UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DO BIOLOGIA ANIMAL

Evolution of *Hox* 3'UTR regulation by alternative poly-adenylation and microRNA regulation within twelve *Drosophila* genomes

Pedro Miguel Queirós do Patrocínio Patraquim

MESTRADO EM BIOLOGIA EVOLUTIVA E DO DESENVOLVIMENTO

2010

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS

DEPARTAMENTO DO BIOLOGIA ANIMAL

Evolution of *Hox* 3'UTR regulation by alternative poly-adenylation and microRNA regulation within twelve *Drosophila* genomes

Pedro Miguel Queirós do Patrocínio Patraquim

Dissertação de Mestrado Orientada por:

Dr. Élio Sucena (Faculdade de Ciências da Universidade de Lisboa)

Dr. Claudio R. Alonso (University of Sussex)

MESTRADO EM BIOLOGIA EVOLUTIVA E DO DESENVOLVIMENTO

2010

ACKNOWLEDGEMENTS/AGRADECIMENTOS

Ao *camarada* Igor Guerreiro, com saudade.

I'd like to thank Claudio Alonso, Stefan Thomsen, Richard Kaschula and João Picão-Osório for reviewing this manuscript, and the other colleagues at the Alonso Lab, namely Ana Bontorim, Casandra Villava and Elvira Lafuente, for a very interesting and creative scientific environment during this year.

Ao Élio Sucena pela Ciência e paciência, ao Cláudio Alonso pela dedicação incansável e pela bolsa de assistente de investigação que ajudou ao meu sustento neste ano.

Aos meus pais Rui Paulo e Maria Satanela, à Marta, aos meus tios Luís Carlos e Paula, ao mano André Luís e ao *Zé* Miguel pela ternura.

ao *Cocas*, *Picão*, Sandra *Alice*, Mafalda, Bruce, Elvira, Diana e Emília *Mi* pelo companheirismo incondicional,

ao Rulfo, Drummond, Mingus, *Bird*, Cormac, Ortega y Gasset, Rick Danko, Richard Manuel, Levon Helm, Garth Hudson, Robbie Robertson, *Chico* Buarque de Hollanda, Djavan e José Cardoso Pires pelo conhecimento.

TITLE

Evolution of Hox 3'UTR regulation by alternative polyadenylation and microRNA regulation within twelve Drosophila genomes

RESUMO

O desenvolvimento é um processo biológico generativo que integra a informação genotípica e o seu contexto ambiental para originar um organismo adulto, com um fenótipo particular. Dado que o fenótipo de um ser vivo é a porção do mesmo que está exposta à acção da selecção natural e o mesmo é em grande parte o resultado dos processos de desenvolvimento, torna-se fulcral estudar a evolução das espécies no contexto das mudanças evolutivas que ocorrem nos mecanismos de desenvolvimento.

A evolução dos processos de desenvolvimento dá-se por dois tipos gerais de mudança, de natureza distinta: criação de novos genes efectores (as proteínas) ou mudanças na regulação das mesmas. Vários estudos apontam para a predominância da mudança regulatória enquanto mecanismo para a evolução do desenvolvimento. Entre estes destacam-se os resultados da comparação do complemento proteico do chimpanzé (*Pan troglodytes*) e do humano (*Homo sapiens*), revelando que mais de 90% destas são semelhantes entre as duas espécies. Torna-se no entanto necessário entender de que natureza são estas mudanças regulatórias, para que se possa formular um modelo mecanístico de como o desenvolvimento evolui.

O contexto da descoberta mencionada acima, pouco depois do início da era da biologia molecular pelas descobertas de Jacob e Monod, que demonstraram que as bactérias E.Coli controlam a expressão dos seus genes através de mudanças na transcrição dos mesmos, tornou o nível transcricional o mais estudado como o *locus* evolutivo por excelência. Nesta linha, surgiram vários estudos que demonstraram que mudanças espaciais e temporais na transcrição de genes específicos, nomeadamente nas sequências em *cis* que promovem a mesma, podem ser responsáveis por diferenças morfológicas entre espécies de vários grupos animais, dos insectos aos peixes.

Quanto ao modelo apresentado para a predominância da mudança *cis*-regulatória da transcrição na modificação evolutiva do desenvolvimento, este defente que estas sequências (baptizadas de sequências *enhancer* ou *potenciadoras*) têm um

grande potencial evolutivo, pela sua estrutura compacta e modular (curtas sequências de nucleótidos) escapando assim às consequências deletérias da pleiotropia, assim como pela sua capacidade de recrutar diversos tipos de factores de transcrição (proteínas que promovem o início da transcrição dos genes) de modo combinatório, podendo assim diversificar facilmente o padrão de transcrição dos genes.

No entanto, outros níveis menos explorados de regulação da expressão génica também possuem as mesmas características. O nível pós-transcricional, explorado por este trabalho, é um destes casos. Nos eucariotas, após transcrição de um dado gene codificante, o transcrito de ARN é processado de vários modos, por excisão de intrões não codificantes, assim como por modificações nas extremidades 5' e 3' da molécula (*capping* e poliadenilação) que sinalizam o início e o fim do transcrito, respectivamente. O transcrito processado (ARN mensageiro ou ARNm) é translocado do núcleo para o citoplasma, onde é reconhecido pelos ribossomas, dando-se o início do processo de tradução, que descodifica a mensagem contida no ARNm dando origem a uma proteína. Durante o seu ciclo-de-vida, a concentração de cada tipo de ARNm é também regulada, assim como a taxa da tradução do mesmo, para que a quantidade de proteína produzida seja controlada.

Vários estudos independentes apoiam o modelo de que grande parte da informação que controla estes eventos está contida na 3'UTR (*3'untranslated region* ou região não-traduzida em 3'). Como o nome indica, esta sequência está presente nos ARNm dos eucariotas não com o propósito de codificar uma sequência proteica, mas sim porque contém a informação pós-transcricional necessária para a sua regulação.

Um dos modos pelos quais a taxa de tradução, assim como a concentração de um dado mRNA são reguladas no citoplasma é a ligação a microARNs. Estes últimos são moléculas curtas de ARN, contendo em média 22 ribonucleótidos, que se associam a complexos proteicos no citoplasma para detectar ARNm específicos, por complementaridade de bases com determinadas sequências-alvo nas suas regiões 3'UTR. Após a sua a ligação às 3'UTRs, os microARNs promovem a deadenilação dos ARNm, assim como a paragem do processamento do mesmo pelo ribossoma, tendo assim um efeito negativo na procução de proteínas.

As sequências-alvo dos micro-ARN partilham com os *potenciadores* da transcrição as características, anteriormente mencionadas, que tornaram os últimos nos principais candidatos a promover mudanças evolutivas na expressão génica. São modulares (8 ribonucleótidos), tendo também capacidade de promover eventos de regulação combinatória (cada 3'UTR tem sequências-alvo para microARNs distintos).

Como tal, torna-se claro que a informação *cis*-regulatória pós-transcricional é também candidata a um factor potenciador da mudança evolutiva nos processos de desenvolvimento.

O nosso estudo baseou-se em descobertas recentes referentes à regulação pós-transcricional de genes *Hox*, efectuadas por outros membros do nosso laboratório, assim como por investigadores de outros grupos.

Os genes *Hox* codificam uma família de factores de transcrição que operam durante o desenvolvimento dos animais com simetria bilateral, regulando vários genes-alvo ao nível da transcrição para dirigir programas de desenvolvimento que geram diferentes identidades segmentares ao longo do eixo antero-posterior. Para além desta função conservada, os genes *Hox* já foram implicados em vários eventos de diversificação evolutiva.

Resultados recentes no modelo animal *Drosophila melanogaster* indicam que este grupo de genes é regulado no nível pós-transcricional, tanto por microARNs (do complexo iab4/iab8) como pela geração de diferentes transcritos codificando a mesma proteína, mas contendo isoformas de 3'UTR distintas (geradas por sinais em *cis* que medeiam eventos de poliadenilação alternativa).

Em mais detalhe, os ARNm dos genes *Hox antennapedia*, *abdominal-a*, *abdominal-b* e *ultrabithorax* sofrem poliadenilação alternativa durante o desenvolvimento embrionário de *Drosophila melanogaster*, dando origem a transcritos codificando a mesma proteína mas contendo informações regulatórias distintas. Estas isoformas de 3'UTR são reguladas no espaço e no tempo, e medeiam regulação diferencial por microRNAs.

Neste trabalho, recorremos a análises bioinformáticas para explorar a evolução destes eventos de regulação pós-transcricional dos genes *Hox*, com o intuito de entender quais as forças evolutivas que intervém na evolução das sequências 3'UTR dos genes *Hox* do género Drosophila. Esta informação, gerada *in silico*, será depois usada para guiar uma investigação *in vitro* mais informada, tendo em vista um conhecimento mais profundo da evolução da regulação génica pós-transcricional, e das suas consequências no fenótipo dos animais.

A grande quantidade de ferramentas e dados já disponíveis desde o início da era genómica, permitiram-nos um estudo extenso desta questão.

Recorremos primeiro a doze genomas de espécies do género Drosophila, isolando as sequências 3'UTR dos genes *Hox* referidos acima. Alinhámos depois estas mesmas sequências e observámos que existe uma extensa variação no posicionamento das sequências em *cis* que promovem a poliadenilação alternativa, havendo quase sempre, no entanto, dois sinais alternativos. Estes resultados indicam que a capacidade de gerar duas isoformas de 3'UTR é fulcral para a regulação dos genes Hox em Drosophila, segregando entre as duas informação regulatória diferencial, mas também que existe alguma plasticidade evolutiva no tipo de informação que cada isoforma contém. O tamanho relativo das isoformas também paree ter evoluído substancialmente, apoiando esta ideia.

De seguida, examinámos a estrutura secundária dos ARNs das 3'UTRs dos genes *Hox* para as 12 espécies, e encontrámos um padrão conspícuo de conservação, ao contrário do que acontece ao nível da sequência primária. Isto indica que a realidade tridimentional em que as 3'UTRs destes genes se encontram, ao longo da vida do ARNm, exerce uma pressão selectiva forte para a manutenção de uma estrutura que seja reconhecida pelos reguladores em *trans*.

A regulação por microARNs foi também abordada. Nesta secção, concentrámo-nos em *ultrabithorax*, usando um sofware desenvolvido para o efeito (PITA) assim como informação de expressão dos miARNs para gerar uma lista de candidatos a reguladores pós-transcricionais. Analisámos em seguida a evolução das sequências-alvo que medeiam a regulação pelos microARN-candidatos e encontrámos dinâmicas quantitativas, que sugerem que houve uma mudança significativa nas sequências 3'UTR no sentido de acomodar diferentes potenciais regulatórios. A dinâmica individual destas sequências-alvo sugere outros paralelos com o modelo de evolução transcricional: observámos que existe, tal como no caso dos enhancers, uma sequência-alvo predominante, responsável pela maioria da afinidade da 3'UTR para cada microARN, assim como sequências acessórias que intervêm pouco na regulação e têm uma evolução mais rápida. Na transcrição, os *enhancers* acessórios são funcionais apenas na ocorrência de *stress* ambiental. Dado que os miARNs já foram implicados na robustez do desenvolvimento ao *stress* ambiental, sugerimos que a existência de sequências-alvo acessórias faça parte do mecanismo pelo qual os microARNs exercem esta função.

Finalmente, tentámos formular um modelo geral para a regulação pós-transcricional dos genes *Hox*. Para isto, investigámos a estrutura secundária de todo o transcrito de ARNm de cada *Hox*, nos diferentes contextos gerados pela poliadenilação alternativa. As 3'UTRs parecem ter uma estrutura modular, estando segregadas tridimensionalmente do resto do transcrito. Este resultado apoia a ideia de que a estrutura secundária é fulcral para a regulação dos genes *Hox* e ajuda também a explicar os resultados da comparação evolutiva das estruturas secundárias.

Para além disso, a região onde se encontra o primeiro sinal de poliadenilação parece sofrer uma remodelação na sua estrutura secundária, que afecta a probabilidade de interacção ARNm-microARN ao mudar a acessibilidade das sequências-alvo aí contidas.

Assim, a poliadenilação alternativa parece estar conservada na linhagem Drosophila, apesar de ter diversificado a informação que é segregada para cada uma das isoformas. A evolução da regulação por microARNs parece ter mudado significativamente durante os 60 milhões de anos de evolução destas espécies, e a lista de candidatos que gerámos abre as portas para estudos *in vivo* de evolução do desenvolvimento por mudanças na regulação em *cis* do nível pós-transcricional.

Adicionalmente, a estrutura secundária do ARN das 3'UTRs parece ser muito importante ao longo da evolução, dada a sua conservação, e é um factor que terá que ser tido em conta, aquando de outros estudos do género. Estamos neste momento a realizar testes *in vivo*, tendo em vista a validação destes resultados.

Palavras-chave

Hox, microARN, Evolução do Desenvolvimento, *Drosophila melanogaster*, 3'UTR, Estrutura secundária do ARN.

ABSTRACT

The Hox genes encode a family of transcriptional regulators that operate differential developmental programs along the anteroposterior axis of animal bodies. Regulatory changes affecting Hox gene expression are believed to have been crucial for the evolution of animal body plans. In *Drosophila melanogaster*, Hox expression is post-transcriptionally regulated by microRNAs (miRNAs) acting on target sites located in Hox 3' untranslated regions (3'UTRs). Notably, recent work has shown that during development *Hox* genes produce mRNAs with variable 3'UTRs (short and long forms) in different tissues as a result of alternative polyadenylation; importantly, Hox short and long 3'UTRs contain very different target sites for miRNAs. Here we use a computational approach to explore the evolution of Hox 3'UTRs treated with especial regard to Hox miRNA regulation. Our work is focused on the twelve *Drosophila* species for which genomic sequences are available, and shows, first, that alternative polyadenylation of Hox transcripts is a feature shared by all Drosophilids tested in the study. Second, that the regulatory impact of miRNAs is evolving very fast within the *Drosophila* group, and, third, that in contrast to the low degree of conservation observed at the level of primary sequence Hox 3'UTR regions show very similar RNA topology indicating that RNA structure is under strong selective pressure. Finally, we also demonstrate that alternative polyadenylation leading to the formation of short and long Hox 3'UTRs can remodel the control regions seen by miRNAs by at least two mechanisms: by gradually adding target sites to a short 3'UTR form, as well as modifying the regulatory value of multiple miRNA target sites simultaneously through changes in RNA secondary structure.

Keywords:
Hox, microRNAs, Evolution of Development, Drosophila melanogaster, 3'UTR, RNA secondary structure.

Index

1. INTRODUCTION

1.1) Gene regulation and the evolution of development.

