Do *DNA* lixo ao *DNA* luxo: sequências codificantes diminutas ganham destaque na ciência

Diego Guerra-Almeida[1]. Diogo Antonio Tschoeke[1], Rodrigo Nunes-da-Fonseca[1]

AUTHOR AFILIATIONS

1 – Instituto de Biodiversidade e Sustentabilidade (NUPEM), Universidade Federal do Rio de Janeiro

CONTACT

diegoguerra@ufrj.br

ABSTRACT

The development of large-scale DNA sequencing methods enabled the investigation of eukaryotic genomes previously classified as irrelevant to the adaptive value of species, the so called “junk” DNA. Junk DNA contains several random elements originated by aleatory nucleotide arrangement and rearrangement. Junk DNA also contains a large amount of genomic fossils, sequences that were once active, but that lost their functions in the course of evolution. In this context, small Open Reading Frames (smORFs) are small nucleotide sequences (<100 codons) mainly occurring in a random pattern in genomes. The prediction of smORF-encoded bioactive peptides is extremely challenging and has been largely ignored by gene prediction programs, generating a wide gap in sequence databases. However, the recent improvement of genome and transcriptome sequencing technologies associated to evolutionary, genetic and functional analysis, have demonstrated the existence of several smORF-encoded bioactive peptides in diverse biological species, ranging from insect disease vectors to humans. In this manuscript, we contextualize this serendipity approach towards junk DNA exploration in the search of these new small coding jewels.

Keywords: small Open Reading Frames; Junk DNA, Next-Generation Sequencing; Gene evolution.

RESUMO

O sequenciamento de *DNA* em larga escala possibilitou a exploração de trechos dos genomas eucarióticos classificados como irrelevantes para o valor adaptativo das espécies, o chamado *DNA* lixo. No *DNA* lixo, encontram-se diversos elementos aleatórios, além de diversos fósseis genômicos, ou seja, sequências que perderam funcionalidade no decorrer da evolução. Entretanto, um número crescente de dados evidenciam que no *DNA* lixo podem existir elementos importantes, cujas funções ainda são desconhecidas. Neste contexto, as *small Open Reading Frames* (*smORFs*), ou pequenas janelas abertas de leitura, são sequências nucleotídicas com tamanho diminuto (< 100 códons) que ocorrem, em

sua grande maioria, de forma aleatória nos genomas. Por este motivo, a predição de *smORFs* funcionais é extremamente desafiadora e sua detecção tem sido ignorada por programas de predição gênica nos últimos anos, gerando uma enorme lacuna nos bancos de dados de sequências. Todavia, o recente avanço das tecnologias de sequenciamento de genomas e transcriptomas, acoplados a análises evolutivas, genéticas e funcionais, demonstraram a existência de diversos peptídeos biologicamente relevantes codificados a partir de *smORFs* em diversas espécies, desde insetos vetores de doenças a humanos. Nesta revisão, contextualizaremos esta recente corrida científica de exploração do *DNA* lixo em busca de novas pequenas joias codificantes.

Palavras-chave: *Small Open Reading Frames*; *DNA* lixo, Sequenciamento de Nova Geração; Evolução gênica.

INTRODUÇÃO

Discorrer sobre a imensa relevância que o sequenciamento de ácidos nucleicos possui para a ciência não é uma tarefa trivial. No mais fundamental dos níveis, pode-se dizer que o advento das tecnologias de sequenciamento possibilitou a análise de uma das principais propriedades capazes de orquestrar uma vida terrestre e diferenciá-la de outras: o *DNA* (ou o *RNA*) (Heather & Chain, 2016). Apropriadamente, Frederick Sanger e Walter Gilbert receberam o prêmio Nobel em Química do ano de 1980 por suas contribuições como "pais" do sequenciamento de ácidos nucleicos. Contudo, a área só ganhou forte projeção após o anúncio do sequenciamento do genoma humano, em 2001 (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001).

Os genomas possuem o código genético, o "livro da vida" que define, juntamente a aspectos ambientais, as características básicas dos organismos (Lesk, 2012). Desde o seu primórdio, o estudo dos genomas cresce como um abrangente ramo da ciência, assim como o estudo dos transcriptomas, isto é, das sequencias de *RNA* transcritas em uma célula ou em uma população de células, que atualmente também tem se expandido enormemente. Todos estes avanços foram possíveis graças ao desenvolvimento da bioinformática e o advento das tecnologias de sequenciamento de nova geração, que permitem que novos dados genéticos sejam constantemente produzidos, fomentando a geração novos

conhecimentos funcionais. Aproximadamente 50 mil genomas estão indexados na seção "*Genome*" do *National Center for Biotechnology Information* (*NCBI*) <ncbi.nlm.nih.gov/genome> (Dezembro de 2019), abrangendo uma vasta gama de táxons e favorecendo análises comparativas. Entretanto, este número de genomas ainda é uma parcela muito pequena diante de toda a biodiversidade existente.

Embora os genomas possuam grande diversidade de regiões regulatórias (sequencias que não são necessariamente transcritas e atuam regulando a expressão dos genes) e genes expressos atuando de forma dinâmica dentro dos mais variados sistemas biológicos, nem todo conteúdo genômico, sobretudo em eucariotos, é considerado relevante biologicamente. Estima-se, por exemplo, que apenas 10% do genoma humano seja funcional (Rands et al., 2014), sendo aos cerca de 90% restantes, historicamente, atribuída a controversa definição de *DNA* "lixo" (Sridhar et al., 2011), por não possuírem atividade biológica ou simplesmente por ainda não possuírem suas funções elucidadas. Portanto, ainda que o sequenciamento de genomas e transcriptomas seja fundamental, decifrá-los ainda é uma árdua e necessária tarefa. Neste contexto, as tecnologias de sequenciamento de nova geração (*NGS* do inglês *Next-Generation Sequencing*) abriram as portas para o sequenciamento massivo e profundo de genomas e transcriptomas, permitindo a visualização de verdadeiros "submundos" genômicos antes inexplorados (Shendure et al., 2017). Estudos voltados para regiões de *DNA* pouco exploradas estão em voga na genética e tendem a crescer com o aumento da disponibilidade de genomas e transcriptomas sequenciados, possibilitando a descoberta de novos genes e funções associadas às moléculas de *DNA*, como, por exemplo, os genes de small Open Reading Frames, que serão abordados nesta revisão.

