes da família SmFAM_10, Locus_0061 e Locus_0062, são curtos (74 e 82 resíduos de aa, respectivamente), apresentam peptídeo sinal e a contribuição relativa de seus μ Ex na sequência codificante são de 38 e 59%, respectivamente. Estes genes são vizinhos, como ocorre com algumas das famílias MEGs multigênicas previamente descritas.

Curiosamente, nossa classificação agrupou as famílias 17 e 25 (SmFAM_06) e também 8 e 28 (SmFAM_02).

No caso da **SmFAM_06**, o que ocorre é que o MEG-25 original era uma sequência incompleta. Detectamos uma nova isoforma com ORF maior e completa para o MEG-25 e que tem alta similaridade com MEG-17. Buscamos, por blastn, loci genômicos com similaridade ao exon final do MEG-17 e 25 e verificamos existe sobreposição entre todos eles (loci reportados na Tabela 6).

Já no caso da **SmFAM_02**, que uniu MEGs 8 e 28, verificamos no output no blastp que não há hits entre MEG-28 e os MEG-8. O agrupamento desses genes o-

correu devido a introdução de um novo gene (Locus_0060) que tem similaridade tanto com o MEG-28 quanto com os demais MEG-8, criando esse novo cluster. O alinhamento da sequência de aminoácido desses genes revela similaridade na porção C-terminal desses genes (Figura 12).

3.6 TaMEGs detectados na coleção de anotações de genomas

Detectamos TaMEGs em 125 organismos somando um total de 494 genes, sendo que 241 deles tem uma contribuição de μ Ex na sequência codificante superior a 50% (Tabela 7). 200 TaMEGs foram agrupados em famílias de 2 ou mais genes, cujas características principais estão resumidas na Tabela 11. Os demais 294 genes formam famílias monogênicas.

Detectamos uma média de 4 genes por organismo, que é a quantidade máxima de genes detectados em 96 (75%) dos organismos. Apenas 11 organismos apresentam 10 ou mais TaMEGs (Tabela 8).

3.6.1 Genes e famílias das espécies com maior abundância de TaMEGs

A descrição a seguir de alguns genes e famílias das 11 espécies com 10 ou mais TaMEGs foi feita com auxílio das seguintes planilhas do arquivo “tabelas_extra.xlsx”: sheet3, que contém a linhagem completa dos 125 organismos detectados aqui; sheet4, que contém informações sobre cada gene detectado, como tamanho da proteína, contribuição absoluta e relativa de resíduos codificados por μ Ex, número de resíduos com tendência a desordem, presença de peptídeo sinal, domínios transmembrana, cluster, espécie, identificadores no ensembl e link para o ensembl; sheet5, que consiste numa versão mais completa da Tabela 11, incluindo as famílias monogênicas. Outro arquivo consultado foi o “clusters_top_species.xlsx” anexo, que consiste em um recorte da planilha sheet5 somente com as famílias de uma espécie específica para os 11 organismos com mais TaMEGs.

Dentre os 11 organismos com 10 ou mais TaMEGs, 5 são animais invertebrados, 3 são fungos, 2 são protistas e 1 deles é um animal vertebrado.

Os seis animais invertebrados pode ser subdivididos em:

- 4 artrópodos, sendo um zooplâncton (*Daphnia pulex*), uma centopéia (*Strigamia maritima*), uma borboleta (*Melitaea cinxia*) e um afídio (*Acyrtosiphon pisum*);
- 1 nematelminto, *Onchocerca volvulus*, agente etiológico da oncocercose.

A borboleta *M. cinxia* é o organismo com maior número de TaMEGs, sendo que 6 deles estão agrupados em famílias:

- 2 na família FAM_012, que inclui um gene de cada um dos seguintes organismos da ordem Lepidoptera (borboletas e mariposas): o bicho da seda (*Bombyx mori*), e outras duas borboletas (*Danaus plexippus* e *Heli-*

conius melpomene). É um grupo particularmente interessante, cuja contribuição média dos μ Ex na ORF é de 78%, sem domínios conhecidos e a maioria apresenta peptídeo sinal (Tabela 11) – apenas um dos parálogos de *M. cinxia* não apresenta peptídeo sinal.

- 3 na família FAM_019, família exclusiva de *M. cinxia*, todos com peptídeo sinal, sem domínio Pfam e contribuição média dos μ Ex na ORF de 67% (Tabela 11).
- 1 na família FAM_024, cluster de apenas 2 genes com contribuição média dos μ Ex na ORF de 81% e sem domínio Pfam (Tabela 11). O gene de *M. cinxia* apresenta peptídeo sinal, enquanto o outro, da borboleta *D. plexipus*, não.