Development is a generative process whereby cells sharing the same genotypic information and with a proximate common ancestry - the unicellular zygote - act in a coordinated manner to organize and distance themselves functionally from each other (division of labour). This in turn produces a mature organism in which the phenotype is thus a direct result of the integrated diversity of cell-types As the phenotype is the fraction of the organism that directly determines its interaction with the environment, and thus the component visible to natural selection, it becomes clear what S.J. Gould meant, in the introduction of the foundational book *Ontogeny and Phylogeny* (GOULD, 1977) when quoting Van Valen (VAN VALEN, 1973): "A plausible argument could be made that evolution is the control of development by ecology*"* (ALONSO, 2008; PATRAQUIM & SUCENA 2008).

At this point, it is important to note that other processes linking genotype to phenotype, like physiology, fall out of this set of concepts. While it is true that development causes physiology - take for instance an adult human being, in which the vital abilities to metabolise toxins or produce proteins are dependent on the liver, an organ generated during embryonic development by specialised cell-types like the hepatocytes– it is also true that physiology can be considered to support development. A good example of this are the extraembryonic tissues of placental mammals, which nurture the developing animal by allowing for nutrient and gas exchanges among other things. These are present only as a means to achieve the physiological viability that allows for and thus in part causes the development of a viable adult animal.

Despite this, one can argue that developmental processes, and the changes in their genotype-to-phenotype mapping properties across generations, can be accountable for a great amount of the variability of life in form and fuction. This old promise of a throrough multi-dimensional view of life by relating embryology to evolution (WADDINGTON, 1957; GARSTANG, 1922; DE BEER, 1971), coupled with the recent good understanding of developmental processes at the molecular level (WILKINS, 2002) has brought the emerging discipline of evo-devo to a prominent position in biology.

In order to address the evolution of development, one has to start by asking what is the nature of developmental change, in the sense of understanding exactly *the quality* of the genotypic changes that elicit different developmental processes, in turn generating novel phenotypes.

In 1975, M.-C. King and A.C. Wilson published a landmark paper in which they present and discuss the results of a comparison between an extended number of proteins pertaining to the species *Homo sapiens* (human) and *Pan troglodytes* (chimpanzee). Given the well-understood differences amongst these species in aspects ranging from anatomy to behaviour, which arose since the rather-recent split from their most-recent common-ancestor (MRCA), it came as a great surprise that the proteins, the macromolecules that act as the main effectors to produce these very distinct phenotypic outcomes, seemed to be almost identical across these species, regardless of the biochemical assessment method in use (KING & WILSON, 1975). This intriguing result proposed in a very persuasive and relatable manner the idea that changes of a regulatory nature, as opposed to changes in the composition of the coding-sequence of genes *per se*, seem to be powerful enough to explain radical differences in phenotype, and thus can be considered as a prominent mechanism by which development evolves.

Given that the understanding of gene regulation was at its start at the time, it comes as no surprise that in the study of how gene regulation affects evolution, the main focus was given to the transcriptional level of gene expression (ALONSO, 2008). This is because the molecular biology field was at that time profoundly influenced by the ground-breaking proposal of the *lac*-operon model for gene expression control (JACOB & MONOD, 1961). According to this work, the first mechanistic model for the regulation of gene expression, bacteria control the quantity and quality of the proteins available in their cells in direct response to different environmental contexts, repressing or de-repressing the *transcription* of genes depending on how necessary or unnecessary their protein products are at a given moment, this being estimated based on environmental cues like nutrient availability. This, coupled with the King and Wilson study mentioned above, prompted extensive research into the evolution of transcriptional mechanisms as related to change in developmental processes.

The most recent incarnation of this paradigm – that the evolution of phenotypic diversity arises by changing the onset and space of developmental transcription of specific genes - relies on the current model for eukaryotic gene expression of protein-coding genes at the transcriptional initiation level. According to this, there are control sequence modules proximal (in *cis*) to each protein-coding DNA sequence, that mediate gene expression by functioning as molecular attractors to specific transcription factors (TFs). TF's (the *trans*-regulatory elements) then bind these cis-regulatory elements, for which they have affinity, in a specific way to recruit the RNA-polymerase II as well as associated co-factors to the transcription start-site. As such, the transcription of a given protein-coding gene is controlled in time and space by *cis*-modules, in turn making these a prime candidate for the evolution of gene expression patterns. In addition, the fact that the transcriptional *cis*-regulatory modules are discrete and can mediate a combinatorial input (*via* different TF's binding close *cis*-elements), would mean that these are modular and capable of great specificity in function, two important characters directly linked to evolvability - the diversifying potential of a genetic system.

Consistent with this framework, there is now a great wealth of cases in which changes in *cis*-regulation of developmental genes were shown to underlie morphological evolution. The examples include diverse groups of metazoans: in freshwater stickleback fish, recently derived from marine populations, the evolution of skeletal reduction seems to rely, in independent instances, on regulatory mutations in a single enhancer which effectively halts the production of a transcription factor (*pitx1*) involved in bone formation (SHAPIRO ET AL., 2004); the evolutionary divergence of larval dorsal hairs within closely-related Drosophilids was shown to have occurred by loss of expression of a TF (*svb*) in a specific manner, caused by *cis*-regulatory changes in transcriptional control regions (SUCENA & STERN, 2000; MCGREGOR ET AL., 2007); the evolutionary novelty of adult lactose tolerance in some human populations, a trait absent in other hominids, was show to have occurred mainly by three single-nucleotide changes in a transcriptional control region residing in one intron of the MCM6, the gene immediately 5' to *lactase* (*lct*), which is the locus that encodes the enzyme responsibe for lactose hydrolization (TISHKOFF ET AL. 2007). Taken together, these results indicate that transcriptional *cis*-regulatory changes can be an important factor driving the evolution of development.

1.2) Gene-specific regulation at the post-transcriptional level.

With the expansion of our understanding of the molecular control of development, it has become clear that the supposed idiosyncrasies that made transcriptional control regions a good candidate for developmental evolution are actually common features shared with other gene regulatory levels (ALONSO & WILKINS, 2005). During DNA transcription in eukaryotes, the nascent pre-mRNA undergoes a series of processing steps that ultimately lead to a mature mRNA, in order to be recognized by the intracellular environment as codifying a tranlatable message. These mainly include addition of a 5'cap (a guanine nucleotide bound by an 5'-5' triphosphate link to the beginning of the transcript), the splicing, trans-splicing, editing and polyadenylation or the transcript (LICATALOSI & DARNELL, 2010) – see **Figure 1**.

Splicing consists of a ribonucleoprotein-mediated (RNP) removal of introns from the nascent RNA, leading to the colinearity between each group of 3 ribonucleotides – the triplet codon – and the protein that will be originated from the message during translation. In some cases, the genome encodes for different proteins within the same gene, and their composition is regulated at this level – a mechanism called alternative splicing (LICATALOSI & DARNELL, 2010). In turn, polyadenylation consists of the addition of multiple adenosine monophosphate (AMP) nucleotides to the 3'-end of the transcribed message, functioning as a protection mechanism against mRNA degradation, mediating successful nuclear export of mRNA messages and aiding in translation efficiency in the cytoplasm (LICATALOSI & DARNELL, 2010).

Like splicing, polyadenylation (See **Figure 1**) can occur alternatively amongst mRNAs transcribed from the same locus; this is mediated by alternative polyadenylation signals lying in the 3'untranslated-region (3'UTR) of the gene in question. Polyadenylation signal sequences (PASs) consist of a U-rich hexamer (the consensus sequence for mammals and insects is UUAUUU), and usually lie close to an upstream AU-rich region that is important for their recognition, as well as a U-rich region downstream of the PAS, where the nascent RNA transcript is effectively truncated and polyadenylated (LICATALOSI & DARNELL, 2010). Genome-wide studies in vertebrates suggest that most genes originate alternatively polyadenylated messages (LICATALOSI & DARNELL, 2010).

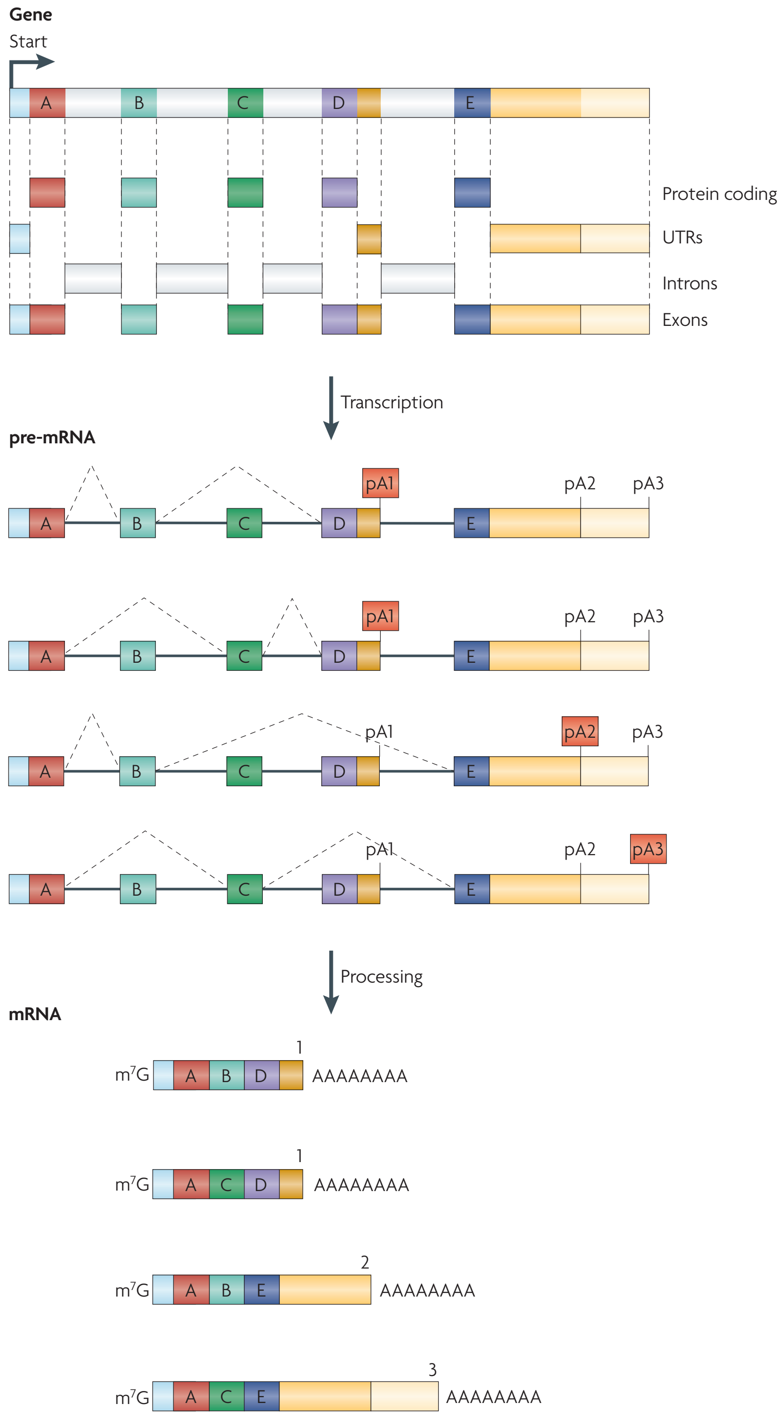


Figure 1 (from LICATALOSI & DARNELL, 2010). Co-transcriptional RNA processing. Here, alternative splicing and alternative polyadenylation, co-transcriptional regulatory events, are depicted. Their ability to generate different proteins (former) of mRNA transcripts with the same protein-coding information but different post-transcriptional control (latter) is depicted.

The regulatory consequences of alternative splicing are conspicuous – the production of different protein products from the same locus allows for diversity and specificity in function without the need to create novel protein-coding materials. In the case of alternative polyadenylation, its effects on gene regulation might be less obvious. In order to better understand the functional consequences of the apparently widespread alternative 3'-end formation, we should first understand the nature of the gene-regulatory information that is conveyed by 3'UTRs.

A revealing, integrated and recent example of 3'UTR-mediated regulation of gene expression is that of the *p27* tumour-supressor mRNA. The product of *p27* mediates cell-cyle arrest and it was shown to be downregulated only in non-quiescent cells by microRNAs (miRNAs) - small RNAs which decrease the probability of translation of a given gene by complementarity-based targeting of its transcript. Also, it had been previously shown that Pumilio-1 (PUM-1), a RBP (RNA-Binding Protein), mediated the downregulation of *p27* in non-quiescent cells after 3'UTR-binding. Proceeding from pattern to mechanism, Kedde and colleagues (KEDDE ET AL., 2010) showed that PUM1 recognizes a particular secondary structure conformation within the 3'UTR of p27; this is a local double-stranded region, a consequence of the base-pairing by neighboiring ribonucleotides of the p27 mRNA. Additionally, this region contains a latent target-site for miRNAs miR-221 and miR-222, protected and thus unaccessible to miRNAs in the normal RNA conformation. Upon PUM-1 binding, the RNA was shown to undergo a change in its local conformation unpairing the double-stranded region aforementioned. It is only then that the p27 mRNA is targeted by miR-221/miR-222, given that the target-region is now single-stranded and thus free to pair with the miRNAs in question.

This case study highlights the diversity of regulatory events that can be mediated by 3'UTRs (miRNA-targeting, RBP targeting) as well as the kind of information that is needed for these events (miRNA/RBP target-sites, secondary structure in the form of accessibility to the regulatory-molecule binding).

Additionally, the noncoding nucleotide sequences of 3'UTRs are also reported to influence gene-expression regulatory steps such as mRNA transcript localization and transport, as in the well-documented case of *gurken* mRNAs in the *Drosophila melanogaster* oocyte, where *cis*-acting sequences in the 3'UTR mediate RBP regulation that effectively localizes the transcript in the dorsoanterior section of the cytoplasm, an event that triggers the definition of the dorsal-ventral axis (MACDOUGALL ET AL., 2003).

As mentioned before, the modular nature of transcriptional enhancer regions of a given gene allows for the evolution of pleiotropy and thus for diversitication in phenotypic outcomes without the creation of *de novo* protein material, as each enhancer might drive transcription in different timepoints/tissues during development; the combinatorial possibilities of enhancers further expand the specificity that can be achieved with this type of gene expression control. These characteristics can be said to be present in 3'UTR *cis*-acting sequences (ALONSO & WILKINS, 2005). miRNA/RBP target-sites also present a modular structure within 3'UTRs, and can mediate a combinatorial regulatory input (ALONSO & WILKINS, 2005). Because of these properties, post-transcriptional regulatory steps are considered to possess an evolutionary potential as an agent of gene expression diversification at least comparable to transcription (ALONSO & WILKINS, 2005). Additionally, and unlike transcriptional control regions, 3'UTRs are discrete entities defined by PAS positioning, making them a more manageable model to tackle the evolution of gene regulation.