DNA "lixo": um território de enigmas genômicos

Embora controverso, o termo *DNA* lixo pode ser compreendido como regiões genômicas irrelevantes para o valor adaptativo de um organismo, mas que possivelmente em algum momento da história evolutiva assumiram e/ou

assumirão papéis biológicos importantes, podendo abrigar fósseis genômicos (adaptado de Eddy (2012). Na língua inglesa, onde a expressão surgiu, este conceito é traduzido como *junk DNA.* Contudo, o significado da palavra "*junk*" é constantemente confundido com o termo "*garbage*", que se refere diretamente ao que é inútil e descartável, enquanto a palavra "*junk*" é melhor associada, por exemplo, aos objetos antigos acumulados nos fundos de garagens (Eddy, 2012). Esta confusão semântica afastou pesquisadores por muitos anos desta área de estudos e ainda em dias atuais é tópico de intensos debates, mesmo diante dos avanços tecnológicos.

Nos primórdios da genômica, o dogma do "um gene, uma proteína" (Ingram, 1957) promoveu a ideia de que as proteínas eram o cerne das atividades moleculares dos organismos e os *RNAs* meros carreadores de informação genética (revisto por Crick, 1970; Nam, Choi & You, 2016). Atualmente, o conceito de "gene" é atribuído não só às regiões codificantes, mas também às regiões regulatórias e trechos transcritos em *RNAs* não codificantes (*ncRNAs*, do inglês *non-coding RNAs*) (Gerstein et al., 2007). Entretanto, este paradigma contribuiu, historicamente, para a desvalorização das regiões genômicas não precursoras de proteínas, mesmo quando estas eram transcritas em *ncRNAs* (Zanet et al., 2016). Favorecendo este cenário, em 1972, no Simpósio de Biologia de Brookhaven nos Estados Unidos, o biólogo evolucionista Susumu Ohno cunhou o termo *DNA* lixo ao classificar pseudogenes não codificantes de proteína. Porém, rapidamente o conceito passou a ser atribuído a todas as regiões não codificantes do genoma (Palazzo & Gregory, 2014).

Um estudo do consórcio *ENCODE* (*The Encyclopedia of DNA Elements*) sugeriu que 80% do genoma humano possui algum tipo de função bioquímica (Dunham et al., 2012), refutando a existência de grandes aglomerados de *DNA* lixo no genoma. Contudo, o conceito de "função" utilizado pelo grupo gerou controvérsias, pois não estaria necessariamente associado a papéis sob pressão seletiva e com relevância biológica (Doolittle, 2013; Graur et al., 2013). Posteriormente, um estudo mais conservador propôs que apenas 10% do genoma humano seja efetivamente funcional (Rands et al., 2014).

Independentemente destes debates, sugere-se que entre 75 a 90% do genoma humano possa ser transcrito e menos de 2% codifiquem proteínas (Birney et al., 2007; Costa, 2010; Djebali et al., 2012). Embora grande parte deste número de transcritos possa representar produtos estocásticos, estes dados sugerem que genes e processos importantes ainda sejam desconhecidos.

Atualmente, a importância de se estudar os componentes do *DNA* lixo é reconhecida, neste contexto, seus principais representantes a serem reconceituados como biologicamente relevantes são os genes de *ncRNA*, pois a hipótese do mundo de *RNA* (Gilbert, 1986), somada ao avanço biotecnológico, contribuiu para a elucidação funcional de muitas destas moléculas. Entre os diversos processos moleculares em que os *ncRNAs* estão envolvidos, encontram-se a ativação de fatores de transcrição, regulação epigenômica, silenciamento gênico, regulação pós-transcricional e modulação de *splicing* alternativo, além de comprovada relação com o surgimento e desenvolvimento de tumores (Pop-Bica et al., 2017; Hu et al., 2018).

Porém, os *ncRNAs* não são os únicos elementos oriundos do *DNA* lixo a possuírem sua relevância elucidada. Diferentes estudos vêm sendo desenvolvidos nos últimos anos sobre regiões de heterocromatina; zonas repetitivas; elementos de transposição; zonas intrônicas; regiões regulatórias como *enhancers*, promotores e silenciadores (*silencers*); genes em sobreposição; pseudogenes e; mais recentemente, genes de *smORFs* codificantes (do inglês *small Open Reading Frames*) (Balakirev & Ayala, 2003; Okamura et al., 2007; De Koning et al., 2011; Pink et al., 2011; Qin et al., 2015; Burns, 2017; Chugunova et al., 2019; Elkon & Agami, 2017; Podgornaya, Ostromyshenskii & Enukashvily, 2018).

A descoberta de que *RNAs* anotados como não codificantes podem sofrer tradução, canônica ou não, de seus pequenos quadros abertos de leitura (*smORFs*), configura-se como uma das mais desafiadoras novidades da genômica moderna. Deste modo, as *smORFs* são uma nova classe de genes codificantes, com padrões genéticos, evolutivos e funcionais pouco conhecidos, que emerge do submundo do *DNA*

lixo (Nam et al., 2016; Couso & Patraquim, 2017; Lu et al., 2019).

O que define uma *small Open Reading Frame*?

Para definir o conceito de *smORF* é preciso, fundamentalmente, discorrer sobre o significado de *frame* (em português, quadro/janela/fase) e *ORF* (do inglês, *Open Reading Frame*; em português, janela/quadro/fase aberta(o) de leitura).

Em genética molecular, uma *frame* é uma leitura nucleotídica em tripletes (códons) não sobrepostos. Sequências nucleotídicas podem ser lidas em três *frames*, considerando-se, quando aplicável, as outras três *frames* de sua sequência antissenso (totalizando-se seis possibilidades de leitura). A composição de códons de cada *frame* é distinta uma das outras (Figura 1A) (Lesk, 2012).