Os TaMEGs detectados nos demais invertebrados formam famílias exclusivas à suas respectivas espécies, sendo a maioria famílias monogênicas.

Os três fungos são ascomycetes: *Colletotrichum orbiculare*, *Trichoderma reesei* e *Fusarium pseudograminareum*.

Tanto *C. orbiculare* quanto *F. pseudograminareum* tem representantes na família FAM_005, uma família de genes de ascomycetes que inclui genes de *Botrytis cinerea*, *Sclerotinia scleroiorum* e outras espécies do gênero *Colletotrichum*. Todos apresentam um N-terminal com exons de tamanho 12,24,26,42 e 55, mas que contribuem apenas com 12% do total da ORF. São tubulinas, dada a presença dos domínios PF00091, PF03953.

A família FAM_027 tem dois genes, sendo um de *F. pseudograminareum* e *F. graminareum*, sem assinaturas Pfam e apresentam contribuição média de seus μ Ex de 71 % da ORF.

T. reesei apresenta 4 TaMEGs que compoem a família FAM_013, família de 5 genes que apresenta um membro não fungo: *Thalassiosira pseudonana*. Possivelmente uma família mal classificada, dada a discrepância de tamanho do gene do protista: quase 2 vezes e meia maior que a média dos 4 genes de *T. reesei* e com menor contribuição de μ Ex na ORF.

As demais famílias de *C. orbiculare*, *F. pseudograminareum* e *T. reesei* são monogênicas.

Ambos os protistas são oomicetes, eucariotos unicelulares similares aos fungos, mas que pertencem ao reino Chromista: *Phytophthora kernoviae* e *Pythium ultimum* (também conhecido como *Globisporangium ultimum*). Ambos são patógenos de plantas.

P. ultimum apresenta seis famílias não monogênicas exclusivas de seu gênero (Tabela 11), havendo um TaMEG do *P. iwayamal* contribuindo com a família FAM_016. Nenhum dos genes de *P. ultimum* apresenta domínios conhecidos ou peptídeo sinal, mas 9 deles tem domínio transmembrana.

P. kernoviae tem apenas famílias monogênicas, sendo que duas apresentam assinaturas funcionais do Pfam. Estas ocorrem em exons grandes destes genes.

Finalmente, o único animal vertebrado do grupo de organismos com maior abundância de TaMEGs é a ovelha (*Ovis aries*). Seis de seus genes estão incluídos nas quatro maiores famílias de TaMEG, FAM_001 a 004, que abordaremos adiante.

Três TaMEGs da ovelha estão presentes na FAM_008, uma família gênica de mamíferos com baixa contribuição média de μ Ex (11% da ORF), domínio transmembrana em todos os genes e, de seus 7 genes, 6 apresentam domínio PFam PF13853 (receptor olfatório).

Um dos TaMEGs de ovelha participa da família FAM_015, também uma família de genes de mamíferos com contribuição média de μ Ex na ORF de 51%. Três dos quatro genes dessa família apresentam peptídeo sinal e nenhum domínio conhecido.

3.6.2 Famílias TaMEG de Cordados

Embora os vertebrados apresentem poucos genes por espécie, os TaMEGs detectados neles formam os maiores grupos (Tabela 11). FAM_001 a 004 são formadas exclusivamente por genes de organismos do filo Chordata, sendo que três dessas famílias são exclusivas de mamíferos.

Algo particularmente interessante dessas quatro famílias é que elas apresentam representantes no ser humano e em camundongo. Consultamos a anotação presente no Pfam e Uniprot desses genes.

A primeira família, FAM_001 é composta de caseínas (caseína alfa-s1, alfa-s2 e beta) (69% delas apresentam a assinatura Pfam característica das caseínas, PF00363), proteínas que são o principal composto do leite produzido pelos mamíferos e são responsáveis pelo transporte de cálcio fosfato. A contribuição média de μ Ex à ORF das caseínas é equiparável à dos primeiros MEGs descritos (59%) e em 91% dos genes detectamos peptídeo sinal. Interessantemente, um peptídeo derivado da alfa-s2 inibe o crescimento de bactérias.

Todos os membros da FAM_002 apresentam um domínio ART (PF01129), que tem atividade ADP-ribosiltransferase e são conhecidas como ART3. O que distingue ART3 das demais ARTs é a presença de microexons de 30 a 33 nt em tandem separando o domínio ART e a porção C-terminal da proteína. Em média, a contribuição dos μ Ex para a proteína é de apenas 20%, e eles coincidem com os resíduos com tendência à desordem.