3'UTR regulation by miRNAs involves 3'UTR target-recognition and direct binding based on base-complementarity between the miRNA sequence and the 3'UTR of the targeted genes. Analysis of this regulatory event thus presents the possibility of a greater predictive value than TF-enhancer interactions, where the *cis*-sequences recognized by protein regulators appear to be less well-defined.

In the next section, we will develop the notion of miRNA regulation as a good and manageable candidate for the evolution of gene expresion patterns, integrating this regulatory step with other post-transcriptional events.

What are microRNAs?

microRNAs or miRNAs are a recently discovered family of endogenously expressed single-stranded RNA molecules, pervasive in multicellular eukaryotes. These small RNAs are 21-24 ribonucleotides long and act on the translation of mRNAs to negatively regulate gene expression of protein-coding genes (BARTEL, 2009).

miRNAs are produced from transcripts generated by the RNA Polymerase II complex, after a series of stepwise processing steps. First, the immature transcript is capped, polyadenylated and spliced, as other RNA Pol II products, yeilding a primary transcript (pri-miRNA), which can range from hundreds up to thousands of ribonucleotides in length. As other RNA molecules, the pri-miRNA molecule is stabilized by the generation of minimum free-energy conformations, achieved by the formation of mostly local secondary structures, although long-range base-pairing is possible between ribonucleotides, distant at the primary sequence level.

The region of the transcript that will give rise to the miRNA is approximately 70-100 ribonucleotides in length and folds into a stereotypical secondary structure, the stem-loop, which is recognized within the pri-miRNA by the *microprocessor complex*, comprising the nuclear proteins *drosha* and *pasha*. *drosha*, an RNAse, cleaves the stem-loop structure, effetively separating it from the rest of the pri-miRNA. This processing step yields a double-stranded RNA haipin molecule of about 65-70 ribonucleotides in length, the pre-miRNA (BARTEL, 2009).

The pre-miRNA is then exported to the cytoplasm by Exportin 5, a nuclear membrane protein that recognizes a two-nucleotide overhang typical of Drosha pri-miRNA processing. The nulcear export event is energy-dependent, relying on cofactor Ran-GTP. In the cytoplasm, pre-miRNAs are recognized by Dicer, a RNAse of the RISC complex (RNA-induced silencing complex) which interacts with the 3'end of the pre-miRNA molecule to recognize and cleave its characteristic loop. This cleavage step generates a double stranded RNA molecule, composed of a miRNA strand and a complementary sequence (BARTEL, 2009).

Only the mature miRNA is loaded into the RISC complex, and this selection is apparently based on its greater thermodynamic instability. The other strand, called passenger or star (*) strand, is usually degraded shortly after the mature miRNA strand choice. Nevertheless, star sequences have been shown to be used as functional miRNAs in some cases, indicating that other regulatory steps might act on RISC miRNA selection (BARTEL, 2009).

How do miRNAs regulate their targets?

Another member of the RISC complex Argonaute (Ago), binds the miRNA and directs it to an accessible region of the RNP complex, where the miRNA will function to recognize its target mRNAs by base complementarity (BARTEL, 2009). The majority of described functional miRNA targets lie within the 3'untranslated region (3'UTR) of the regulated mRNAs (MAJOROS & OHLER, 2007).

It is important to note that there is assymetrical importance within the 21-24 miRNA sequence in respect to target-recognition, so that the ribonucleotides in the 5'end-most region, more specifically in positions 2-8, have been shown to be more important for target-recognition. This region, called the miRNA *seed*, functions as an anchor that acts co-operatively with the flanking miRNA region to induce a zip-like nucleation event, effectively binding the target mRNA to the RISC complex. After target recognition by the RISC complex, two consequences are currently well-suported:

1) The target mRNA is degraded, either directly by Ago or indirectly, *via* RISC-bound co-factors, or
2) The translation of the target mRNA is halted (VALENCIA-SANCHEZ ET AL. 2006).

As such, miRNA regulation acts to repress the expression of target genes by interfering negatively with the production of a protein product.

Genome organization and evolution of miRNA genes

The most recent estimates of miRNA gene number, based both on expression and bioinformatic analyses place this class of gene regulators amongst the most represented in eukaryotic genomes. miRNAs are thought to represent 1-5% of all animal genes (NIWA & SLACK, 2007). For instance, the human genome harbours more than 1000 miRNAs, while *Drosophila melanogaster* has 171 miRNAs as compared to about 3.000 protein coding genes. Their pervasiveness, as well as the abillity to regulate multiple targets – pleiotropy -, a result of the small miRNA *seed* size needed to recognize targets within mRNA molecules. Also, the number of miRNAs in a given genome appears to be tightly correlated with the complexity of that organism (KOSIK, 2009). Taken together, these facts make this class of regulators a good candidate to explain gene regulatory events during both ontogeny and evolution.

The Drosophila genome is estimated to habour more than 110.000 sequences that are predicted to fold into a miRNA-like hairpin if transcribed (LU ET AL., 2008b). Given that most of the genome is known to be transcriptionally active and that 90% to

98.3% of the miRNAs in Drosophila appear to have originated from non-miRNA sequences (instead of miRNA-gene duplication, for instance) (LU ET AL., 2008b) it remains obscure why only 171 of these sequences are detected in RNA-sequencing experiments. One possible explanation is that additional regulatory steps, other than the typical miRNA stem-looped secondary structure, are necessary for a RNA region to be recognized by the miRNA-processing machinery as a valid regulatory sequence. This might include the requirement of a strong polyadenylation sequence in the transcript that carries the putative miRNA.

Comparative genomics approaches using the recently sequenced genomes of 12 Drosophila species have helped understand on the dynamics that underlie miRNA evolution. These DNA sequence databases comprise species that diverged from a common ancestor around 60 million years ago (60 Mya). This, along with the to RNA expression datasets, it has been recently estimated by various research groups that there are 0.3 to 1 novel miRNAs appearing every 1 million years within the Drosophilid lineage (LU ET AL., 2008b). Of these, apparently only 2.5% to 4% become fixed in the genomes in the long-run (LU ET AL., 2008a), the average half-life of each novel miRNA being 1.96 Myr. This points to a high turnover of genetic material at the level of miRNA loci.

Nevertheless, the overwhelming majority of conserved miRNA loci within the drosophililds also shows that if miRNA regulation influences the evolution of development within this group, this is expected to occur mostly by changes in *cis*-regulatory modules of the targeted 3'UTRs, consistent with the model presented above for the predominance of regulatory evolution in transcription. After the appearance of a novel miRNA, it is expected that a period of strong natural selection for or against the targeting of specific mRNAs follows – the target selection step. Ultimately, the effects of the novel target-site interactions on organismal fitness will decide whether the novel miRNA locus is kept or lost from the genome.

Few studies have focused on the evolution of miRNA-target sequences. This stems from the miRNA target-site prediction techiques, that mostly rely on target-site conservation as well as miRNA-mRNA complementarity (BARTEL, 2009; KERTESZ ET AL. 2007), a fact that artificially steers research on miRNA regulation from accessments on their role on evolutionary diversification. 3'UTR targeting by miRNAs is predicted to be widespread in metazoans (KOSIK, 2009), and significantly correlated with both complexity and multicellularity (KOSIK, 2009). Recent studies have shown a link of miRNA regulation with other 3'UTR regulatory events: alternative polyadenylation is predicted to significantly affect the miRNA target-site content of each 3'UTR isoform (MAJOROS & OHLER, 2007), in terms of number and pattern of target-site distribution. Also, the accessibility of miRNA target-site within the 3'UTR, a function of secondary structure, has been shown to a strong predictor of the success of miRNA regulation, as highlighted by the p27 study mentioned above (KERTESZ ET AL. 2007). 3'UTR length, like miRNA gene number, is also positively correlated with complexity in multicellular organisms (CHEN ET AL., 2010), pointing to a role for the expansion of miRNA regulatory information in 3'UTRs as a putative mechanism underlying the evolution of gene expression.

As such, the analysis of 3'UTR evolutionary dynamics of miRNAs target-site diversification for conserved miRNA loci, as well as its relationship to RNA secondary structure and alternative polyadenylation, is expected to provide a representative and informative view of post-transcriptional evolution in Drosophilids.

1.3) Hox genes.

Hox genes encode a family of transcription factors that operate during the early development of bilateral metazoans, driving the expression of through a myriad target genes (PEARSON ET AL., 2005) to instruct developmental programs that generate differential identities along the segments of the anterioposterior axis.

The members of this gene family are identified by both structural and functional characteristics. First, Hox genes bear a stereotypical 180 nucleotide sequence called *homeobox* within the 3'-most protein-coding exon that encodes for a 60 aminoacid helix-turn-helix DNA-binding protein motif, the homeodomain. This portion of the Hox protein products is responsible for the recognition of transcriptional regulatory sequences of target-genes in *cis*, acting alone or in coordination with other transcription factors.

Although the homeobox is not exclusive to Hox genes, the misregulation of Hox gene expression can generate transformations of one body segment into the likeness of another, a class of phenotypes called homeotic from which the homeodomain derives its name.

In 1915, Calvin Bridges discovered the first of such homeotic transformations, the *Ultrabithorax* phenotype in the dipteran *Drosophila melanogaster* (D.mel.). While the wild-type animal bears one pair of wings in the second thoracic segment (T2) and a pair of reduced flight-control organs called halteres in T3, mutations that affect regulatory regions of the Hox gene *ultrabithorax* (*Ubx*), change the segmental identity of T3 to that of its immediately anterior neighbour, by loss of *Ubx* protein expression in this segment, rendering a homeotic fly with two sets of wings. As the *Ubx* expression domain extends from the posterior compartment of T3 to the anterior portion of the first abdominal segment (A1), these experiments established a role of *Ubx* in the control of segment identity. Another example of homeosis by Hox misexpression is the head to thorax partial transformation, achieved by mutations that induce ectopic expression of the Hox gene *antennapedia* (*antp)* in the head, instead of its wild-type expression domain in the T2. The two most marked morphological characteristics of the T2 segment are the presence of both a pair of wings laterally, as stated above, and a pair of legs in a more ventral position. When *antp* is expressed in the early head development, the primordia that differentiate the head appendages change their idenity to that of the T2 segment, rendering an adult fly with two legs in place of antennae. The converse experiment, that of *antp* loss-of-expression in its wild-type expression domain renders an adult with two antennae in the T2 segment, in place of legs.

Homeosis by Hox misexpression is also known to ocurr in birds, reptiles and mammals (PEARSON ET AL., 2005; GILBERT, 2010). Taken together, these results solidify the notion that Hox genes act as selector genes (LEWIS, 1978) that dictate segment identity in bilaterian animals.

Evolution of Hox genes

The aforementioned conservation of homeotic effects upon Hox misexpression points to a conserved and ancestral role for Hox genes in providing positional information across the anteroposterior axis in bilaterians. Perhaps paradoxically, Hox genes have also been shown to be involved in the diversification of developmental strategies across animal evolution.

In arthropods, the diversification of Hox developmental gene expression patterns has been shown to be directly involved in evolutionary innovations. The dipterans share a common ancestor with a evolutionary innovation in body-paln, wherein the T3 segment - which in other adults arthropods exhibits a set of wings like the T2 – gives rise to halteres. As mentioned before, *Ubx* is involved in haltere specification, and when its gene product is absent, dipterans develop an ectopic set of wings in the T3. In lepidoptera however, a large insect order that includes butterflies and moths, the adults have two sets of wings. As in Drosophila melanogaster – a dipteran - *Ubx* expression was shown to occur in T3 imaginal discs, implying that the developmental changes leading to the generation of an haltere in T3, instead of wings, is mainly the result of the way the Hox message is interpreted by downstream targets (WARREN ET AL., 1994).

Changes in Hox expression were also shown to be correlated with the evolution of the crustacean body-plan. The thoraxic segments that do not express *Ubx* and *abdominal-a* give rise to maxillipeds, and the evolution of the *Ubx*/*abd-a* thoraxic expression domains, a regulatory change, gives rise to adult animals with 0 to 3 thoraxic maxillipeds (AVEROF & PATEL, 1997). This shows that the evolution of Hox developmental expression patterns themselves can drive evolutionary change in morphology.

In the case of onychophora, an edysozoan phylum that includes animals with many repeated pairs of abdominal legs, *Ubx* protein products were shown to lack limb-repression function. As such, the posterior embryonic expresison of *Ubx*, a pattern shared with dipterans, still allows for limb-formation. This evolutionary change was shown by two groups to lie in the carboxy-terminal domain of *Ubx*, which shows a novel abdominal-limb repression domain in insects (GILBERT, 2010). This example highlights the fact that changes in Hox proteins can also cause the evolution of development, leading to changes in morphology (GILBERT, 2010; HUGHES & KAUFMAN, 2002).

The mechanistic aspects of Hox-related evolution of development in arthropods can thus be categorized as:

a) evolution in *cis*

1) *Changes in Hox protein-sequences, eliciting differential developmental programs.*

2) *Evolution of Hox cis-regulatory sequences, eliciting evolutionary Hox expression changes in time and space.*

b) *Evolution of Hox target-genes, wherein the same Hox code is read differently in different species* (GILBERT, 2010; HUGHES & KAUFMAN, 2002)

1.4) Post-transcriptional regulation of Hox genes in Drosophila melanogaster.

Recentlty, Hox genes were shown by colleagues in the host lab to produce alternatively polyadenylated transcripts in a developmentally controlled way (spatially and temporally) in *Drosophila melanogaster* (THOMSEN ET AL., 2010). In stage 10 embryos, in-situ hybridizations for the 3'UTRs of Hox genes *Ubx*, antennapedia (*antp*), abdominal-a, (*abd-a*) and abdominal-b (*abd-b*) show that a short constitutive 3'UTR is present. In stage 15, however, there is expression of a transcript bearing a longer 3'UTR in the CNS (a change from 951 to 2.400 nucleotides in *Ubx*, for example) in all of the analysed Hox (THOMSEN ET AL., 2010).

The different 3'UTR isoforms are predicted to harbour different miRNA-regulatory information, and their establishment is independent of miRNA regulation, in the case of *Ubx* (THOMSEN ET AL., 2010). These results indicate that the developmental 3'UTR remodelling of Hox genes is a general molecular strategy in *Drosophila melanogaster*.

Additionally, miRNAs from the iab-4 and iab-8 loci, transcribed from complementary strands of the same genomic region (TYLER ET AL., 2008; STARK ET AL., 2008; BENDER, 2008) were shown to downregulate *Ubx* expression when ectopically expressed in the halteres (RONSHAUGEN ET AL., 2005) and cell-cultures (TYLER ET AL., 2008). Since these miRNAs are co-expressed in time and space with *Ubx* (THOMSEN ET AL., 2010.), during the developmental stages mentioned above, and were shown to change *Ubx* expression patterns in the embryo (BENDER, 2008), this raises the possibility that alternative *Ubx* polyadenylation might elicit differential visibility of Hox transcripts to miRNA regulation, and supports the more general notion that post-transcriptional regulation, via miRNA regulation of distinct 3'UTR isoforms, underlies developmental expression patterns of Hox genes and can be responsible for the establishment of phenotypes in ontogeny on their change during evolution.