De acordo com o dogma da biologia molecular, os genes podem ser transcritos em *RNAs*, e *RNAs* podem ser traduzidos em proteínas, podendo haver transcrição reversa. As *ORFs* são um dos principais atores deste processo, pois é sobre elas que a tradução efetivamente ocorre (Figura 1) (Lesk, 2012). Será utilizado neste artigo o conceito mais empregado de *ORF* em estudos sobre *smORFs*, que a define como uma sequência que, em mesma *frame*, seja iniciada por um códon de iniciação da tradução, em geral um ATG, e finalizada pelo códon de terminação da tradução mais próximo, em geral TAG, TGA, TAA (Figura 1A,B). Definições alternativas podem ser encontradas em Sieber, Platzer & Schuster (2018).

Cada códon de uma *ORF* (com exceção dos códons de terminação em traduções canônicas) dá origem a um aminoácido durante a tradução (Figura 1B). Em genes eucarióticos, *ORFs* podem se sobrepor a introns, o que promove sua excisão e possível mudança de *frame* após eventos de *splicing*, podendo modificar completamente a sequência do RNA maduro e da proteína resultante. As *ORFs* efetivamente codificantes presentes em *RNAs* maduros são chamadas de *CDSs* (do inglês, *Coding DNA Sequences*) (Figura 1C).

A principal diferença conceitual entre *ORFs* e *smORFs* é o tamanho (em códons). Contudo, não há um consenso científico que

estabeleça um limiar. O critério amplamente utilizado estabelece que as *smORFs* possuem entre o mínimo limite teórico de dois códons até 100 códons (Andrews & Rothnagel, 2014; Chekulaeva & Rajewsky, 2018). Entretanto, há estudos que definem limites máximos de tamanho entre 150 a 250 códons (Yang et al., 2011; Andrews & Rothnagel, 2014; Chu, Ma & Saghatelian, 2015).

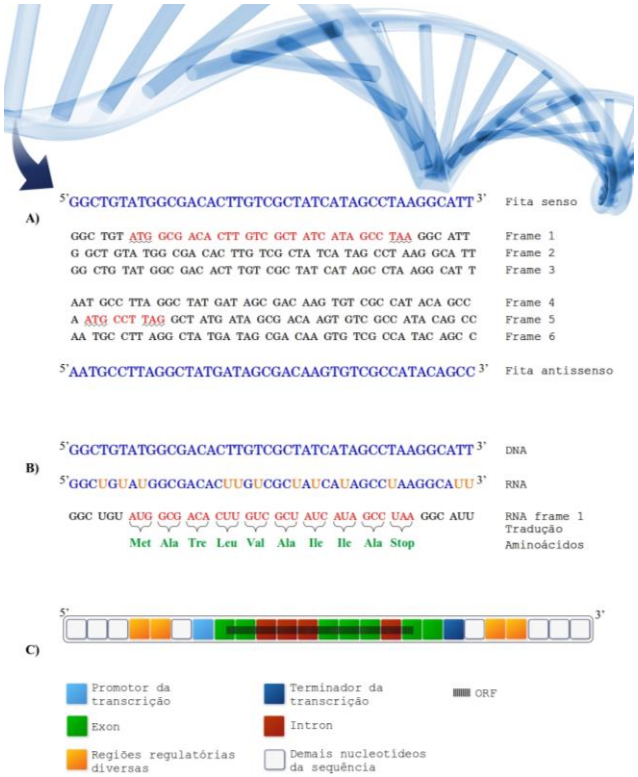

Figura 1 – Conceito, localização e processamento de *ORFs*. **A** – Fita senso e antissenso de uma sequência de *DNA* hipotética e suas respectivas *frames*: Uma fita de *DNA* pode ser dividida em três *frames*, que são divisões em códons (tripletes) não sobrepostos em que a composição de códons de cada *frame* é distinta uma da outra. A fita antissenso é reversa e também complementar à sua respectiva fita senso (evidenciando a organização predominantemente dupla fita do *DNA*) e também pode ser dividida em três outras *frames*. Um gene pode possuir função atrelada ao seu respectivo produto de transcrição da fita senso e também antissenso. A tradução de genes é realizada sobre *ORFs* (em vermelho), que consistem em sequências que, na mesma *frame*, sejam iniciadas e terminadas por códons de iniciação e finalização da tradução, respectivamente (sublinhados). **B** – Modificação de uma sequência hipotética de *DNA* até sua tradução. Em processos de transcrição de um gene em seu respectivo *RNA*, os nucleotídeos de Timina (T) são substituídos por nucleotídeos de Uracila (U). Os *RNAs* produzidos após a transcrição estão sujeitos a sofrer tradução ribossomal de suas *ORFs* (em vermelho), em que cada códon de uma *ORF* codificará um aminoácido de acordo com o código genético universal, com exceção dos códons de terminação da tradução. No momento em que surge um códon de iniciação da tradução em uma sequência, abre-se uma janela de leitura ribossomal, em que os códons podem ser traduzidos em aminoácidos na produção de proteínas (peptídeos), encerrando-se o processo no surgimento do primeiro códon de finalização da tradução presente na mesma *frame*. **C** – Estrutura geral e simplificada de um gene eucariótico. Uma *ORF* pode se localizar em sobreposição a introns em um gene, podendo haver modificação, nos eventos de *splicing*, da *frame* e da composição de códons desta *ORF* após a produção do *RNA* maduro, em que introns são deletados (com exceção de eventos de *splicing* alternativo). *ORFs* podem ser inativadas neste processo ou novas *ORFs* não existentes no gene de origem podem surgir em nível de *RNA*. As *ORFs* efetivamente codificantes de proteína estão integralmente presentes no seu respectivo *RNA* mensageiro maduro, onde passam a se chamar *CDSs*. A maioria das *smORFs* codificantes começam e terminam em exons, não sendo processadas em eventos de *splicing* (Ladoukakis et al., 2011).