A FAM_003 contém, além de genes de mamíferos, genes de aves (*Gallus gallus*) e seus principais domínios são o Ig V-set e SPRY (respectivamente, PF07686 e PF00622). Todos os genes de galinha (7 genes) deste grupo apresentam o domínio Ig V-set e são parálogos do B-G, o MHC-IV, exclusivo de aves. Outros 10 genes, de mamíferos, apresentam ambos os domínios aparentam ser todos membros da família

ERMAP, proteína presente nas hemácias e responsável pelo grupo sanguíneo “Scian-na”. A maioria destes genes apresenta exons de 21 a 27nt e contribuem, em média, com 29% da ORF.

Finalmente, a família FAM_004 é uma das mais intrigantes. Todas recebem a anotação genérica “c6orf10” em alusão à sua localização no genoma humano e não apresentam assinatura de domínio conhecido. A maioria de seus microexons são de 21nt e contribuem, em média, com 35% da ORF desses genes. Sua média de resíduos tendendo a desordem é alta (69%).

3.7 Contribuição de μ Ex à ORF

Ao comparar a contribuição de relativa de μ Ex à sequência codificante (μ Ex%) nos MEGs originais e nos TaMEGs detectados neste trabalho (Figura 13), fica claro que os TaMEGs tem uma tendência a ter uma menor contribuição de μ Ex. Por exemplo, apenas 25% dos MEGs tem μ Ex% entre menor que 60, enquanto que quase metade dos novos TaMEGs de *S. mansoni* apresentam μ Ex% inferior a 60.

Notamos que, historicamente, a distribuição de μ Ex% foi “descendo” (Figura 14), a ponto de, da publicação de 2012 para a de 2015, a μ Ex% mínima dos MEGs reportados cair de ~40% para ~20%. Assim, se no começo MEGs com μ Ex% < 70 eram exceção, na publicação mais recente isso se tornou regra.

4 Conclusões

Desenvolvemos um pipeline capaz de detectar genes com a arquitetura dos MEGs de *S. mansoni*, com exceção do MEG-21. Como esse pipeline somente considera a arquitetura desses genes, mas não a contribuição de μ Ex na proteína final, chamamos os genes que detectamos de TaMEG.

Detectamos 31 TaMEGs em *S. mansoni*. Acreditamos que todos poderiam ser considerados MEGs dada a contribuição de seus microexons nos resíduos de aminoácidos das proteínas produzidas por eles.

Detectamos 494 TaMEGs na coleção de organismos. A maioria dos organismos apresenta menos de uma dezena de genes com essa arquitetura. Apenas 11 organismos apresentam mais de 10 desses genes, dentre eles invertebrados, protistas e fungos.

5 Bibliografia

Almeida GT, Amaral MS, Beckedorff FCF, Kitajima JP, DeMarco R, Verjovski-Almeida S. 2012. Exploring the *Schistosoma mansoni* adult male transcriptome using RNA-seq. *Experimental Parasitology* **132**: 22.

- Beachy PA, Helfand SL, Hogness DS. 1985. SEGMENTAL DISTRIBUTION OF BITHORAX COMPLEX PROTEINS DURING DROSOPHILA DEVELOPMENT. *Nature* **313**: 545-551.
- Berget SM. 1995. Exon Recognition in Vertebrate Splicing. *Journal of Biological Chemistry* **270**: 2411-2414.
- Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD et al. 2009. The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**: 352.
- Carlo T, Sterner DA, Berget SM. 1996. An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon. *Rna-a Publication of the Rna Society* **2**: 342-353.
- Crick F. 1970. Central Dogma of Molecular Biology. *Nature* **227**: 561-563.
- DeMarco R, Mathieson W, Manuel SJ, Dillon GP, Curwen RS, Ashton PD, Ivens AC, Berriman M, Verjovski-Almeida S, Wilson RA. 2010. Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts. *Genome Research* **20**: 1112.
- Frank-Hansen R, Page SP, Syrris P, McKenna WJ, Christiansen M, Andersen PS. 2008. Micro-exons of the cardiac myosin binding protein C gene: flanking introns contain a disproportionately large number of hypertrophic cardiomyopathy mutations. *European Journal of Human Genetics* **16**: 1062-1069.
- Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**: 307-321.
- Holter J, Davies J, Leresche N, Crunelli V, Carter DA. 2005. Identification of two further splice variants of GABABR1 characterizes the conserved micro-exon 4 as a hot spot for regulated splicing in the rat brain. *Journal of Molecular Neuroscience* **26**: 99-108.
- Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T et al. 2002. The Ensembl genome database project. *Nucleic Acids Res* **30**: 38-41.
- Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol Biol Evol* **33**: 1635-1638.
- Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallieres M, Tapial J, Raj B, O'Hanlon D et al. 2014. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell* **159**: 1511-1523.
- Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nature Reviews Genetics* **11**: 345-355.