1.5) Biological question and tools.

Based on the previous results mentioned in section 4), as well as the concept of post-transcriptional regulation as a plausible mechanism for evolutionary change in development espoused in sections 1) to 3), we developed a bioinformatic approach to answer the following biological question:

What are the evolutionary forces driving 3'UTR function in the Drosophila lineage?

With the goal of generating experimentally-testable hypotheses, we proceeded to the exploration of the following specific points:

1) Is there <u>conservation</u> across the Drosophila genus of the <u>alternative polyadenylation signals</u> (PAS) that were shown to generate disting 3'UTR isoforms in *Drosophila melanogaster*.
2) <u>Is 3'UTR length conserved</u> in Drosophila Hox genes? Does this translate into the <u>conservation</u> of the distal/constitutive <u>3'UTR isoform length ratio</u>?
3) Is <u>secondary structure of Hox 3'UTRs</u> predicted to be <u>conserved</u> in Drosophilids? Does this pattern mirror the <u>conservation at the primary sequence level</u>?
4) How is <u>miRNA-targeting evolution in *cis*</u> predicted to have evolved in Hox 3'UTRs within the Drosophila lineage?
5) Based on the *in silico* results of points 1-4, <u>can we formulate a model for the evolution of *cis*-regulation in Hox 3'UTRs</u> that integrates alternative polyadenylation, transcript length, secondary structure and miRNA targeting? How can this model be <u>tested *in vivo*</u>?

This wide set of questions is tractable in the one year time-frame provided for the MSc thesis, only if we use a bioinformatic approach. The whealth of freely-available computational tools and datasets should allow for an extensive and informative study of these questions, and include:

1) Precomputed whole-genome alignments of the 12 Drosophila genomes (UCSC Genome Browser).
2) Conservation-based predictions of miRNA target-sites.
3) RNA secondary structure prediction alghorithms (HOFACKER, 2003).
4) High confidence multiple-alignments tools for noncoding regions (BRUDNO ET AL., 2003).
5) Quantitative miRNA target-site predictions without assumptions of target-sequence conservation, that incorporate the target mRNA accessibility predictions (KERTESZ ET AL. 2007).

In the next section, we explore how, using these tools, we generated plethora of results that specifically address the aforementioned set of questions.

2. METHODS

Hox 3'UTR primary-sequence alignments for 12 Drosophila species.

We retrieved the 3'UTR sequences homologous to that of *Drosophila melanogaster's Hox 3'UTR* regions from the UCSC Genome Browser. We then obtained our own alignments of these sequences using the LAGAN algorithm embedded in the VISTA tools (BRUDNO ET AL., 2003). This algorithm first identifies blocks of homology across the whole sequences, which it uses as anchors, after which it proceeds to align the remaining stretches of sequence. This method is particularly suitable for noncoding genomic regions, given that these are known to be subjected to highly assynchronous evolutionary change, with conserved "islands" immersed in rapidly changing sequences undergoing neutral evolution.

miRNA targeting evolution of *Hox* 3'UTRs

For miRNA target-site predictions, we used the PITA algorithm (http://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html) (see **Figure 2**). This software does not base miRNA target-site prediction on sequence conservation premises, unlike most miRNA-prediction tools, allowing for a study of miRNA-regulatory diversification. It also has output values ($\Delta\Delta G$) both for individual sites and whole transcripts (net $\Delta\Delta G$) (KERTESZ ET AL. 2007)..

This tool takes into account local RNA accessibility, a relevant secondary structure characteristic, as it has been shown to significantly influence mRNA regulation by both miRNAs and mRNA-binding proteins. The accessibility value of a given region is ascribed by PITA as a $\Delta Gopen$ value, the amount of free energy that is lost by unpairing the local double-stranded RNA structures. Thus, the more positive the $\Delta Gopen$ value the more locally accessible a region of an RNA sequence is. The software then proceeds to calculate the free energy gained by the miRNA-mRNA duplex ($\Delta Gduplex$), subtracting the first to the latter to obtain an energy-based miRNA regulatory value ($\Delta\Delta G$) (**Figure 2**).

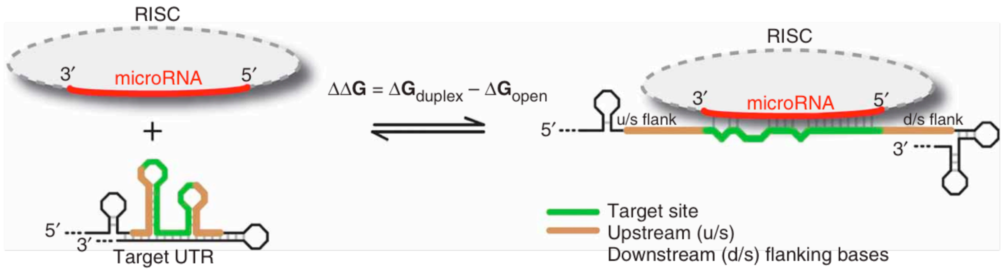


Figure 2 (From KERTESZ ET AL. 2007). The PITA miRNA-targeting prediction energy-based parameters by Kertesz et al. 2007. miRNA-mRNA complementarity and target accessibility matter for the affinity of the interaction.

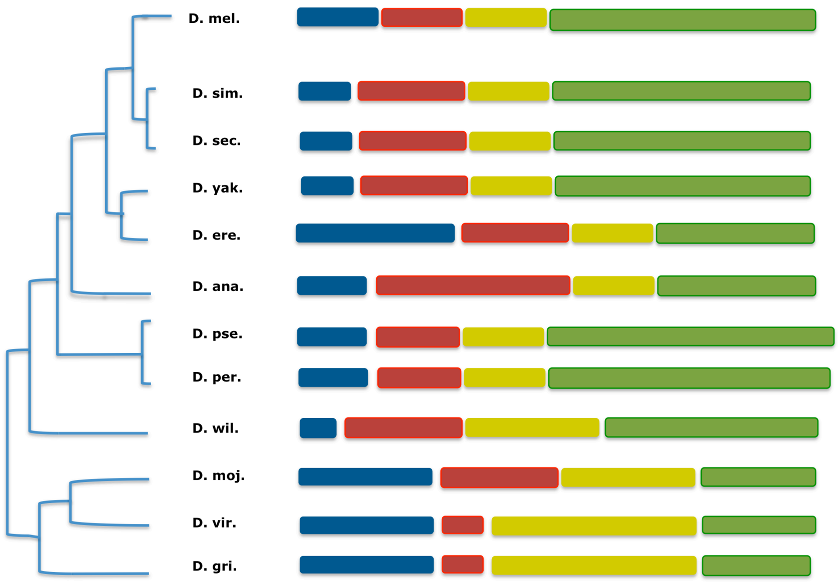


Figure 3. To perform the 3'UTR accessibility alignments, we defined for every species sequences that were homologous to each 200bp window of the *Drosophila melanogaster Hox* 3'UTRs. This allowed us to successfully align the 3'UTR secondary structures of these different species (See Figure 4).

We scanned all the Drosophilid 3'UTR sequences with the PITA online tool only against miRNAs with seeds absolutely conserved across all species. We chose to scan the 3'UTRs for miRNA seeds 6-8 nucleotides in size, allowing for single G:U wobbles and single mismatches in the case of 8 nucleotide seed sequences, since this variability in targeting properties was show to exist *in vivo*. The accessibility of flanking regions was also considered, since it was experimentally shown by (KERTESZ ET AL. 2007) that this significantly improves the algorithm prediction accuracy.

From the PITA outputs, we selected those miRNAs that had ultraconserved *seed* sequences across Drosophilids (freely available information from http://www.miRBASE.org; and (RUBY ET AL. 2007). When information on miRNA conservation was missing or contradictory among these sources, we performed BLAST searches (http://flybase.org/blast/) for both the *Drosophila melanogaster* pri-miRNA and seed sequences. The sequence results with low BLAST E values were folded using the RNAFold algorithm WebServer, part of the Vienna RNA analysis package (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) (HOFACKER, 2003). The stereotypical stem-hairpin structure was used to validate if a given hit was in fact the homologous miRNA in question, as was a minimum free energy value of -20.00 kcal/mol for the 2D RNA structure (RUBY ET AL. 2007), both being outputs of RNAFold (HOFACKER, 2003).

Also, we further refined our results by retrieving only the target-site hits for miRNAs temporally coexpressed with *ubx* during *Drosophila melanogaster* development, based on the Northern-blot information from (RUBY ET AL. 2007).

After this, we undertook a final sieving step by choosing from the remaining miRNA list the ones that showed target-site $\Delta\Delta G$ values above the ones for miR-iab-4/miR-iab-8 miRNAs. This conservative cutoff value was chosen because these miRNA species were already shown to regulate *ubx*, thus lending more confidence to the predictions.

Accessibility alignments for the Drosophilid *Hox* 3'UTRs

We used the 3'UTRs of *antp*, *ubx*, *abd-a* and *abd-b* against the PITA algorithm for all 12 sequenced Drosophilids and retrieved only the $\Delta Gopen$ values across the nucleotide positions of the whole sequences. Since all these sequences independently suffered extensive *indel* and substitution mutations, we generated a correspondence table of homologous regions across these sequences at the primary sequence level, using for this the VISTA-LAGAN output.

For this, 200bp windows of homology were ascribed across Drosophilid *Hox* 3'UTRs as compared to the baseline *Dmel* sequence. This enlarged window allowed us to compare the evolution of accessibility by safely ascribing overall homologous regions despite significant nucleotide divergence (see **Figure 3** for graphical representation of the rationale used for this step of the analysis).

We then calculated the average accessibility values within these 200 bp windows across the whole 3'UTR of each species and plotted the 12 results for a given *Hox* gene. As a negative control for secondary structure evolution, we performed the aforementioned analysis for an intergenic region of the bithorax complex (3R: 12604500-12607000), chosen based on lack of transcriptional activity (information given by the RNA-seq data freely availabe at www.flybase.org), and with the same size as the *Ubx* 3'UTR.

Secondary structure predictions of whole *Hox* mRNA transcripts.

For this, we retrieved the mRNA sequences for *ubx*, *antp*, *abd-a* and *abd-b*, from http://www.Flybase.org, for both *short* and *long* 3'UTR isoforms, after which we used the RNA secondary structure prediction tool RNAFold to generate a visual interpretation of the most stable RNA structure for these different *Hox* mRNAs, using the RNAFold default settings.

3. RESULTS

3.1) Evolution of alternative polyadenilation of Hox genes: conservation and plasticity.

Hypothesising on the evolutionary constraints acting on *Hox* 3'UTR sequence length through gain or loss of the polyadenylation signal sites (PASs) - the modules that define transcript length - we performed and analysed multiple primary-sequence alignments for *ultrabithorax*, *antennapedia*, *abdominal-a* and *abdominal-b* 3'UTRs of the 12 sequenced drosophilids.

With the 3'UTR multiple alignments, we asked whether the sequences for the known functional *Drosophila melanogaster* poly-adenilation signal sequences for the four analysed *Hox* genes were conserved across the Drosophilids.

1) Abdominal-B exhibited ultraconservation of both the first and the second polyadenilation signals (both corresponded to the canonical hexamer AATAAA), while the conservation was less obvious in other *Hox* (**Supplementary Figure S1A**).

2) Antennapedia exhibited an ultraconserved second poly-adenilation signal (henceforth refered to as PAS2) while PAS1 was not found in an exactly homologous position: there is an AT-rich region of approximately 50 bp within which PAS hexamers appear across the analysed species (**Supplementary Figure S1B**).

3) Ultrabithorax 3'UTRs also presentes this pattern, only it is PAS2 and not the first PAS that appears as a *floater* site. In this case, species pertaining to the melanogaster species subgroup (*D. melanogaster*, *D. simulans*, *D. sechelia*, *D. yakuba*, *D. erecta*) appear to retain both *Drosophila melanogaster* PAS sites while more distantly-related species have a very conserved region approximately 200bp upstream of the *Dmel* PAS2 that includes a perfectly conserved AATAAA sequence (refered to here as putative upstream PAS or puPAS). This indicates that the known functional PAS2 is a recent evolutionary novelty. We also found some individual secondary losses of the most conserved PAS sequences in *ultrabithorax*. *Drosophila sechelia* and *Drosophila simulans* share, despite a great degree of similarity with Dmel across the *Ubx* 3'UTR, a CATAAA hexamer in the PAS2 site; the puPAS site of *Drosophila wilistonii* was also distinct from consensus, in this case presenting a GATAAA hexamer (**Figure 4A**).

4) Abdominal-A PAS sites share the overall *ubx* evolutionary pattern, with PAS1 ultraconserved, while containing an homologous region of around 120 bp within which different specific regions emerge as poly-adenilation signaling hexamers through evolutionary time. These results point to the conservation of alternative polyadenilation as a mechanism while allowing for plasticity regarding its precise rules (**Supplementary Figure S1C**).

As such, our analysis showed that polyadenylation signals leading to the production of (at least) two alternative transcripts of distinct length are overall conserved throughout the group. however, the exact position of the polyadenylation signals within each mRNA transcript shows some variation from species to species indicating certain level of plasticity in the mechanism of alternative polyadenylation (**Figure 4A**).

Since VISTA-LAGAN's output is a set of multiple-alignments generated in relation to homology with the *Drosophila melanogaster's* 3'UTR – the baseline - and as such does not graphically transmit the changing size of sequence lengths, we investigated the absolute positioning of polyadenylation signals, as well as the predicted size of the transcripts, to understand if the Long/Short 3'UTR isoform ratio is predicted to change within the Drosophilids. This would be of importance because the current model for *Hox* 3'UTR remodelling is that this regulatory event changes the quality and quantity of *cis*-regulatory information carried by a *Hox* transcript. A significant change in this ratio would mean that the *cis*-information requirements for a successful developmental regulation of different 3'UTR isoforms might be different across species.

We found that the Long/Short 3'UTR isoform ratio changes within the Drosophilids, from a 1:1 relationship in *Drosophila grimshawii*, *Drosophila mojavensis* and *Drosophila virilis* to the 3:2 proportion observed in *Drosophila melanogaster* and its closely-related species (**Figure 4B**).