Contextualização histórica das *smORFs*

O termo "peptídeo" foi primeiramente atribuído às sequências menores do que 50 aminoácidos oriundas da clivagem de proteínas maiores via mecanismos pós-traducionais e proteólise (Figura 2A) (Albuquerque et al., 2015; Saghatelian & Couso, 2015). Diversos peptídeos e pequenas proteínas ativas biologicamente, como neuropeptídeos, hormônios e antígenos, surgem através destes processos (Albuquerque et al., 2015). Entretanto, sabe-se atualmente que

pequenas sequências peptídicas podem surgir diretamente através da tradução ribossomal de *smORFs* (Figura 2B).

Incontáveis *smORFs* potencialmente codificantes têm sido desconsideradas em consequência de um consenso científico aceito até anos recentes, que sugere que *ORFs* menores do que 100 códons não possuem relevância traducional e biológica (Albuquerque et al., 2015; Saghatelian & Couso, 2015). Consequentemente, milhares de *smORFs* codificantes de polipeptídeos importantes são erroneamente anotadas como não-funcionais ou descartadas como produtos traducionais aleatórios de *ncRNAs*, ocasionando grandes lacunas de conhecimento e escassez de informações nos bancos de dados de anotação de sequências em relação às *smORFs* e seus produtos (Ladoukakis et al., 2011; Chekulaeva & Rajewsky, 2018). Entretanto, diversos polipeptídeos importantes oriundos de *smORFs* têm sido descobertos nos três domínios da vida (revisão em Eukarya: Couso & Patraquim, 2017; estudo em Archaea: Prasse et al., 2015; estudo em Bacteria: Miravet-Verde, 2019) e também em genomas virais (Finkel, Stern-Ginossar & Schwartz, 2018), evidenciando as potencialidades deste território genômico desconhecido.

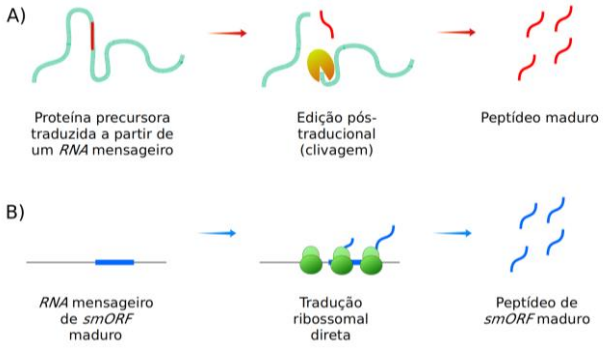

Figura 2 – Esquema simplificado dos processos de produção de peptídeos. **A –** Biossíntese clássica de peptídeos. Hormônios e neuropeptídeos são processados no retículo endoplasmático e complexo de Golgi através da clivagem de longas proteínas precursoras, que são traduzidas por ribossomos a partir de *mRNAs* contendo uma grande *ORF*. Posteriormente, estes peptídeos podem ser secretados a partir de vesículas para ação exterior à célula de origem. **B –** Biossíntese de peptídeos de *smORFs*. Peptídeos de *smORF* são produzidos por tradução ribossomal direta de diversos tipos de *mRNAs* canônicos ou não, incluindo *mRNAs* policistrônicos (*mRNAs* contendo mais de uma *ORF* codificante) e também podem ser secretados para ação exterior à célula de origem.

Desafios preditivos para identificação de *smORFs*

Devido ao seu tamanho diminuto, milhões de *smORFs* podem surgir nos genomas de forma estocástica/aleatória (Ladoukakis et al., 2011). A predição computacional de *smORFs* com potencial codificante em meio a uma miríade de pequenas sequências sem função é uma tarefa extremamente desafiadora (Ladoukakis et al.,

2011; Andrews & Rothnagel, 2014), pois os principais programas de predição gênica utilizam critérios rigorosos de avaliação para eleger uma sequência como potencialmente codificante, dentre os quais, sinais de poliadenilação, códons de iniciação AUG, códons de terminação, sequências promotoras e conservação evolutiva. Contudo, estes padrões ocorrem com menor frequência em genes de *smORFs*, promovendo alto número de resultados falsos negativos (Chu et al., 2015). Além disto, tais algoritmos são desenvolvidos para desconsiderar *ORFs* menores do que 100 códons, por assumir que estas possuam potencial codificante desconsiderável (Olexiouk & Menschaert, 2016; Yeasmin, Yada & Akimitsu, 2018).

A melhor forma de investigar o potencial codificante de uma *smORF* é a conservação evolutiva (Crappé et al., 2013). Entretanto, de acordo com os parâmetros dos principais algoritmos de busca de sequências por similaridade, como o *BLAST* (*Basic Local Alignment Search Tool*) (Altschul et al., 1990), cadeias (sequências) muito pequenas acumulam menor pontuação em termos quantitativos do que sequências maiores, pois a contagem é feita sobre os aminoácidos individualmente, e peptídeos de *smORFs* possuem menor número de aminoácidos. Portanto, a chance de sequências pequenas apresentarem baixos níveis de similaridade, mesmo sendo homólogas, é alta (Couso & Patraquim, 2017). Embora alguns programas já tenham sido desenvolvidos na tentativa de contornar estes problemas, como o *sORF finder* (Hanada et al., 2009) e o *SPADA* (Zhou et al., 2013), há evidências de que a pressão seletiva pode ocorrer apenas sobre a estrutura secundária e terciária dos peptídeos, dificultando ainda mais a detecção de potenciais moléculas através de buscas baseadas em sequência primária (Magny et al., 2013; Zanet et al., 2016).

Outro desafio preditivo descoberto via peptidômica e *ribosome profiling* é a utilização de códons de iniciação não canônicos na tradução de *smORFs*, ou seja, códons diferentes do ATG usualmente encontrado em *RNAs* mensageiros canônicos (Chu et al., 2015; Pueyo, Magny & Couso, 2016). Além disto, estudos de revisão sugerem que peptídeos de *smORFs* podem possuir padrões de composição de aminoácidos

distintos dos encontrados em proteínas maiores usualmente funcionais (Pueyo et al., 2016; Couso & Patraquim, 2017). Evidências da tradução de *smORFs* a partir de *RNAs* atípicos para esta função corroboram a existência de regulações traducionais não canônicas (Zanet et al., 2016; Kim et al., 2018).