- Kersey PJ, Staines DM, Lawson D, Kulesha E, Derwent P, Humphrey JC, Hughes DS, Keenan S, Kerhornou A, Koscielny G et al. 2012. Ensembl Genomes: an integrative resource for genome-scale data from non-vertebrate species. *Nucleic Acids Res* **40**: D91-97.
- Koonin EV. 2009. Evolution of genome architecture. *International Journal of Biochemistry & Cell Biology* **41**: 298-306.
- Li YI, Sanchez-Pulido L, Haerty W, Ponting CP. 2015. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Research* **25**: 1-13.
- McAllister L, Rehm EJ, Goodman CS, Zinn K. 1992. Alternative splicing of micro-exons creates multiple forms of the insect cell-adhesion molecule Fasciclin-I. *Journal of Neuroscience* **12**: 895-905.
- Ovadia G, Shifman S. 2011. The Genetic Variation of RELN Expression in Schizophrenia and Bipolar Disorder. *Plos One* **6**.
- Philippsen GS, Wilson RA, DeMarco R. 2015. Accelerated evolution of Schistosome genes coding for proteins located at the host-parasite interface. *Genome Biology and Evolution*.
- Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, De Silva N, Velarde GS, Anderson TJ, Clark RC et al. 2012. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS neglected tropical diseases* **6**: e1455.
- Rogozin IB, Makarova KS, Natale DA, Spiridonov AN, Tatusov RL, Wolf YI, Yin J, Koonin EV. 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Research* **30**: 4264-4271.
- Sievers F, Wilm A, Dineen D, Gibson T, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**.
- Staley JP, Guthrie C. 1998. Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell* **92**: 315-326.
- Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, Tracey A, Bobes RJ, Fragoso G, Sciutto E et al. 2013. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**: 57-63.
- Volfovsky N, Haas BJ, Salzberg SL. 2003. Computational Discovery of Internal Micro-Exons. *Genome Research* **13**: 1216-1221.
- Wilson RA, Li XH, MacDonald S, Neves LX, Vitoriano-Souza J, Leite LCC, Farias LP, James S, Ashton PD, DeMarco R et al. 2015. The Schistosome Esophagus Is a 'Hotspot' for Microexon and Lysosomal Hydrolase Gene Expression: Implications for Blood Processing. *Plos Neglected Tropical Diseases* **9**: 25.

6 Figuras e Tabelas

Figura 9 – MEG-1.2

Em destaque (ciano), exon de 21nt, alternativo do exon de 138 nt, confirmado em contigs dos projetos SRA e A-ParaDDise

Figura 10 – MEG-21

A) Arquitetura dos exons do MEG-21 e um dos contigs que o estendem, DN13403_c0_g1_i1. A figura segue o mesmo código de cores e escala da Figura 8.

B) Visualização do locus do MEG-21 no Genome Browser.

Figura 11 – Cluster de MEG-18 no cromossomo 2

Visualização no Genome Browser do cluster de MEG-18 no cromossomo 2. A track “MEG-18 loci” foi gerada à partir da Tabela 8. No caso do locus 5, que não detectamos a localização do exon final, marcamos um final arbitrário na posição 290 000. A track “gap” representa sequências de poli-N.

Figura 12 – Destaque do alinhamento múltiplo do cluster SmFAM_02.

Porção C-terminal do AM entre os MEG-8, MEG-28 e o novo gene Locus_0060. Ao lado no nome de cada sequência são mostrados o intervalo de resíduos de aa presentes na figura. As sequências estão ordenadas da mesma forma que na árvore produzida.

Figura 13 - Comparação entre os MEGs originais e os TaMEGs detectados.

Comparamos o a contribuição relativa de μEx à sequência codificante (painel superior) e tamanho das proteínas (painel inferior) dos MEGs originais (azul), dos TaMEGs inéditos de *S. mansoni* detectados neste trabalho e também os detectados na coleção de genomas.

Figura 14 – Comparação entre os MEGs reportados em cada publicação.

Comparamos o a contribuição relativa de μEx à sequencia codificante (painel superior) e tamanho das proteínas (painel inferior) dos MEGs originais e dos TaMEGs inéditos de *S. mansoni*. Nesta figura, separamos os MEGs originais por ano de publicação.

6.2 Tabelas

Tabela 4 – MEGs com código Smp.

Em vermelho, MEGs cujo Smp foi removido na versão recente do genoma.

Em azul, MEGs cuja anotação melhor representa um MEG na versão 5.2.