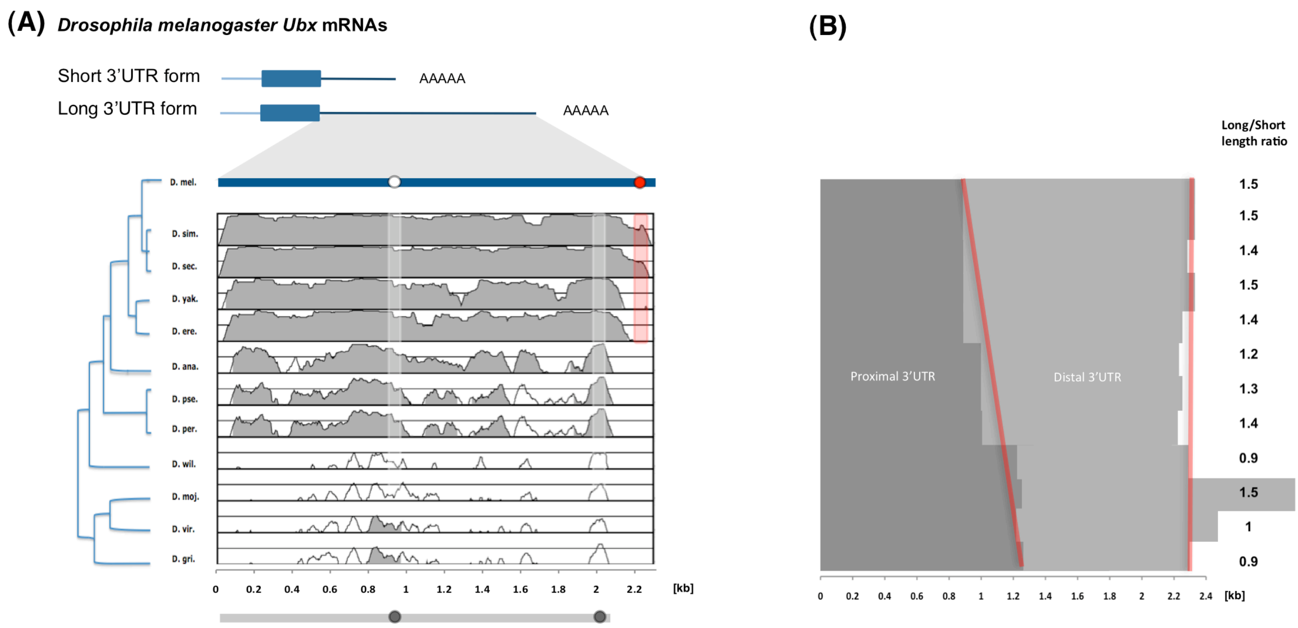


Figure 4. Alternative polyadenilation is conserved within the Drosophilids (A) The *Ubx* gene in *Drosophila melanogaster* produces two alternatively poly-adenylated mRNA forms: *Ubx* short 3'UTR and *Ubx* long 3'UTR (see top diagram). Multiple-alignments for Drosophilid *Ubx* 3'UTR primary sequences using the VISTA-LAGAN software. *Ubx Drosophila melanogaster* 3'UTR sequences (represented by a blue bar – see top) is used as a baseline sequence – see top rectangle. *Drosophila melanogaster* poly-adenilation signals (PAS) are shown: PAS1 in white and PAS2 in red; a putative additional PAS is shown in black with an ultraconserved canonical sequence (AATAAA). Sequence homology is represented on the vertical axis of each aligned sequence, with a minimal value of 50% and a maximum value of 100% (gray regions correspond to segments with 70% or more of sequence similarity). **(B)** The ratio of distal/proximal 3'UTR length is generally greater in species more closely related to *Drosophila melanogaster*, with the first polyadenylation signal site receding approximately 350bp across this time-window.

There is an intermediate group, composed by *Drosophila pseudoobscura, Drosophila persimilis* and *Drosophila ananassae* that shows a Long/Short isoform ratio of 1.2 to 1.4. This points to a the evolution of the 3'UTR isoform ratio within the Drosophila genus.

3.2) RNA accessibility in mRNA 3'UTRs is ultraconserved despite significant change at the primary-sequence level.

We furthered our analysis of the constraints on and nature of 3'UTR sequence evolutionary variation by hypothesising that Drosophilid 3'UTRs might be subject to selective pressures at the secondary structure level. This would be expected, since transcribed sequences necessarily have to encounter two and three-dimensional realities during cellular life, events which have been shown to affect the regulation of mRNA species (see Introduction).

We asked whether these requirements were stringent or relaxed, as compared to the primary-sequence evolutionary profiles, which overall show islands of conservation among very variable regions. Strikingly, we found that each *Hox* has a distinct and generally ultraconserved profile of accessibility across the UTR sequence, despite extensive change at the primary-sequence level (**Figure 5**).

We found that the negative control exhibits, unlike *Hox* 3'UTRs, significant variation across species in accessibility values, within each 200 bp window, despite being of the same size and having a similar primary-sequence conservation profile as the *Hox* 3'UTRs analised (**Figure 5A**).

Also, the negative control presented regions with variance values as low as the ones on *Hox* 3'UTRs; when inspected closely these lowly variant regions corresponded to highly conserved regions at the primary sequence level. Thus, accessibility values are, in the negative control, highly dependent on DNA primary-sequence (**Figure 5B,C**). This is not observed in transcribed regions. For example, the highly conserved accessibility profile 1200-1600 bp into the *ultrabithorax* 3'UTR (see Figure 1A) corresponds to a great degree of erosion at the primary sequence level.

It is also interesting to note that *abd-a* 3'UTRs show highly assymetric variation in accessibility values: while the end of the 3'UTR has very low variance, as in *Ubx*, the beginning of this sequence is the only observed case where the values are almost as variant as the negative control. This points to an interesting local constraint in secondary structure, directed towards the end of the 3'UTR, unlike other *Hox* which appear to have a somewhat low and homogeneous variance in accessibility across the 3'UTRs.

Thus we find that *Hox* 3'UTR regions are predicted to have ultraconserved accessibility values that do not follow from the primary-sequence conservation profiles, pointing to a strong and previously unprobed constraint on the RNA secondary structure level of *Hox* gene expression, and indirectly, on the correspondent genomic region.

3.3) miRNA regulation shows distinct and dynamic evolutionary profiles across Drosophilids.

We previously explored how post-transcriptional *Hox* regulatory inputs might be limited by transcript size and secondary structure constraints. We now asked how individual modules within the 3'UTRs evolved within the 60 million-year evolutionary window available since the sequencing of 12 Drosophilid species.

Given the growing knowledge and bioinformatic tools regarding miRNA regulation, we chose to focus on the miRNA target-site modules. For this, we concentrated on *ubx*, a *Hox* gene which has been shown to be targeted post-transcriptionally by miRNAs of the iab-4/iab-8 complex (see Introduction), and studied how miRNA target-site evolution occurred within Drosophilid *ubx* 3'UTRs.

To have an unbiased approach to the evolution of *ubx* miRNA regulation, we used the PITA algorithm. These analyses generated an extended list of miRNA targets for each Drosophilid species. We refined this list by using only the target hits for miRNAs that had ultraconserved seeds across all species, were co-expressed with *Ubx* and had a high $\Delta\Delta G$ value (**Figure 6**) (See Methods).

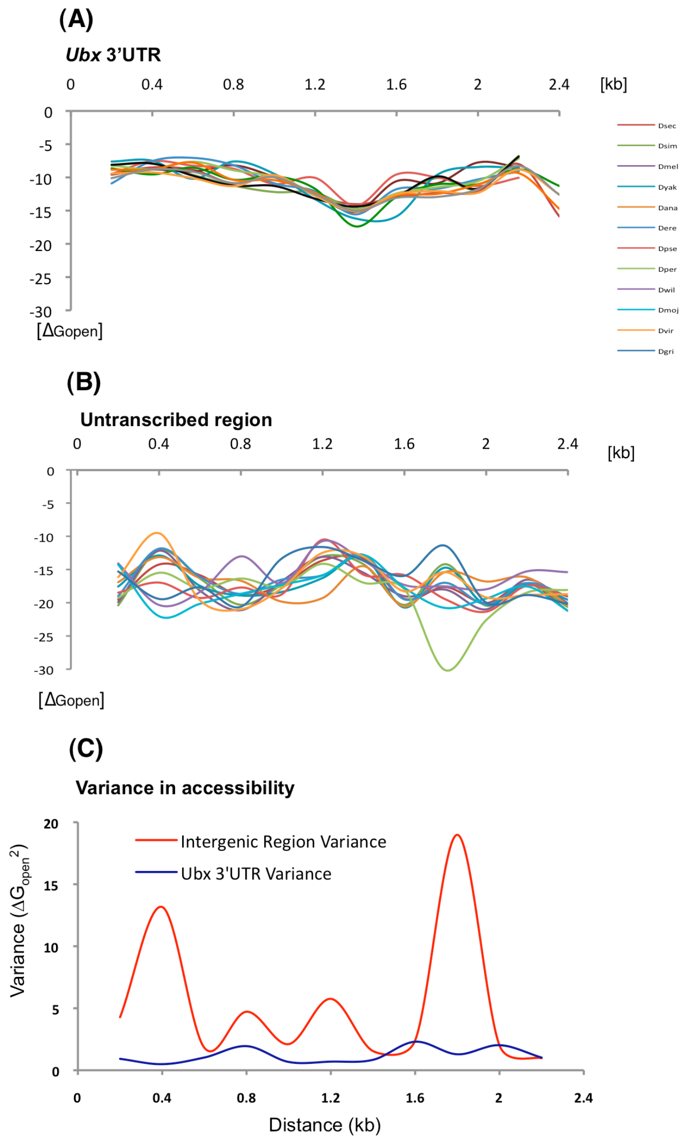


Figure 5. RNA accessibility is conserved across *Hox* 3'UTRs, unlike primary sequence. (A) RNA accessibility alignments for *Ubx* 3'UTRs. A measure of RNA accessibility (ΔGopen) is plotted vs. *Ubx* 3'UTR length. Low ΔGopen values indicate low accessibility. Despite significant divergence at the level of primary sequence, the accessibility of homologous regions of the *Ubx* 3'UTR remains generally constant. **(B)** ΔGopen values for a control sequence extracted from an untranscribed intergenic region in *D. mel*. (Ch3R:12604500-12607000); note the high level of variation in ΔGopen values observed in this case. **(C)** Variance analysis of *Ubx* and the control intergenic segment**;** while *Ubx* variance in accessibility values remains fairly unchanged across the *Ubx* 3'UTR, variance for the control segment shows distinct peaks revealing lack of conservation in secondary structure predictions.

After the described refining steps, we obtained a list of 14 *Drosophila melanogaster* high-ranking miRNAs that are temporally co-expressed with *Drosophila melanogaster ubx* mRNAs.

We then examined the evolutionary patterns of changes in ΔΔ*G* values by plotting the values obtained in each Drosophilid species, for the 14 individual miRNAs, against the Drosophilid phylogenetic tree.

To evaluate quantitative miRNA-targeting evolution, we used two threshold ΔΔG values. The iab-4/iab-8 miRNA predicted to have the lowest average regulatory affinity to *Ubx* 3'UTRs (miR-iab4-5p) was used as the lower threshold. It presented ΔΔ*G* values between 0 and -4 across the phylogeny. On the other hand, the miR-iab-4-3p ΔΔG value in *Drosophila melanogaster* was -8; this was used as the higher threshold.

Thus, we judged other miRNA targeting interactions: those miRNAs with *Ubx* 3'UTR affinity values crossing both threshold values across the phylogeny (0>ΔΔ*G*<-8) were deemed as evolving quantitatively. miRNA regulatory affinitties that did not clearly cross both thresholds were assumed to be conserved (*stasis*).

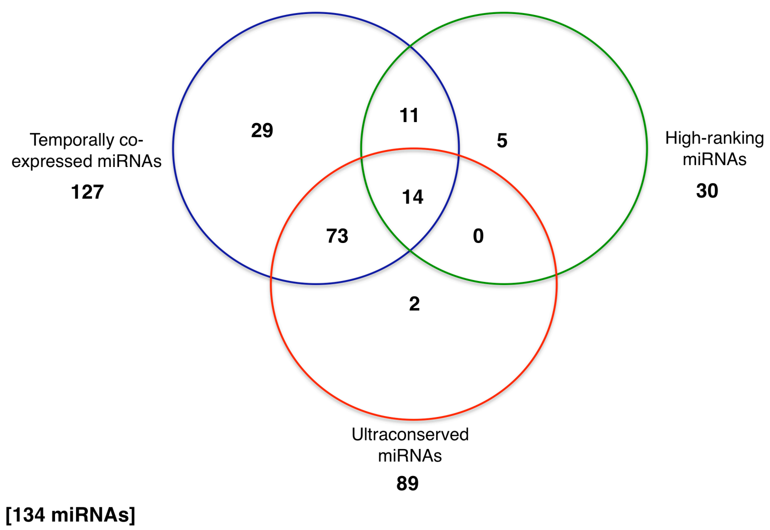


Figure 6. Filtering of PITA miRNA-targeting predictions for further analyses. The PITA *Drosophila melanogaster Ubx* 3'UTR analysis yielded a total of 134 possible miRNA regulators with different affinities. We chose those that were high-ranking (ΔΔG values equal or above those of the lowest-ranking miRNA of the iab-4/iab-8 complex). From these 30 miRNAs, we further selected those that had both a conserved seed sequence throughout this group and were previously shown to be co-expressed with *Ubx*. This yielded a final number of 14 miRNAs for further analyses.

This analysis revealed two distinct miRNA-target evolutionary patterns (Figure 6).

1) 8 of the 14 miRNAs exhibited an evolutionary pattern towards more negative values in D*rosophila melanogaster* (**Figure 7A-H**), and thus gaining in potential regulatory weight (refered henceforth as *positive trend*).

2) the remaining 6 miRNAs showed a *stasis* trend, slightly variating around a given $\Delta\Delta G$ value across the 12 species (**Figure 7I-N**).

3.4) Individual miRNA-target dynamics and alternative polyadenylation.

We next analysed the evolutionary dynamics of individual miRNA target-sites for these 14 miRNAs, to see how their individual evolution might have translated into the previously observed heterogeneity of predicted net regulatory dynamics.

For this we used the PITA outputs for individual target-sites of each of the 14 miRNAs. This software also ascribes a $\Delta\Delta G$ value for each individual site in addition to the net-regulatory values used in the previous section, while also informing on the miRNA-target position on the 3'UTR for each species.

To understand if the predicted miRNA target-sites were homologous, we compared the positions of each individual miRNA site of the PITA output for the 14 miRNAs across all Drosophilids using for this the VISTA-LAGAN alignments generated previously, and found a total of 317 putative sites. We catalogued these target-sites based on their predicted individual regulatory strength, calling them *core* if they had a predicted $\Delta\Delta G$ value equal or below that of miR-iab-4/miR-iab-8 predominant sites ($\Delta\Delta G$=-8), *mild-shadow* sites if they had an intermediate value ($-8<\Delta\Delta G<-4$) and *weak-shadow* sites if the $\Delta\Delta G>-2$.

We then plotted all target-sites for a given miRNA against the Drosophilid phylogenetic tree, using again the *Drosophila melanogaster ubx* 3'UTR as the baseline (see **Supplementary Figure 3**).