A detecção de peptídeos de *smORFs* por análises proteômicas é ineficiente e estocástica, ocorrendo ausência de resultados em até 75% das corridas de um processo (Chu et al., 2015). Os principais métodos de análise proteômica possuem limitações de filtragem para polipeptídeos menores do que 10 kDa. Além disto, a rápida degradação e perda durante os procedimentos de extração e purificação, juntamente aos resultados falsos positivos provocados por peptídeos oriundos da degradação de proteínas maiores, tornam esta abordagem ainda mais desafiadora (Andrews & Rothnagel, 2014; Couso & Patraquim, 2017). Contudo, estima-se que os peptídeos de *smORFs* representem, em média, aproximadamente 0,1% de todo o conteúdo proteico dos tecidos, havendo de 10 a 1.000 moléculas por célula (Slavoff et al., 2013). Outros estudos sugerem que os peptídeos provenientes de *smORFs* representem entre 10 a 20% do conteúdo traduzido em células meióticas de levedura (Hollerer, Higdon & Brar, 2018).

Abordagens genéticas de predição como mutagênese aleatória também podem falhar em introduzir mutações deletérias em *smORFs*, pois sequências diminutas são alvos menos propensos a serem mutados aleatoriamente. Protocolos sistêmicos de mutagênese direcionada também são impraticáveis, pois o número de *smORFs* aleatórias existentes nos genomas é exorbitante. Deste modo, encontrar sequências promissoras é um grande obstáculo para estudos preditivos em larga escala (Andrews & Rothnagel, 2014; Couso & Patraquim, 2017).

Apesar dos desafios preditivos, métodos promissores baseados em sequenciamento profundo de transcriptomas com cobertura ribossomal, como *ribosome profiling* e *poly-ribo-seq*, têm se mostrado altamente promissores na descoberta de novas *smORFs* (Aspden et al., 2014; Weaver et al., 2019), assim como análises de genômica comparativa (Ladoukakis et al., 2011; Mackowiak et al., 2015).

Importantes peptídeos de *smORFs* já descritos

Nos últimos anos, *RNAs* anotados inicialmente como não codificantes têm sido reanotados a partir da descoberta de diversas *smORFs* importantes nestas sequencias, deste modo, muitos *ncRNAs* têm sido reclassificados como genes contendo pequenas sequências precursoras de peptídeos bioativos (*smORFs*) (Aspden et al., 2014).

Um dos principais exemplos é o gene policistrônico *pri*/*mlpt*/*tal* (*polished-rice*/*mille-pattes*/*tarsal-less*, respectivamente), descrito primeiramente no besouro *Tribolium castaneum* como um gene *gap*. Este gene, conservado em Pancrustacea (grupo monofilético contendo insetos e crustáceos), codifica em seu *RNA* mensageiro entre dois a cinco peptídeos de *smORFs* parálogos (duplicados) com motivo LDPTGXY de tamanhos entre 10 a 30 aminoácidos (Savard et al., 2006; Pueyo & Couso, 2008). Este é um dos genes de *smORFs* mais estudados e sua caracterização contribuiu fortemente para a disseminação da importância das *smORFs* como agentes do desenvolvimento em vertebrados (Albuquerque et al., 2015). Mutações e silenciamentos gênicos de *pri*/*mlpt*/*tal* promovem fenótipos embrionários letais que podem ser diferentes em espécies distintas, como por exemplo, alterações em estruturas epidermais na mosca-do-vinagre *Drosophila melanogaster* (Kondo et al., 2007), modificação no número de pernas em besouros e percevejos (Figura 3) (Savard et al., 2006; Ray et al., 2019; Tobias-Santos et al., 2019) e má-formação das regiões do tarso, a extremidade das pernas de artrópodes (Pueyo & Couso, 2011, Tobias-Santos et al., 2019).

Curiosamente, descobertas recentes com o percevejo *Rhodnius prolixus*, o principal vetor da doença de Chagas, demonstraram o surgimento de novas *smORFs* no *RNA* mensageiro de *pri/mlpt/tal* de hemípteras (percevejos), dentre os quais, uma sequência de cerca de 80 códons, altamente conservada, restrita a genomas de hemípteras e com sequência teórica de aminoácidos distinta dos peptídeos contendo o domínio LDPTGXY, previamente descrito em diversas ordens de insetos (Tobias-Santos et al., 2019). Embora a função desta nova *smORF* seja

desconhecida, este achado demonstra como uma abordagem de genômica comparativa pode levar a identificação de novas *smORFs* em genes previamente conhecidos e anotados.

Dois peptídeos de *smORFs* transmembranares chamados de *sarcolamban* A e B (28 e 29 aminoácidos, respectivamente) também foram descobertos através de um transcrito anotado como possivelmente não codificante em *Drosophila*. Estas *smORFs* teriam surgido a 550 milhões de anos, possuindo ortólogos (*phospholamban* e *sarcolipin*) com função conservada em humanos. Estas *smORFs* estão associadas com a regulação do transporte de cálcio no retículo sarcoplasmático, atuando de forma preponderante na contração do miocárdio, sendo alvos de diversos estudos sobre arritmia cardíaca (Magny et al., 2013). Recentemente, dois peptídeos de *smORFs* da mesma família, a mioregulina (46 aminoácidos), envolvida na regulação de cálcio especificamente em músculos esqueléticos (Anderson et al., 2015) e um outro com função antagônica chamado *DWORF* (34 aminoácidos) (Nelson et al., 2016), também foram descobertos e caracterizados em *RNAs* de rato. Estes transcritos foram previamente anotados como *ncRNAs* e atualmente são alvos de