Em verde, MEGs presentes na versao 5.2, porém com anotação piorada

Family	Name	Smp acc	v3	v5.2
1	MEG-1	Smp_122630	1	1
2	MEG-2	Smp_159800	1	1
2	MEG-2	Smp_159810	1	1
2	MEG-2	Smp_159830	1	1
2	MEG-2	Smp_179860	1	1
2	MEG-2	Smp_180310	1	1
2	MEG-2	Smp_180320	2	2
2	MEG-2	Smp_180330	1	1
2	MEG-2	Smp_180340	1	1
2	MEG-2	Smp_181510	2	2
3	MEG-3.1	Smp_138060	1	1
3	MEG-3.2	Smp_138070	1	1
3	MEG-3.3	Smp_138080	1	1
3	MEG-3.4	Smp_138090	1	0
4	MEG-4	Smp_085840	2	1
4	MEG-4	Smp_163630	1	2
5	MEG-5	Smp_152580	1	1
6	MEG-6	Smp_163710	1	1
7	MEG-7	Smp_165050	1	1
8	MEG-8	Smp_171190	1	1
8	MEG-8	Smp_172180	1	1
9	MEG-9	Smp_125320	1	1
10	MEG-10	Smp_152590	1	1
11	MEG-11	Smp_176020	1	1
12	MEG-12	Smp_152630	1	2
13	MEG-13	Smp_127990	1	1
14	MEG-14	Smp_124000	1	1
15	MEG-15	Smp_010550	1	1
16	MEG-16	Smp_158890	1	1
17	MEG-17	Smp_180620	1	1
18	MEG-18	Smp_032560	1	0
32	MEG-32.1	Smp_123100	1	1
32	MEG-32.2	Smp_123200	1	1
isoforms			36	35
genes			33	31

Tabela 5 – Composição dos exons internos de MEGs representativos

Contagem do número de exons internos com a característica de cada coluna. No caso das contagens “em tandem”, é considerada a maior sequência de exons com a característica estudada.

μEx – micro exon

μExS – micro exon simétrico

ExS – exon simétrico

MEG	Smp	#	μEx ≤ 36	μEx ≤ 60	μEx ≤ 36 tan-dem	μEx ≤ 60 tan-dem	μExS ≤ 36 tan-dem	μExS ≤ 60 tan-dem	ExS
1	Smp_122630	13	9	13	8	13	8	12	12
2	Smp_159800	12	12	12	12	12	11	11	11
3	Smp_138080	19	19	19	19	19	19	19	19
4	Smp_163630	18	18	18	18	18	16	16	16
5	Smp_152580	6	5	6	3	6	3	3	5
6	Smp_163710	9	9	9	9	9	9	9	9
7	Smp_165050	8	6	7	4	7	4	7	7
8	Smp_172180	15	15	15	15	15	13	13	14
9	Smp_125320	5	5	5	5	5	3	3	4
10	Smp_152590	4	4	4	4	4	3	3	3
11	Smp_176020	8	8	8	8	8	8	8	8
12	Smp_152630	5	5	5	5	5	5	5	5
13	Smp_127990	5	5	5	5	5	5	5	5
14	Smp_124000	13	13	13	13	13	12	12	12
15	Smp_010550	12	10	12	5	12	5	11	11
16	Smp_158890	6	4	6	3	6	3	4	5
17	Smp_180620	8	8	8	8	8	3	3	6
18	Smp_032560	4	4	4	4	4	4	4	4
mínimo		4	4	4	3	4	3	3	3
razão mínima		-	0.67	0.88	0.42	0.88	0.38	0.38	0.75

Tabela 6 – Sobreposição do MEG-17 e MEG-25.

Coordenadas genômicas onde há sobreposição entre as coordenadas do último exon do MEG-17 e MEG-25.

Cromossomo	Início	Fim
Chr_1	33618460	33618617
Chr_1	33721377	33721534
Chr_1	65414641	65414798
Chr_1.unplaced.SC_0153	176183	176340
Chr_2.unplaced.SC_0108	1141680	1141837
Chr_4	17315007	17315163
Chr_6	19857263	19857420
SC_0041	1896292	1896444
SC_0212	209934	210091
SC_0236	37022	37177
SC_0236	197246	197399
SC_0271	52891	53049
SC_0271	68788	68945
SC_0300	15684	15840
SC_0430	16161	16316

Tabela 7 – Contagem de genes detectados por grupos taxonômicos.

Total de genes detectados nos grandes grupos taxonômicos presentes na coleção de genomas. Chamamos de “meglike” genes cuja contribuição de μEx na sequência codificadora é $> 50\%$.

division	species	genes	meglike
vertebrates	51	195	50
invertebrates	21	122	78
fungi	28	99	58
protists	12	56	42
plants	13	22	13
Total	125	494	241

Tabela 8 – Contagem de genes detectados por espécie.