1) We found that in all cases, with the exception of miR-210 which has two cooperative target-sites, each miRNA is predicted to have a predominant *core* target-site on the *ubx* 3'UTR, contributing largely to the net $\Delta\Delta G$ value, followed by latent or *shadow* target-sites that contribute to this value less strongly (see **Supplementary Figure 3**).

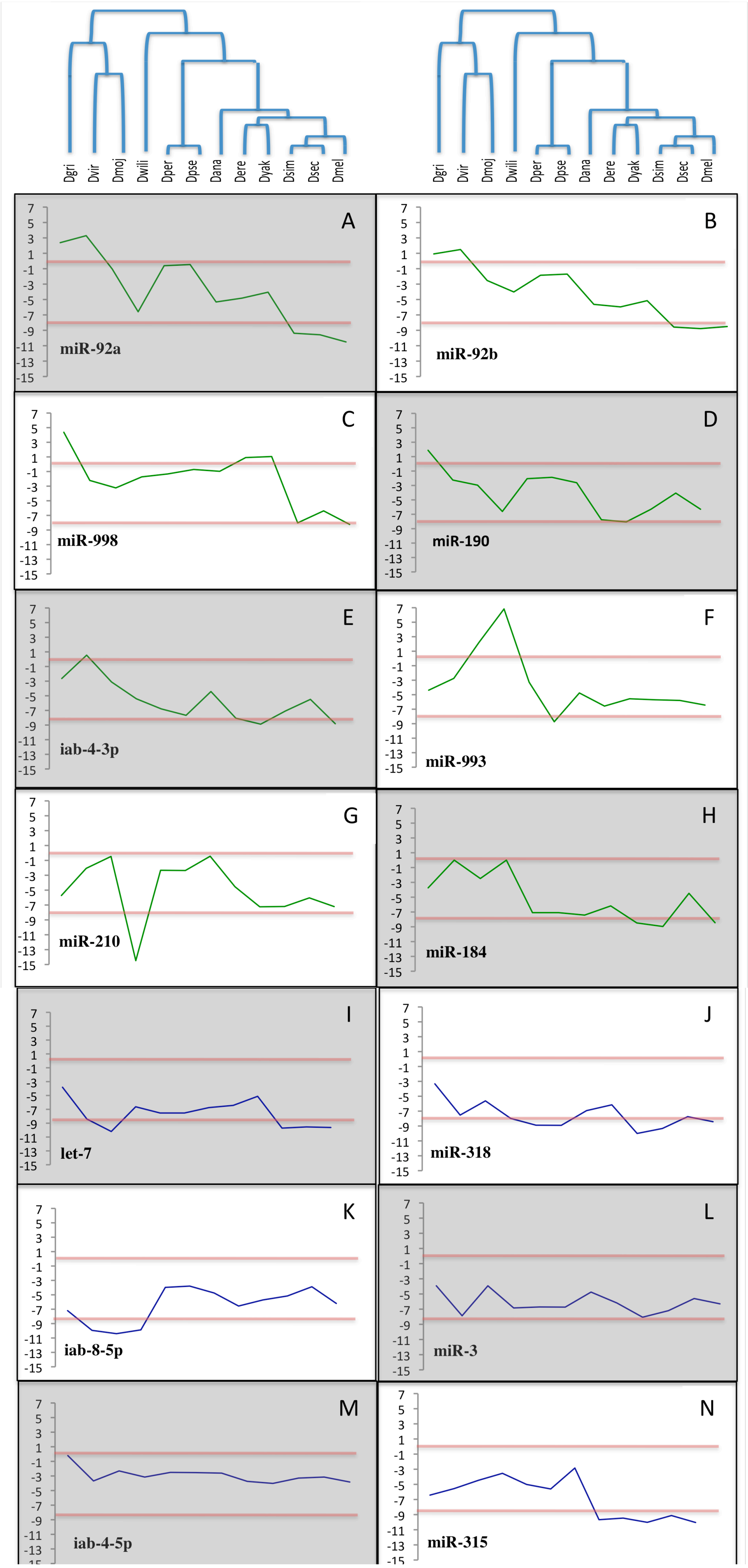

Dgri
Dvir
Dmoj
Dwili
Dper
Dpse
Dana
Dere
Dyak
Dsim
Dsec
Dmel
A
miR-92a
B
miR-92b
C
miR-998
D
miR-190
E
iab-4-3p
F
miR-993
G
miR-210
H
miR-184
I
let-7
J
miR-318
K
iab-8-5p
L
miR-3
M
iab-4-5p
N
miR-315

Figure 7 (Continued from previous page). Quantitative evolution of miRNA regulation of *Ubx*. The figure shows the regulatory evolution of the 14 miRNAs selected for further study. The Y axis represents strength of regulatory interactions in ΔΔG – the more negative the value, the stronger the interaction is. The two thresholds used to judge directional evolution are represented in red in each of the 14 graphs. miRNA regulatory trends were judged as (**A-H**) **directional if they cross both thresholds**, increasing in predicted regulatory effects across species. If they do not cross both thresholds, this indicates (**I-N**) a ***stasis*** trend, showing no significant change in predicted regulatory effects. Notice that miRNAs produced from the iab-4/iab-8 locus – which have been experimentally shown to target *Ubx* mRNAs – have distinct evolutionary trends regarding the targeting of *Ubx 3'UTR:long* mRNAs.

Table 1 – Analysis of individual miRNA target-sites

	***Core* sites**			***Shadow* sites**					
				Mild			**Weak**		
miRNAs	**Total**	Conserved	Lost	**Total**	Conserved	Lost	**Total**	Conserved	Lost
miR-92a	**3**	3	0	**3**	0	3	**3**	0	3
miR-92b	**3**	3	0	**3**	2	1	**16**	13	3
miR-iab-4-3p	**4**	3	1	**4**	2	2	**3**	0	3
miR-190	**0**			**7**	5	2	**12**	3	9
miR-998	**2**	2	0	**1**	1	0	**6**	6	0
miR-993	**1**	1	0	**5**	5	0	**4**	3	1
miR-210	**2**	0	2	**7**	6	1	**12**	4	8
miR-184	**2**	2	0	**11**	11	0	**9**	8	1
let-7	**5**	5	0	**11**	9	2	**17**	9	8
miR-iab-4-5p	**0**			**0**			**17**	8	9
miR-iab-8-5p	**4**	4	0	**5**	3	2	**54**	31	23
miR-315	**5**	5	0	**7**	2	5	**24**	3	21
miR-3	**0**			**11**	11	0	**9**	8	1
miR-318	**7**	7	0	**6**	4	2	**5**	4	1
Total	**38**	**35**	**3**	**81**	**61**	**20**	**191**	**100**	**91**

Legend: Table 1 shows the categorization of the 310 indentified individual target-sites that underlie the net regulatory predictions presented in section 2.2) for each miRNA. We categorized each individual site according to **strength** (*Core* vs *Shadow*) and **conservation**.

2) miRNA target-sites tended to have a polyadenylation isoform identity, in each species and throughout evolution, showing a tendency to remain within either the long or the short form of the alternatively polyadenilated 3'UTR (**Figure 8**).

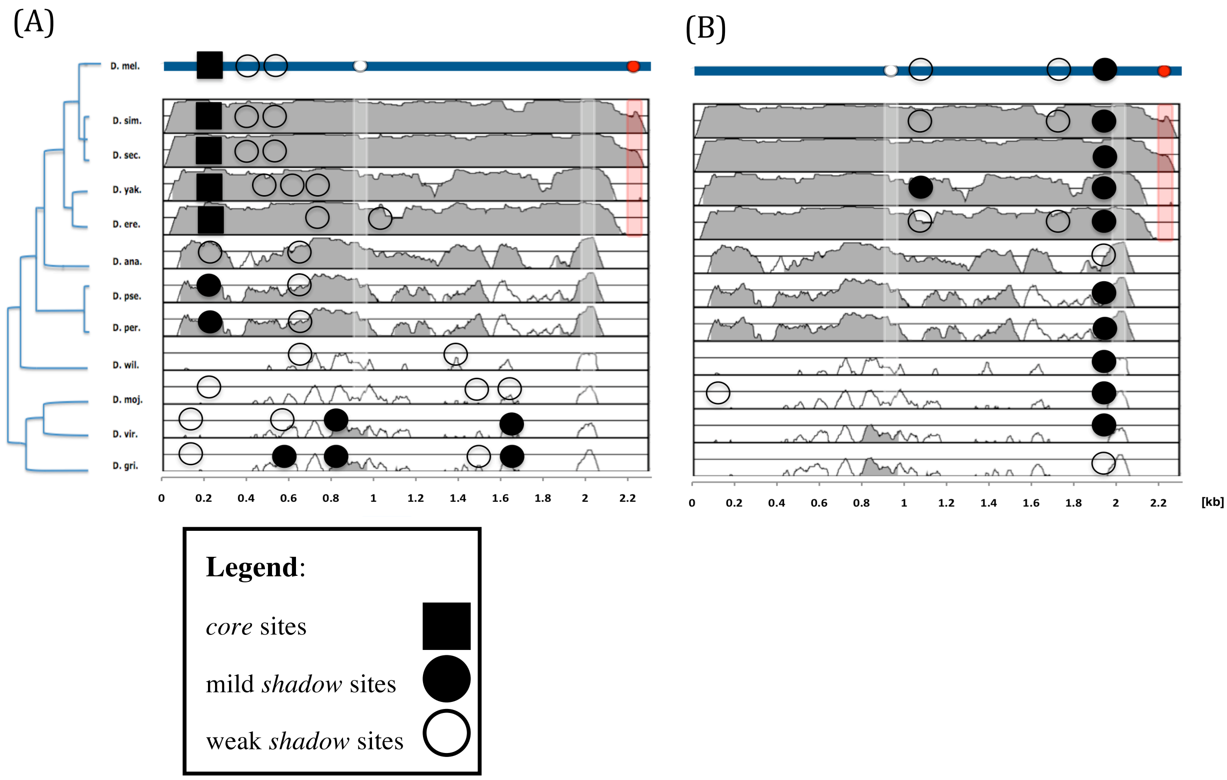


Figure 8. Isoform-identity of miRNA targets throughout evolution: two examples. Here, two examples are shown to highlight the statistically-supported result that miRNA targets tend, for a given miRNA, to occur within one of the *Ubx* 3'UTR isoforms, despite significant evolutionary change in their precise positioning **(A)** *miRNA-315* target-sites show a tendency to occur within the proximal *Ubx* 3'UTR. **(B)** *miRNA-3* target-sites show a tendency to occur within the distal *Ubx* 3'UTR. For the target-site evolution of other miRNAs please see **Supplementary Figure S5**.

76.7% of the miRNA target-sites, either newly-formed or conserved, remained within the UTR isoform in which the predominant target-site for that specific miRNA lies in *Drosophila melanogaster*. The performed χ^2 test supported this idea by rejecting the null hypothesis for α=0.01 (χ^2=7.679>6.635).

3) We considered a given miRNA target-site as conserved if it was found in homologous positions in 3 or more of the 12 species with any strength. Based on this, we found that *core* sites tend to be less evolutionary volatile. Conversely, *shadow* sites show an energic dynamic of emergence and erosion. For instance, out of the 38 strong miRNA target-sites found in all species for all of the 14 miRNAs, only in three cases did the site disappear completely from the UTR (92.1% of conservation) (**Table 1**). The χ^2 test supported this for α=0.001 (χ^2=15.535>10.827).

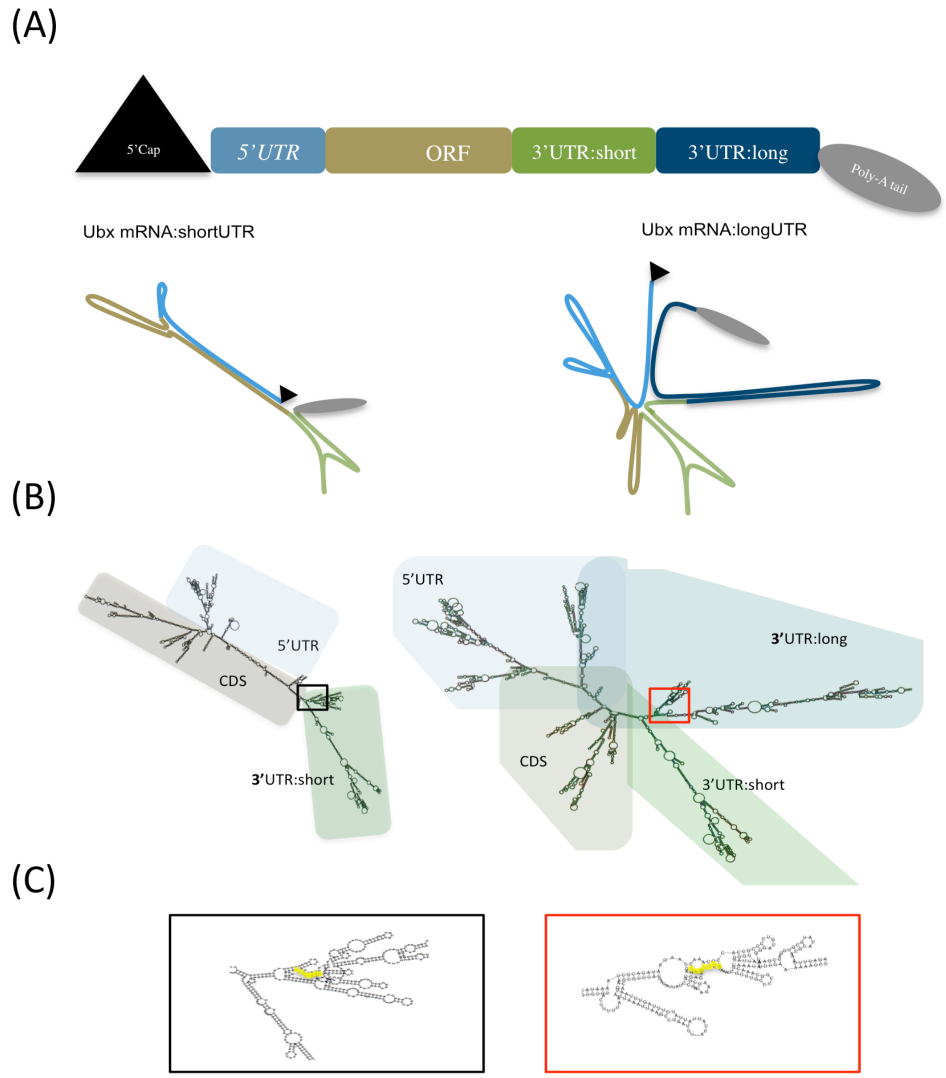
On the other hand, focusing on *shadow* sites, out of the 81 *mild* sites found in all species, 61 were conserved (75.3%), while in the case of *weak* sites only 52% of the identified target-sites were conserved in three or more species (100 out of the pool of 191).