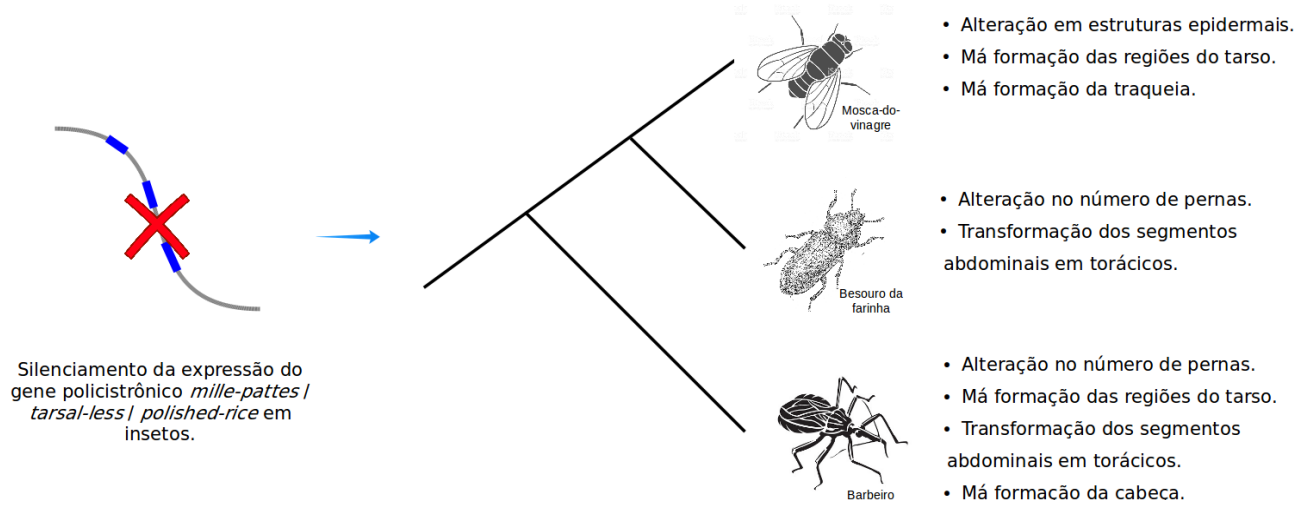

Figura 3 – Diferentes fenótipos descritos em insetos após o silenciamento do gene policistrônico *mille-pattes*/*tarsal-less*/*polished-rice*. Faixas azuis no gene *mille-pattes / tarsal-less / polished-rice* representam as *smORFs* duplicadas. Dados retirados de Savard et al., 2006; Kondo et al., 2007; Pueyo & Couso, 2011; Tobias-Santos et al., 2019.

estudos em desempenho muscular (Chekulaeva & Rajewsky, 2018).

O peptídeo de *smORF* humanina (24 aminoácidos), encontrado em humanos, é codificado pelo *rRNA* mitocondrial *16S* e foi descoberto através de um estudo sobre o mal de Alzheimer que buscava *cDNAs* relacionados à morte celular no sistema nervoso. Após um *cDNA* demonstrar atividade na prevenção à apoptose, descobriu-se que o mesmo codifica um peptídeo nomeado humanina, que interage com o regulador apoptótico *Bax*, prevenindo a ativação desta molécula (Guo et al., 2003).

O *rRNA* mitocondrial *12S* codifica o peptídeo *MOTS-c* (16 aminoácidos). Em ratos, o tratamento com *MOTS-c* previne a resistência à insulina e a obesidade induzida por alimentação, além da reversão da resistência muscular à insulina atrelada à idade (Lee et al., 2015; Kim et al., 2018). Outros exemplos de *smORFs* importantes foram descritas nos últimos anos, como a *hemotin (88 aminoácidos),* que regula a fagocitose a partir do controle da maturação do endossoma em moscas-do-vinagre, e *Stannin* (88 aminoácidos em humanos), descrito como atuante na resposta a metais pesados em vertebrados, descrito como homólogo de *hemotin* (Pueyo et al., 2016), sendo mais um caso de uma importante *smORF* descoberta através de genômica comparativa.

Conclusão

O presente artigo busca destacar para a comunidade científica brasileira uma nova fronteira bastante negligenciada da biologia molecular/genética moderna. Com a grande quantidade de genomas e transcriptomas disponíveis gratuitamente, é possível identificar novos genes contendo *smORFs* que podem se constituir como futuros alvos para controle de vetores e agentes etiológicos de doenças, além de estudos biomédicos. A partir de uma abordagem multidisciplinar, o estudo das *smORFs* encontra-se na fronteira da era pós-genômica, vislumbrando-se grandes avanços nos próximos anos.

Referências Bibliográficas

Albuquerque, J.P. et al. (2015). small ORFs: A new class of essential genes for development. *Genetics and Molecular Biology*, 38(3), 278-283. doi: 10.1590/S1415-475738320150009.

Altschul, S.F. et al. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Anderson, D.M. et al. (2015). A micropeptide encoded by a putative long noncoding RNA regulates muscle performance. *Cell*, 160(4), 595–606. doi: 10.1016/j.cell.2015.01.009.

Andrews, S.J. & Rothnagel, J.A. (2014). Emerging evidence for functional peptides encoded by short Open Reading Frames. *Nature Reviews Genetics*, 15(3), 193–204. doi: 10.1038/nrg3520.

Aspden, J.L. et al. (2014). Extensive translation of small Open Reading Frames revealed by poly-ribo-seq. *eLife*, 3(e03528), 1–19. doi: 10.7554/eLife.03528.

Balakirev, E.S. & Ayala, F.J. (2003). Pseudogenes: are they "junk" or functional DNA? *Annual Review of Genetics*, 37(1), 123–151. doi: 10.1146/annurev.genet.37.040103.103949.

Birney, E. et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146), 799–816. doi: 10.1038/nature05874.

Burns, K.H. (2017). Transposable elements in cancer. *Nature Reviews Cancer*, 17(7), 415–424. doi: 10.1038/nrc.2017.35.

Chekulaeva, M. & Rajewsky, N. (2018). Roles of long noncoding RNAs and circular RNAs in translation. *Cold Spring Harbor Perspectives in Biology*, 3(11). doi: 10.1101/cshperspect.a032680.