A lista está em ordem decrescente da quantidade de genes. Chamamos de “meglike” genes cuja contribuição de μEx na sequência codificadora é $> 50\%$.

specie	division	genes	meglike
Melitaea cinxia	invertebrates	26	18
Pythium ultimum	protists	24	22
Colletotrichum orbiculare	fungi	22	19
Strigamia maritima	invertebrates	15	9
Acyrtosiphon pisum	invertebrates	15	12
Fusarium pseudograminearum	fungi	14	12
Ovis aries	vertebrates	13	3
Trichoderma reesei	fungi	13	8
Phytophthora kernoviae	protists	12	8
Daphnia pulex	invertebrates	10	3
Onchocerca volvulus	invertebrates	10	8
Capitella teleta	invertebrates	9	8
Danio rerio	vertebrates	9	4
Sus scrofa	vertebrates	8	2
Bos taurus	vertebrates	8	2
Oryctolagus cuniculus	vertebrates	8	2
Ornithorhynchus anatinus	vertebrates	8	0
Macaca mulatta	vertebrates	8	1
Myotis lucifugus	vertebrates	7	1
Latimeria chalumnae	vertebrates	7	0
Gallus gallus	vertebrates	7	2
Heliconius melpomene	invertebrates	7	4
Felis catus	vertebrates	7	3
Sarcophilus harrisii	vertebrates	6	1
Equus caballus	vertebrates	6	1
Amphimedon queenslandica	invertebrates	6	2
Mus musculus	vertebrates	6	3
Trichoderma virens	fungi	5	2
Phytophthora lateralis	protists	5	5
Rattus norvegicus	vertebrates	5	3
Homo sapiens	vertebrates	5	1

specie	division	genes	meglike
Anolis carolinensis	vertebrates	4	2
Rhodnius prolixus	invertebrates	4	1
Thalassiosira pseudonana	protists	4	0
Trichoplax adhaerens	invertebrates	4	3
Otolemur garnettii	vertebrates	4	1
Nomascus leucogenys	vertebrates	4	1
Mustela putorius furo	vertebrates	4	1
Gorilla gorilla	vertebrates	4	1
Dasypus novemcinctus	vertebrates	4	1
Fusarium solani	fungi	4	1
Aspergillus terreus	fungi	3	1
Oryza nivara	plants	3	2
Aspergillus niger	fungi	3	2
Danaus plexippus	invertebrates	3	3
Oryza rufipogon	plants	3	2
Fusarium graminearum	fungi	3	3
Chlamydomonas reinhardtii	plants	3	2
Ailuropoda melanoleuca	vertebrates	3	1
Magnaporthe oryzae	fungi	3	2
Callithrix jacchus	vertebrates	3	1
Canis familiaris	vertebrates	3	2
Erinaceus europaeus	vertebrates	3	2
Petromyzon marinus	vertebrates	3	0
Pongo abelii	vertebrates	3	1
Neurospora crassa	fungi	3	1
Oryza glumaepatula	plants	3	2
Puccinia graminis	fungi	2	1
Anas platyrhynchos	vertebrates	2	0
Tuber melanosporum	fungi	2	0
Aspergillus flavus	fungi	2	1
Toxoplasma gondii	protists	2	0
Emiliana huxleyi	protists	2	1
Phytophthora infestans	protists	2	1
Oreochromis niloticus	vertebrates	2	0
Microbotryum violaceum	fungi	2	2
Poecilia formosa	vertebrates	2	0
Lottia gigantea	invertebrates	2	2
Atta cephalotes	invertebrates	2	1
Cavia porcellus	vertebrates	2	0
Leptosphaeria maculans	fungi	2	0
Ixodes scapularis	invertebrates	2	1
Astyanax mexicanus	vertebrates	2	0
Zymoseptoria tritici	fungi	2	0
Xenopus tropicalis	vertebrates	2	0
Pan troglodytes	vertebrates	2	1
Tarsius syrichta	vertebrates	2	1
Colletotrichum graminicola	fungi	2	0