3.5) How do alternative polyadenilation events, mRNA secondary structure and miRNA-targeting coexist in Hox post-transcriptional regulation: a model?

Next we asked whether alternative polyadenilation, given the radical remodelling the 3'end of a transcript, could be responsible for a significant change in mRNA secondary structure, and as such, in the change in importance of relevant regulatory modules sitting on the 3'UTR of *Hox* genes, such as miRNA target-sites, thus changing the post-transcriptional regulatory landscape of the mRNA in a non-additive manner.

The structure prediction for the mRNA transcript with the longest *ubx* 3'UTR (*ubx:*long) was first analysed. We superimposed gene sequence anatomy information on the 2D structure prediction and found that both the proximal and distal tracts of the 3'UTR (before and after PAS1) folded mostly individually (in a modular manner) albeit with a minor area of mutual overlapping around *ubx* PAS1. This was not observed in both the 5'UTR and the coding sequence (CDS), which mainly form double-stranded structures with each other. The observation of an small area of RNA-RNA interaction between the proximal and distal tracts of the *ubx*:long 3'UTR prompted us to look at the minnimum-free energy 2D prediction for the *ubx*:short 3'UTR isoform.

We found that the aforementioned small area of superimposition between the ribonucleotides of the two 3'UTR isoforms had different predicted shapes and accessibility values.



microRNA	*Position*	ΔΔG
miR-313	**919**	-12.01
miR-92a	919	-10.51
miR-92b	919	-8.51
miR-312	**919**	-6.41

microRNA	*Position*	ΔΔG
miR-313	**919**	**-10.37**
miR-92a	919	-8.87
miR-92b	919	-6.87
miR-312	**919**	**-4.77**

Figure 9. Integrated model of *Ubx* post-transcriptional regulation in Drosophila. Here, we show the RNAFold secondary structure predictions for the *Ubx* mRNA transcripts, and how these change as a consequence of alternative polyadenylation. **(A)** mRNA model with a colour-code correspondence to the a graphical representation results for the *Ubx* mRNA:short3'UTR and the *Ubx* mRNA:long3'UTR that should guide the following sections of the figure. (**B**) The secondary structure predictions for *Ubx* mRNA:short3'UTR and the *Ubx* mRNA:long3'UTR. Notice the modular property of the 3'UTR isoforms in terms of folding, an how, on the contrary, other regions of the transcript (5'UTR, coding-sequence) mostly fold with one another. The squares represent the same region (black in the short 3'UTR, red in the long 3'UTR). In the case of the *Ubx* mRNA bearing the long 3'UTR isoform, this rgion shown a partial superimposition between the proximal and the distal 3'UTR tracts.

(C) When analysed carefully, the region of partial complementary between the proximal and the distal 3'UTR tracts show differences in secondary structure depending on the context. The secondary structure of this region is more complex (i.e. more loops and "bubbles") when the transcript has only the proximal 3'UTR tract. This enhances the accessibility of the region. When miRNA target-site predictions were performed for this region, lying about 30 nucleotides upstream of the first polyadenylation signal, we found that target-sites within this region decrease their affinity to miRNAs when in the context of the *Ubx*:long3'UTR mRNA transcript, even though they sit in the proximal tract. This area of secondary structure "instability" was also confirmed for the other alternatively-polyadenylated *Hox* genes (data not shown).

Since 3'UTR accessibility is important for miRNA regulation, we performed miRNA target-site predictions using PITA for the two *ubx* 3'UTR isoforms in order to understant if the $\Delta\Delta G$ values for this region were changed by alternative polyadenylation. We found that the miRNA target sites within the region starting approximately 65 bps upstream of the PAS1 are predicted to change their regulatory strength, sometimes significantly. More specifically, the $\Delta\Delta G$ miRNA target-site values for miR-92a, miR-92b, miR-312 and miR-313 (in position 919) changed significantly. As an example, miR-92a and miR-92b target-sites decreasing in strength by a value of approximately $\Delta\Delta G$=2 (from -10.51 to -8.77 and -8.51 to -6.77, respectively). As such, the addition of approximately 1300 bp to the *ubx* mRNA by alternative polyadenylation is not only predicted to add new regulatory modules to the transcript but also to change the modules of the constitutive tract of the 3'UTR, by remodelling the secondary structure of the region around the first polyadenylation site.

We proceeded to analyse the other *Drosophila melanogaster Hox* genes that are known to undergo alternative polyadenylation to understand how general is the observed RNA structural instability around the *ubx* first polyadenylation signal site. Iterating the procedure described above for *antp*, *abd-a* and *abd-b*, we found that all these *Hox* genes present the same general pattern. *abd-a* showed a region of structural instability with a similar size, while in *abd-b* and *antp* this region extended to 75 bp and 80 bp upstream of the PAS1, respectively. The strength of miRNA target-sites within these regions was also changed, sometimes significantly as with *Ubx* (Data not shown).

4. DISCUSSION AND FUTURE DIRECTIONS

In this work, we expand the current knowledge of DNA sequence evolution by addressing the variation of Drosophila Hox 3'UTRs, sequences in *cis* that are known to play an important role in the post-transcriptional regulation of the transcripts.

We first show that the alternative polyadenylation patterns found in *Drosophila melanogaster Hox* developmental expression patterns (see Introduction) are conserved throughout the Drosophilids, suggesting that *Hox* alternative polyadenylation is likely to be a feature present in the common ancestor of the group. Interestingly, we also find that the total transcript length was approximately maintained within the group, while the ratio of 3'UTR isoform length has undergone significat change: in those species where the proximal 3'UTR is shorter, the distal 3'UTR is extended. We are now performing extractions of embryonic RNA, followed by RT-PCR in the species *Drosophila melanogaster, Drosophila simulans, Drosophila ananassae, Drosophila pseudoobscura* and *Drosophila virilis* to validate the predictions on alternative polyadenylation signal usage and thus also on 3'UTR isoform ratio evolution. (*Drosophila simulans* populations were kindly provided by the Sucena Lab; other species were kindly provided by John Roote at the Department of Genetics, Cambridge University)

Also, upon probing the *Ubx* 3'UTRs of the Drosophilid *Ceratitis capitata* and insects *Aedes aegypti*, *Anopheles gambiae*, *Bombyx mori*, *Tribolium castaneum*, *Apis mellifera* and *Nasonia vitripenis* (Flybase BLAST followed by VISTA-LAGAN alignments; data not shown), we found that even though the exact positioning of the Drosophila polyadenylation signals seems unconserved across these species, there are always two strong alternative polyadenylation signals present. These results support the idea that 3'UTRs, as well as the ability to generate alternative versions of these for a given transcript, are functionally very important for gene expression patterns.

The vertebrate paralogous *Hox* genes HOXa7 and HOXb7 are *Ubx* orthologs (PEARSON ET AL., 2005) show a marked distinction in 3'UTR size and sequence composition (*NCBI:nucleotide search*; data not shown). This is consistent with the hypothesis that there is a selective pressure to maintain differential post-transcriptional information in *Ubx*, since it is expected that after the generation of redundancy by gene duplication, either loss of one of the paralogs occurs or division of labour between the two is achieved, through complementary degeneration.

As such, alternative polyadenylation appears important as a system that generates different transcripts across the constantly changing molecular context that defines development, as it can provide temporal resolution to the control of gene expression, by allowing the same gene to be differently recognized as its expression time progresses – the onset of embryonic *Hox* expression and its end can thus be differentiated by *Hox trans*-regulators.

However, the described change in the precise positioning of the polyadenylation signals that underlie alternative polyadenylation, as well as the 3'UTR isoform length ratio changes, point to some degree of plasticity within this mechanism. The *readout* of this system appears thus relatively free to change within the context of alternatively polyadelylated 3'UTRs.

We have shown that many co-expressed miRNAs predicted to target the *Ubx* 3'UTR in *Drosophila melanogaster* appear to have changed significantly, sometimes from no affinity to a valued predicted to signifficantly affect translation (KERTESZ ET AL., 2007). Others, like let-7, a miRNA that was shown to control the developmental transition from late larval to adult stages (NIWA & SLACK, 2007), seem to maintain a strong targeting value across evolutionary time. Only miRNAs from the iab-4/iab-8 bidirectional locus were shown to target *Ubx* (see Introduction). As such, this study provides candidate miRNA genes for the post-transcriptional control of *Ubx*, as well as the evolution of this regulatory level. Our results strengthen the hypothesis that miRNA target-site evolution can be quantitative.

It is interesting to note that the three iab-4/iab-8 complex miRNAs identified as top-ranking by our analysis show a very distinctive *Ubx* target-site evolution. Although these miRNAs were shown to downregulate *Ubx* (see Introduction), there is no information about *which* miRNA of each of the two forms, the 5p or the star miRNAs, is actually responsible for this effect. This study provides a novel hypothesis in this respect. While iab-4-5p and miR-iab-8-5p apparently maintain their *Ubx* 3'UTR targeting value relatively constant across evolutionary time, affinity to other miRNA from the same locus, miR-iab-4-3p, has apparently undergone a signifficant quantiutative change. Given that this loci are conserved across insects and possibly arthropods (RONSHAUGEN ET AL. 2008), it seems thus plausible that the post-transcriptional control of *Hox* genes by this locus might be a novelty within the

Diptera. The strong effect on *Ubx* observed in *Drosophila melanogaster* is thus hypothesized to occur by miR-iab-4-3p targeting, and to have undergone significant change during Drosophila evolution.

One could argue that the quantitative miRNA-targeting evolution data is in accordance with the MD (Muller-Dobzhansky) model of hybrid incompatibillity, wherein after the divergence of two populations from a common ancestor, two sets of interacting genes are expected to coevolve idependently in each population, and are thus expected to be incompatible when a hybrid occurs. In fact, a recent study of the miR-310 cluster shows that this does occur in the case of miRNA-target-gene co-evolution, in this case between *Drosophila melanogaster* and *Drosophila virilis*. The miR-310 cluster miRNAs show seed-sequence divergence and cause a misexpression of the target-genes (predicted to be the same across species) when expressed in an heterospecific manner (TANG ET AL., 2010). However, we selected miRNAs with ultraconserved seeds, supporting the idea that our results represent a true quantitative change in miRNA targeting.

We are currently developming an algorithm, in colaboration with Mafalda Dias (Theoretical Physics Department, University of Sussex), to simulate the evolution of the 12 *Ubx* 3'UTR sequences of Drosophila in neutral circumstances. Even though we observe ultraconservation of modules that extend up to 200 base-pairs in length (Results, **Figure 4**), we will compare the negative control with our data to access how these results can be explaine by natural selection.

We also studied individual miRNA target-site dynamics, and found that targets for a given miRNA tend to occur in one of the 3'UTR isoforms, despite significant evolutionary change in their precise positioning. This is again in accordance with the model espoused above, wherein the segregation of post-transcriptional regulatory information amongst disting 3'UTR isoforms is an important mechanism for the control of gene expression patterns in ontogeny and evolution.

Additionally, most ot the targeting values for a given miRNA can be explained by one of the many target-sites present in the 3'UTR for that miRNA, the others being accessory or *shadow* target-sites, that contribute marginally to the miRNA visibility of the 3'UTR. This mirrors the discoveries in recent studies on the eukaryotic transcriptional control by redundant enhancers for a given gene. For instance, Dorsal, a TF that is involved in dorsal-ventral patterning of the early *Drosophila melanogaster* embryo was shown to have two enhancers that activate its transcription in the same tissues (HONG ET AL., 2008). The secondary or *shadow* enchancers were shown to suffer rapid evolutionary turnover within the 12 Drosophilids, while the primary of *core* enhancers seem more throughly conserved. This is in accordance with our results for miRNA target-sites, in which *core* sites appear more conserved that *shadow* sites.

Also, recent study of the transcriptional control of *svb* (see Introduction) has shown that *shadow* enhancers can function as a robustness mechanism. The mutant *shadow* enhancers show no phenotypic effect unless the organism is exposed to environmental stress caused by high temperatures (FRANKEL ET AL., 2010). This points to a role of *shadow* enhancer sequences in the achievent of developmental robustness, a principle that is in accordance with our data and could thus be extended to the treatment of post-transcriptional *cis*-regulation and its evolution.

RNA secondary structure is usually disregarded in developmental evolutionary studies. This is justified by the fact that most of these studies deal with transcription, a regulatory event controlled at the level of the DNA sequence. We aligned the accessibility values (See Methods and Results) of four *Hox* 3'UTRs (*Ubx*, *antp*, *abd-a* and *abd-b*) across the 12 Drosophilid genomes available, and found that the conservation in secondary structure is more conspicuous than its primary sequence counterpart. This points to an previously overlooked constraint in RNA evolution, and can explain, in conjuction with primary-sequence module analysis, the evolutionary constraints acting on 3'UTRs. This secondary structure constraint might be specific to 3'UTRs or can otherwise be a property of the trascriptionally active genome. We are planning a bioinformatic study that addresses this question, with the aim of further understanding the particularities of secondary structure in post-transcriptional control and its evolution.

Since post-transcriptional regulatory information is mainly deposited in the 3'UTR, we tried to formulate an integrated model for *Hox* 3'UTR regulation in Drosophila that included secondary structure, miRNA regulation and alternative polyadenylation. We studied the RNA folding of the whole *Ubx* mRNA, and found that 3'UTRs tend to fold in a modular way (i.e. base-pairing occurring essentially within each of the isoforms), unlike the rest of the transcript. Also, the minimal region of overlapping between proximal and distal isoforms is precidted to change the miRNA predictions for the proximal 3'UTR tract.

These results point to a previously unknown post-transcriptional mechanism, wherein the addition of a nucleotide stretch to the 3'-terminal untranslated region of a transcript changes the structure of the constitutive 3'UTR, thus making it possible for alternative polyadenylation to remodel the regulatory landscape of the mRNA molecule in a non-linear manner, instead of acting as a simple addition of novel regulatory modules to *Hox* mRNAs. Also within this framework, the spatial segregation of the 3'UTR could be important as mRNA *trans*-regulators would be able to readily recognize and regulate a transcript.

Note: *The present study has yielded a research paper, currently undergoing the second iteration of the reviewing process in the journal* Molecular Biology and Evolution (**Patraquim & Alonso**, *Molecular Biology and Evolution (*2010) – *in revision (see annex S4).*

5. BIBLIOGRAPHY

ALONSO, C.R. & WILKINS, A.S., 2005. The molecular elements that underlie developmental evolution. *Nature Reviews Genetics* **6**, 709–715.

ALONSO, C.R., 2008. The molecular biology underlying developmental evolution *in Evolving Pathways: Key Themes in Evolutionary Developmental Biology*. Cambridge University Press. Cambridge.

AVEROF M. & PATEL N.H., 1997. Crustacean appendage evolution associated with changes in Hox gene expression. *Nature*, **388**, 682-686.