Chugunova, A. et al. (2019). *LINC00116* codes for a mitochondrial peptide linking respiration and lipid metabolism. *Proceedings of the National Academy of Sciences*, 116(11), 4940-4945. doi: 10.1073/pnas.1809105116.

Chu, Q., Ma, J. & Saghatelian, A. (2015). Identification and characterization of sORF-encoded polypeptides. *Critical Reviews in Biochemistry and Molecular Biology*, 50(2), 134–141. doi: 10.3109/10409238.2015.1016215.

Costa, F.F. (2010). Non-coding RNAs: meet thy masters. *BioEssays*, 32(7), 599–608. doi: 10.1002/bies.200900112.

Couso, J. & Patraquim, P. (2017). Classification and function of small Open Reading Frames. *Nature Reviews Molecular Cell Biology*, 18(9), 575-589. doi: 10.1038/nrm.2017.58.

Crappé, J. et al. (2013). Combining *in silico* prediction and ribosome profiling in a genome-wide search for novel putatively coding sORFs. *BMC Genomics*, 14(648). doi: 10.1186/1471-2164-14-648.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258), 561-563. doi: 10.1038/227561a0.

De Koning, A.P.J. et al. (2011). Repetitive elements may comprise over Two-Thirds of the human genome. *PLoS Genetics*, 7(12). doi: 10.1371/journal.pgen.1002384.

Djebali, S. et al. (2012). Landscape of transcription in human cells. *Nature*, 489(7414), 101–108. doi: 10.1038/nature11233.

Doolittle, W.F. (2013). Is junk DNA bunk? A critique of ENCODE. *Proceedings of the National Academy of Sciences*, 110(14), 5294–5300. doi: 10.1073/pnas.1221376110.

Dunham, I. et al. (2012). An integrated encyclopedia of DNA elements in the human

genome. *Nature*, 489(7414), 57–74. doi: 10.1038/nature11247.

Eddy, S.R. (2012). The C-value paradox, junk DNA and ENCODE. *Current Biology*, 22(21), R898–9. doi: 10.1016/j.cub.2012.10.002.

Elkon, R. & Agami, R. (2017). Characterization of noncoding regulatory DNA in the human genome. *Nature Biotechnology*, 35(8), 732–746. doi: 10.1038/nbt.3863.

Finkel, Y., Stern-Ginossar, N. & Schwartz, M. (2018). Viral short ORFs and their possible functions. *Proteomics*, 18(10), 1–33. doi: 10.1002/pmic.201700255.

Gerstein, M.B et al. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Research*, 17(6), 669–681. doi: 10.1101/gr.6339607.

Gilbert, W. (1986). Origin of life: The RNA world. *Nature*, 319, 618–618. doi:10.1038/319618a0.

Graur, D. et al. (2013). On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of encode. *Genome Biology and Evolution*, 5(3), 578–590. doi: 10.1093/gbe/evt028.

Guo, B. et al. (2003). Humanin peptide suppresses apoptosis by interfering with Bax activation. *Nature*, 423(6938), 456–461. doi: 10.1038/nature01627.

Hanada, K. et al. (2009). sORF finder: A program package to identify small Open Reading Frames with high coding potential. *Bioinformatics*, 26(3), 399–400. doi: 10.1093/bioinformatics/btp688.

Heather, J.M. & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107(1), 1–8. doi: 10.1016/j.ygeno.2015.11.003.

Hollerer, I., Higdon, A. & Brar, G.A. (2018). Strategies and challenges in identifying function for thousands of sORF-Encoded peptides in meiosis. *Proteomics*, 8(10). doi: 10.1002/pmic.201700274.

Hu, X. et al. (2018). The role of long noncoding RNAs in cancer: the dark matter matters. *Current Opinion in Genetics and Development*, 48. 8–15. doi: 10.1016/j.gde.2017.10.004.

Ingram, V.M. (1957). Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin. *Nature*, 180(4581), 326-328. doi: 10.1038/180326a0.

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921. doi: 10.1038/35057062.

Kim, K.H. et al. (2018). The Mitochondrial-encoded peptide MOTS-c translocates to the nucleus to regulate nuclear gene expression in response to metabolic stress. *Cell Metabolism*, 28(3), 516-524. doi: 10.1016/j.cmet.2018.06.008.

Kondo, T. et al. (2007). Small peptide regulators of actin-based cell morphogenesis encoded by a polycistronic mRNA. *Nature Cell Biology*, 9(6), 660-665. doi: 10.1038/ncb1595.

Ladoukakis, E. et al. (2011). Hundreds of putatively functional small Open Reading Frames in *Drosophila*. *Genome Biology*, 12(11). doi: 10.1186/gb-2011-12-11-r118.

Lee, C. et al. (2015). The mitochondrial-derived peptide MOTS-C promotes metabolic homeostasis and reduces obesity and insulin resistance. *Cell Metabolism*, 21(3), 443–454. doi: 10.1016/j.cmet.2015.02.009.

Lesk, A.M. (2012). *Introduction to Genomics* (2a. ed). Nova York: Oxford University Press.

Lu, S. (2019). A hidden human proteome encoded by 'non-coding' genes. *Nucleic Acids Research*, 47(15), 8111–8125. doi: 10.1093/nar/gkz646.

Mackowiak, S.D. et al. (2015). Extensive identification and analysis of conserved small ORFs in animals. *Genome Biology*, 16(179), 1–21. doi: 10.1186/s13059-015-0742-x.

Magny, E.G. et al. (2013). Conserved regulation of cardiac calcium uptake by peptides encoded in small Open Reading Frames. *Science*, 341(6150), 1116–1120. doi: 10.1126/science.1238802.

Miravet-Verde, S. et al (2019). Unraveling the hidden universe of small proteins in bacterial genomes. *Molecular Systems Biology*, 15(2). doi: 10.15252/msb.20188290.