specie	division	genes	meglike
Ochotona princeps	vertebrates	2	0
Oryza barthii	plants	2	1
Botrytis cinerea	fungi	2	0
Lepisosteus oculatus	vertebrates	2	1
Papio anubis	vertebrates	2	0
Loxodonta africana	vertebrates	2	0
Pelodiscus sinensis	vertebrates	1	0
Pediculus humanus	invertebrates	1	1
Megaselia scalaris	invertebrates	1	1
Musa acuminata	plants	1	0
Colletotrichum gloeosporioides	fungi	1	0
Microcebus murinus	vertebrates	1	0
Phaeosphaeria nodorum	fungi	1	1
Colletotrichum higginsianum	fungi	1	0
Vicugna pacos	vertebrates	1	0
Pythium aphanidermatum	protists	1	1
Procavia capensis	vertebrates	1	0
Sorghum bicolor	plants	1	0
Echinops telfairi	vertebrates	1	0
Tribolium castaneum	invertebrates	1	0
Xiphophorus maculatus	vertebrates	1	0
Magnaporthe poae	fungi	1	0
Loa loa	invertebrates	1	1
Sclerotinia sclerotiorum	fungi	1	0
Albugo laibachii	protists	1	1
Arabidopsis lyrata	plants	1	0
Arabidopsis thaliana	plants	1	1
Nematostella vectensis	invertebrates	1	0
Aspergillus nidulans	fungi	1	0
Pythium iwayamai	protists	1	1
Tetraodon nigroviridis	vertebrates	1	1
Gadus morhua	vertebrates	1	1
Bombyx mori	invertebrates	1	0
Brassica rapa	plants	1	0
Ictidomys tridecemlineatus	vertebrates	1	1
Oryza punctata	plants	1	1
Dictyostelium discoideum	protists	1	1
Drosophila simulans	invertebrates	1	0
Oryza meridionalis	plants	1	1
Ficedula albicollis	vertebrates	1	0
Fusarium fujikuroi	fungi	1	0
Fusarium oxysporum	fungi	1	1
Fusarium verticillioides	fungi	1	1
Gaeumannomyces graminis	fungi	1	0
Guillardia theta	protists	1	1
Leersia perrieri	plants	1	1
Oryzias latipes	vertebrates	1	1

Tabela 9 – Loci do MEG-18

Coordenadas dos loci do MEG-18 em ambas as versões do genoma de *S. mansoni*. Dentre as colunas para cada versão de genoma, o triângulo indica o tamanho do locus referente, **coords** as coordenadas (no caso da v3.1 o nome Smp_scaff foi omitido para brevidade) e **strand** a orientação. As demais colunas tem o mesmo valor/significado para ambas as versões, sendo **seq** o nome do exon do MEG-18 (exons 2, 3 e 4 apresentam exatamente a mesma sequência), **seq-len** seu tamanho (em nt) e **id** a identidade.

locus	▲ v5.2	coords v5.2	strand v5.2	seq	seq-len	id	coords v3.1 (Smp_scaff)	strand v3.1	▲ v3.1
1	3247	Chr_2:168032-168143	-	exon_6	111	96.4%	000073:107373-107484	-	111
1		Chr_2:171188-171279	-	exon_1	91	97.8%	//	//	
2	1986	Chr_2:186239-186330	+	exon_1	91	100%	000073:123198-123289	+	6257
2		//	//	exon_2-3-4	36	100%	000073:124855-124891	+	
2		//	//	exon_2-3-4	36	100%	000073:127536-127572	+	
2		//	//	exon_2-3-4	36	100%	000073:128331-128367	+	
2		Chr_2:187821-187848	+	exon_5	27	100%	000073:129051-129078	+	
2		Chr_2:188114-188225	+	exon_6	111	100%	000073:129344-129455	+	
3	4757	Chr_2:197333-197444	-	exon_6	111	100%	000073:138026-138137	-	4757
3		Chr_2:201999-202090	-	exon_1	91	94.5%	000073:142692-142783	-	
4	8350	Chr_2:239495-239606	-	exon_6	111	92.8%	000073:186863-186974	-	4234
4		Chr_2:240834-240870	-	exon_2-3-4	36	100%	000073:188202-188238	-	
4		Chr_2:247761-247845	-	exon_1	84	83.5%	000073:191013-191097	-	
5	84	Chr_2:307105-307189	-	exon_1	84	84.6%	018868:2140-2224	-	84
6?	-	//	//	exon_6	111	90.1%	018658:10697-10808	+	111

Tabela 10 – Famílias de TaMEGs de *S. mansoni*.

Famílias dos TaMEGs detectados em *S. mansoni*. As famílias monogenicas foram tratadas como se fossem uma grande família. As colunas pepsize, μ Ex % e dis % são médias, respectivamente, do tamanho da sequencia peptídica, contribuição relativa de μ Ex à ORF e de resíduos com tendência à desordem. Pep. sinal e transm. contabiliza o percentual de genes que apresentam, respectivamente, peptídeo sinal e domínio transmembrana.

cluster	genes	novos genes	MEGs	pepsize	μ Ex %	dis %	pep. sinal	transm.
SmFAM_01	9	0	MEG-2	75	80	0	67	0
SmFAM_02	6	1	MEG-8, MEG-28	109	60	17	67	17
SmFAM_03	6	4	MEG-10	70	50	2	83	67
SmFAM_04	5	4	MEG-26	84	28	0	100	0
SmFAM_05	5	1	MEG-3	147	90	0	80	0
SmFAM_06	3	0	MEG-17, MEG-25	59	78	7	67	0
SmFAM_07	3	1	MEG-1	126	69	0	0	100
SmFAM_08	3	1	MEG-32	87	28	3	67	100
SmFAM_09	2	0	MEG-4	164	70	20	50	50
SmFAM_10	2	2		78	48	0	100	0
SmFAM_11	2	1	MEG-5	84	56	10	100	50
SmFAM_12	2	1	MEG-31	62	84	25	100	0
Monogenic	33	15	MEGs: 6, 7, 9, 11-16, 18-20, 22-24, 27, 29, 30	87	62	24	42	27

Tabela 11 – Famílias da coleção de genomas.