Nam, J., Choi, S. & You, B. (2016). Incredible RNA: dual functions of coding and noncoding. *Molecules and Cells*, 39(5), 367–374. doi: 10.14348/molcells.2016.0039.

Nelson, B.R. et al. (2016). A peptide encoded by a transcript annotated as long noncoding RNA enhances SERCA activity in muscle. *Science*, 351(6270), 271-275. doi: 10.1126/science.aad4076.

Okamura, K. et al. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell*, 130(1), 89–100. doi: 10.1016/j.cell.2007.06.028.

Olexiouk, V. & Menschaert, G. (2016). Identification of small novel coding sequences, a proteogenomics endeavor. *Proteogenomics*, 926, 49–64. doi: 10.1007/978-3-319-42316-6_4.

Palazzo, A.F. & Gregory, T.R. (2014). The case for junk DNA. *PLoS Genetics*, 10(5). doi: 10.1371/journal.pgen.1004351.

Pink, R.C. et al. (2011). Pseudogenes: Pseudo-functional or key regulators in health and disease. *Rna*, 17(5), 792–798. doi: 10.1261/rna.2658311.

Podgornaya, O.I., Ostromyshenskii, D.I. & Enukashvily, N.I. (2018). Who needs this junk, or genomic dark matter. *Biochemistry*, 83(4), 450–466. doi: 10.1134/S0006297918040156.

Pop-Bica, C. et al. (2017). Understanding the role of non-coding RNAs in bladder cancer: from dark matter to valuable therapeutic targets. *International Journal of Molecular Sciences*, 18(7). doi: 10.3390/ijms18071514.

Prasse, D. et al. (2015). First description of small proteins encoded by spRNAs in *Methanosarcina mazei* strain Gö1. *Biochimie*, 117, 138–148. doi: 10.1016/j.biochi.2015.04.007.

Pueyo, J.I. & Couso, J.P. (2011). Tarsal-less peptides control Notch signalling through the Shavenbaby transcription factor. *Developmental Biology*, 355(2), 183–193. doi: 10.1016/j.ydbio.2011.03.033.

Pueyo, J.I., Magny, E.G. & Couso, J.P. (2016). New peptides under the s(ORF)ace of the genome. *Trends in Biochemical Sciences*, 41(8), 665–678. doi: 10.1016/j.tibs.2016.05.003.

Pueyo, J.I. & Couso, J.P. (2008). The 11-aminoacid long tarsal-less peptides trigger a cell signal in *Drosophila* leg development. *Developmental Biology*, 324(2), 192–201. doi: 10.1016/j.ydbio.2008.08.025.

Pueyo, J.I. et al. (2016). Hemotin, a regulator of phagocytosis encoded by a small ORF and conserved across metazoans. *PLoS Biology*, 14(3), 1–34. doi: 10.1371/journal.pbio.1002395.

Qin, J. et al. (2015). Intronic regions of plant genes potentially encode RDR (RNA-dependent RNA polymerase)-dependent small RNAs. *Journal of Experimental Botany*, 66(7), 1763–1768. doi: 10.1093/jxb/eru542.

Rands, C.M. et al. (2014). 8.2% of the human genome Is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genetics*, 10(7). doi: 10.1371/journal.pgen.1004525.

Ray, S. et al. (2019). The mlpt/Ubr3/Svb module comprises an ancient developmental switch for embryonic patterning. *eLife*, 8(e39748), 1–28. doi: 10.7554/eLife.39748.

Saghatelian, A. & Couso, J.P. (2015). Discovery and characterization of smORF- encoded bioactive polypeptides. *Nature Chemical Biology*, 11(12), 909–916. doi: 10.1038/nchembio.1964.

Savard, J. et al. (2006). A segmentation gene in *Tribolium* produces a polycistronic mRNA that codes for multiple conserved peptides. *Cell*, 126(3), 559–569. doi: 10.1016/j.cell.2006.05.053.

Shendure, J. et al. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550(7676), 345–353. doi: 10.1038/nature24286.

Sieber, P., Platzer, M. & Schuster, S. (2018). The definition of Open Reading Frame revisited. *Trends in Genetics*, 34(3), 167-170. doi: 10.1016/j.tig.2017.12.009.

Slavoff, S.A. et al. (2013). Peptidomic discovery of short Open Reading Frame-encoded peptides in human cells. *Nature Chemical Biology*, 9(1), 59–64. doi: 10.1038/nchembio.1120.

Sridhar, J. et al. (2011). Junker: An intergenic explorer for bacterial genomes. *Genomics, Proteomics and Bioinformatics*, 9(4–5), 179–182. doi: 10.1016/S1672-0229(11)60021-1.

Tobias-Santos, V. et al. (2019). Multiple roles of the polycistronic gene tarsal-less/mille-pattes/polished-rice during embryogenesis of the kissing bug *Rhodnius prolixus*. *Frontiers in Ecology and Evolution*, 7(10), 1–16. doi: 10.3389/fevo.2019.00379.

Venter, J.C. et al. (2001). The sequence of the human genome. *Science*, 291(5507), 1304–1351. doi: 10.1126/science.1058040.

Weaver, J. et al. (2019). Identifying small proteins by ribosome profiling with stalled initiation complexes. *mBio*, 10(2), e02819-18. doi: 10.1128/mbio.02819-18.

Yang, X. et al. (2011). Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. *Genome Research*, 21(4), 634–641. doi: 10.1101/gr.109280.110.

Yeasmin, F., Yada, T. & Akimitsu, N. (2018). Micropeptides encoded in transcripts previously identified as long noncoding RNAs: A new chapter in transcriptomics and proteomics. *Frontiers in Genetics*, 9(114), 1–10. doi: 10.3389/fgene.2018.00144.

Zanet, J. et al. (2016). Small peptides as newcomers in the control of *Drosophila* development. *Current Topics in Developmental Biology*, 117. doi: 10.1016/bs.ctdb.2015.11.004.

Zhou, P. et al. (2013). Detecting small plant peptides using SPADA (Small Peptide Alignment Discovery Application). *BMC Bioinformatics*, 14(335), 1-16. doi: 10.1186/1471-