Chamamos de “meglike” genes cuja contribuição de μEx na sequência codificadora é $> 50\%$. As colunas pepsiz, $\mu\text{Ex}\%$ e $\text{dis}\%$ são médias, respectivamente, do tamanho da sequência peptídica, contribuição relativa de μEx à ORF e de resíduos com tendência à desordem. Pep. sinal e transmb. contabiliza o percentual de genes que apresentam, respectivamente, peptídeo sinal e domínio transmembrana. As três últimas colunas dizem respeito à classificação taxonômica das espécies presentes na família gênica. Na coluna “division” temos o(s) grupo(s) equivalente(s) às divisões do ensembl. As colunas tax_rank e rank_name são, respectivamente, o menor rank taxonômico comum às espécies presentes na família gênica e o nome desse rank.

cluster	genes	meglike	species	pepsiz	$\mu\text{Ex}\%$	$\text{dis}\%$	pep. sinal	transmb.	division	tax_rank	rank_name
FAM_001	32	24	22	207	59	28	91	0	vertebrates	class	Mammalia
FAM_002	22	0	22	373	20	16	86	68	vertebrates	class	Mammalia
FAM_003	19	3	12	362	29	11	26	63	vertebrates	phylum	Chordata
FAM_004	13	1	12	489	35	69	23	69	vertebrates	class	Mammalia
FAM_005	9	0	7	446	12	8	0	0	fungi	phylum	Ascomycota
FAM_006	8	0	7	425	17	22	0	25	vertebrates	phylum	Chordata
FAM_007	8	8	1	151	68	6	0	0	invertebrates	genus	Capitella
FAM_008	7	0	5	360	11	0	0	100	vertebrates	class	Mammalia
FAM_009	7	6	7	111	64	62	100	0	plants	family	Poaceae
FAM_010	6	3	6	152	55	18	100	0	vertebrates	class	Mammalia
FAM_011	6	6	1	31	100	0	0	0	protists	genus	Globisporangium
FAM_012	5	4	4	128	78	43	80	0	invertebrates	order	Lepidoptera
FAM_013	5	4	2	253	83	8	0	60	protists, fungi	multi-kingdom	Fungi,Chromista
FAM_014	4	4	4	142	80	74	100	0	vertebrates	class	Actinopterygii
FAM_015	4	3	4	148	51	2	75	25	vertebrates	class	Mammalia
FAM_016	4	4	2	73	100	0	0	50	protists	genus	Globisporangium
FAM_017	4	4	1	64	100	0	0	25	protists	genus	Globisporangium
FAM_018	3	3	3	190	79	94	0	0	vertebrates	class	Mammalia
FAM_019	3	3	1	161	67	2	100	33	invertebrates	genus	Melitaea
FAM_020	3	0	2	595	36	34	0	33	invertebrates,	multi-	Animalia,Chromista

cluster	genes	meglike	species	pepsize	μ Ex %	dis %	pep. sinal	transmb.	division	tax_rank	rank_name
FAM_021	2	2	1	154	63	2	50	0	protists invertebrates	kingdom genus	Acyrtosiphon
FAM_022	2	2	2	80	63	12	0	0	invertebrates, protists	multi- kingdom	Animalia,Chromista
FAM_023	2	0	2	335	14	3	0	100	vertebrates	phylum	Chordata
FAM_024	2	2	2	110	81	8	50	0	invertebrates	family	Nymphalidae
FAM_025	2	1	1	100	34	84	0	0	invertebrates	genus	Daphnia
FAM_026	2	0	2	255	12	21	0	100	vertebrates	class	Mammalia
FAM_027	2	2	2	218	71	25	50	0	fungi	genus	Gibberella
FAM_028	2	0	2	101	30	0	0	0	vertebrates	family	Hominidae
FAM_029	2	2	2	89	54	32	0	0	plants	genus	Oryza
FAM_030	2	0	2	314	10	9	50	100	vertebrates	phylum	Chordata
FAM_031	2	0	2	294	20	50	0	0	vertebrates	phylum	Chordata
FAM_032	2	1	1	96	54	6	0	0	protists	genus	Globisporangium
FAM_033	2	2	1	82	100	0	0	100	protists	genus	Globisporangium
FAM_034	2	2	1	44	100	0	0	100	protists	genus	Globisporangium