BARTEL, D.P., 2009. MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, **136**(2), 215-233.

DE BEER, G., 1971. *Embryos* and *Ancestors*, *3rd ed.*. Clarendon Press. Oxford.

BENDER, W., 2008. MicroRNAs in the Drosophila bithorax complex. *Genes & Development*. **22**(1), 14-9.

BRUDNO, M., DO, C.B., COOPER, G.M. , KIM, M.F. , DAVYDOV, E., N.C.S. PROGRAM, GREEN, E.D., SIDOW, A. & BATZOGLOU, S., 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research* **13**:721-731.

CHEN, C.-Y., CHEN, S.-T., JUAN, H.-F., HUANG, H.-C., 2010. Lengthening of 3'UTR Increases Morphological Complexity in Animal Evolution. *Nature Preceedings*. Sept.

GARSTANG, W., 1922. The theory of recapitulation: A critical restatement of the Biogenetic Law. *Proc. Linn. Soc. Lond. Zool,* 3581-*101*.

GILBERT, S. F., 2010. *Developmental Biology*, *9th Edition*. Sinauer Associates. Sunderland, Massachusetts.

GOULD, S. J., 1977. *Ontogeny and Phylogeny*. Harvard University Press. Cambridge, MA.

HOFACKER, I.L. 2003. Vienna RNA secondary structure server. *Nucleic Acids Research*, **31**, 3429-3431.

HONG, J.-W., HENDRIX, D.A., LEVINE, M.S., 2008. Shadow Enhancers as a Source of Evolutionary Novelty. *Science*, **321**(5894), 1314

HUGHES, C.L. & KAUFMAN, T.C., 2002. Hox genes and the evolution of the arthropod body plan. *Evolution & Development*, **4**, 459–499.

FRANKEL, N., DAVIS, G.K., VARGAS, D., WANG, S., PAURE, F. & STERN, D.L., 2010. Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, **466**, 490–493.

JACOB, F. & MONOD, J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*. **3**, 318–356.

KEDDE, M., VAN KOUWENHOVE, M., ZWART, W., VRIELINK, A.F.O., ELKON, R. & AGAMI, R., 2010. A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nature Cell Biology*, **12**(10), 1014-1020.

KERTESZ et al. The role of site accessibility in microRNA target recognition. Nat Gen. 2007

KERTESZ, M., IOVINO, N., UNNERSTALL, U., GAUL, U. & SEGAL, E.. 2007. The role of site accessibility in microRNA target recognition. *Nature Genetics*. **39**:1278-1284.

KING, M-C. & WILSON, A. C., 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116.

KOSIK, K., 2009. MicroRNAs tell an evo–devo story. *Nature Reviews Neuroscience*, 10, 754-759.

LEWIS, E. B., 1978. A gene complex controlling segmentation in Drosophila. *Nature* **276**, 565-570.

LICATALOSI, D.D., DARNELL, R.B., 2010. RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics*, **11**, 75-87.

LU, J., FU, Y., ZENG, K., XU, A., CARTHEW, R., WU, C., 2008a. Adaptive Evolution of Newly Emerged Micro-RNA Genes in Drosophila. *Molecular Biology & Evolution*. **25**(5), 929-938.

LU, J., SHEN, Y., WU, Q., KUMAR, S., HE, B., SHI, S., CARTHEW, R., WANG, S.M., WU, C.I, 2008b. The birth and death of microRNA genes in Drosophila. *Nature Genetics*, **40**(3), 351-5.

MACDOUGALL, A., CLARK, E., MACDOUGALL, E. & DAVIS, I., 2003. *Drosophila* gurken (TGFalpha) mRNA localizes as particles that move within the oocyte in two dynein-dependent steps, *Devevolpmental Cell* **4**, 307–319.

MCGREGOR, A.P., ORGOGOZO, V., DELON, I., ZANET, J., SRINIVASAN, D.G., PAYRE, F. & STERN, D.L., 2007. Morphological evolution through multiple *cis*-regulatory mutations at a single gene. *Nature*. **448**, 587-590.

MAJOROS, W.H., OHLER, U., 2007. Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics*, **8:**152.

NIWA, R. & SLACK, F.J, 2007. The evolution of animal microRNA function. *Curr. Op. Gen. Dev*., 17, 145–150.

PATRAQUIM, P. & SUCENA, E., 2008. Evolução & Ontogenia. *in Evolução – História e argumentos*. Esfera do Caos Editores. Lisbon.

PEARSON, J.C., LEMONS, D., MCGINNIS, W.. 2005. Modulating Hox gene funtions during animal body patterning. *Nature Reviews Genetics*. **6**, 893-904.

RONSHAUGEN, M., BIEMAR, F., PIEL, J., LEVINE, M., LAI, E.C., 2005. The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. *Genes & Dev*elopment, **19**(24): 2947-52.

RUBY, J.G., STARK, A., JOHNSTON, W.K., KELLIS, M., BARTEL, D.P., & LAI, E.C., 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Research*, 17, 1850-1864.

SUCENA, E. & STERN, D.L., 2000. Divergence of Larval Morphology between Drosophila sechellia and Its Sibling Species Caused by Cis-Regulatory Evolution of Ovo/Shaven-Baby. *PNAS*. **97**(9), 4530-4534.

SHAPIRO, M.D., MARKS, M.E., PEICHEL, C.L., BLACKMAN, B.C., NERENG, K.S., JÓNSSON, B., SCHLUTER, D. & KINGSLEY, D.M., 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* **428**, 717–723.

STARK, A., BUSHATI, N., JAN, C.H., KHERADPOUR, P., HODGES, E., BRENNECKE, J., BARTEL, D.P., COHEN, S.M. & KELLIS, M., 2008. A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. *Genes & Development*. **22**, 8-13.

TANG, T., KUMAR, S., SHEN, Y. ET AL., 2010. Adverse interactions between micro-RNAs and target genes from different species. *PNAS*, early edition.

THOMSEN, S., AZZAM, G., KASCHULA, R., WILLIAMS, L.S., ALONSO, C.R., 2010. Developmental RNA processing of 3′UTRs in Hox mRNAs as a context-dependent mechanism modulating visibility to microRNAs. *Development*. **137**(17), 2951-2960.

TISHKOFF, S. A., REED, F.A., RANCIARO, A. *ET AL*, 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genet*. **39**, 31–40.

TYLER, D.M., OKAMURA, K., CHUNG, WJ., HAGEN, J. W., BEREZIKOV, E., HANNON, G.J. AND LAI, E.C., 2008. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes & Development*. **22**, 26-36.

VALENCIA-SANCHEZ, M.A., LIU J., HANNON G.J., PARKER R., 2006. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes & Development*, **20**(5), 515-524.

VAN VALEN, L., 1973. Festschrift. *Science* 180, 488.

WADDINGTON, C. H., 1957. *The Strategy of the Genes: a Discussion of Some Aspects of Theroetical Biology*. Macmillan. New York.

WARREN, R.W., NAGY, L., SELEGUE J., GATES, J., CARROLL, S., 1994. Evolution of homeotic gene regulation and function in flies and butterflies. *Nature*. **372**, 458-461.

WILKINS, A., 2002. *The Evolution of Developmental Pathways*. Sinauer Associates. Sunderland, Massachusetts.

6. SUPPLEMENTARY INFORMATION

6.1) Supplementary Figure Legends.

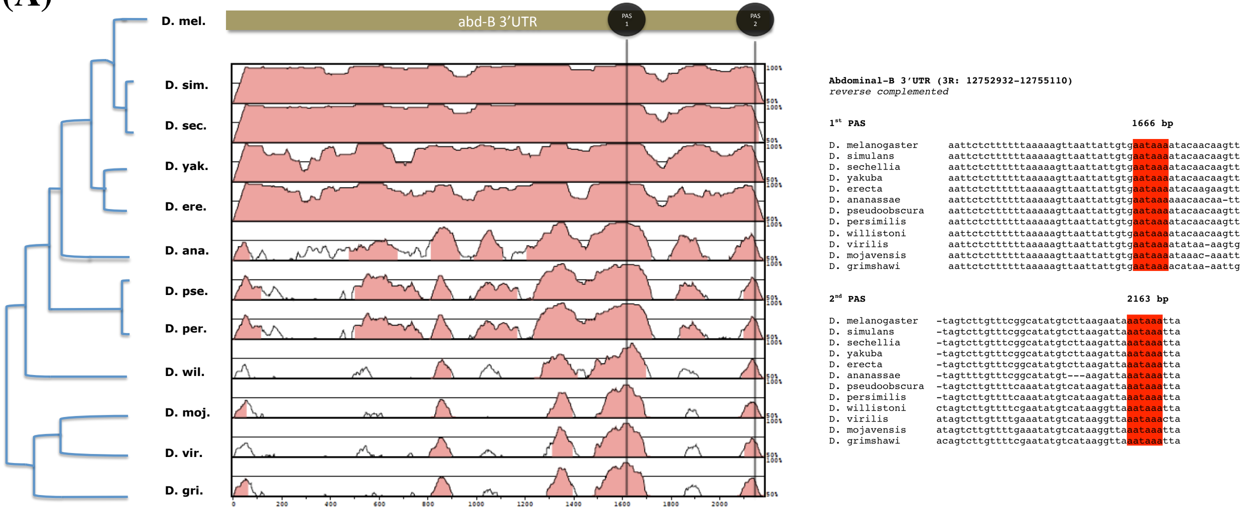
Figure S1. 3'UTR alignments for *abd-b, antp and abd-a.* Here we present the VISTA-LAGAN alignments for **(A)** *abdominal-b*, **(B)** *antennapedia* and **(C)**, *abdominal-a*. In each pannel, there is a detail on the polyadenylation signal site.

Figure S2. PITA Outputs: examples. Here we present two examples of the PITA software outputs **(A)** net targeting for *Ubx 3'UTRs* **(B)** Individual target-sites for the *Ubx* 3'UTR.

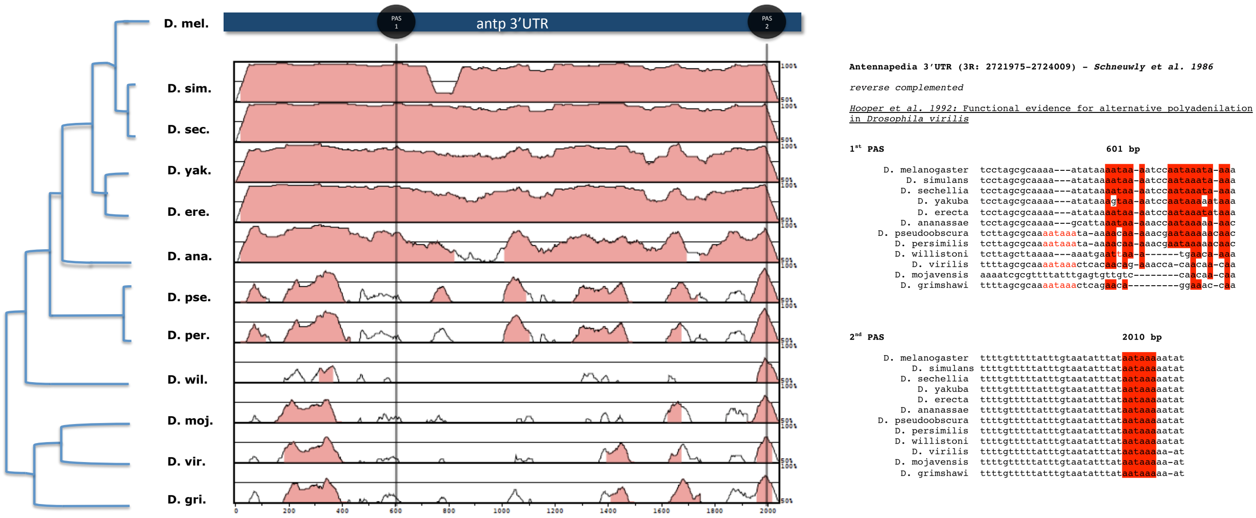
Figure S3. Individual miRNA target-site evolution within Drosophila *Ubx* 3'UTRs. Here we present all the individual target-site evolution profiles for the 14 miRNAs analysed. **(A)** miR-92a **(B)** miR-315 **(C)** miR-iab-4-3p **(D)** miR-92b **(E)** miR-190 **(F)** miR-998 **(G)** miR-3 **(H)** miR-184 **(I)** miR-318 **(J)** miR-993 **(K)** miR-210 **(L)** let-7 **(M)** miR-iab-4-5p **(N)** miR-iab-8-5p

SUPPLEMENTARY FIGURE 1

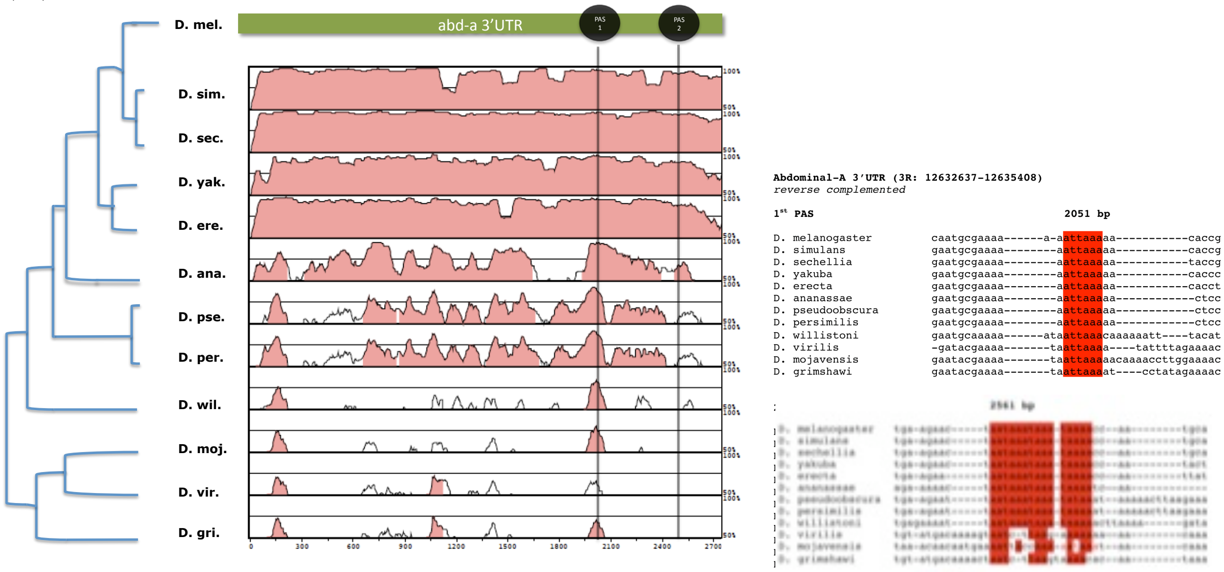
(A)



(B)



(C)



SUPPLEMENTARY FIGURE 2

(A)

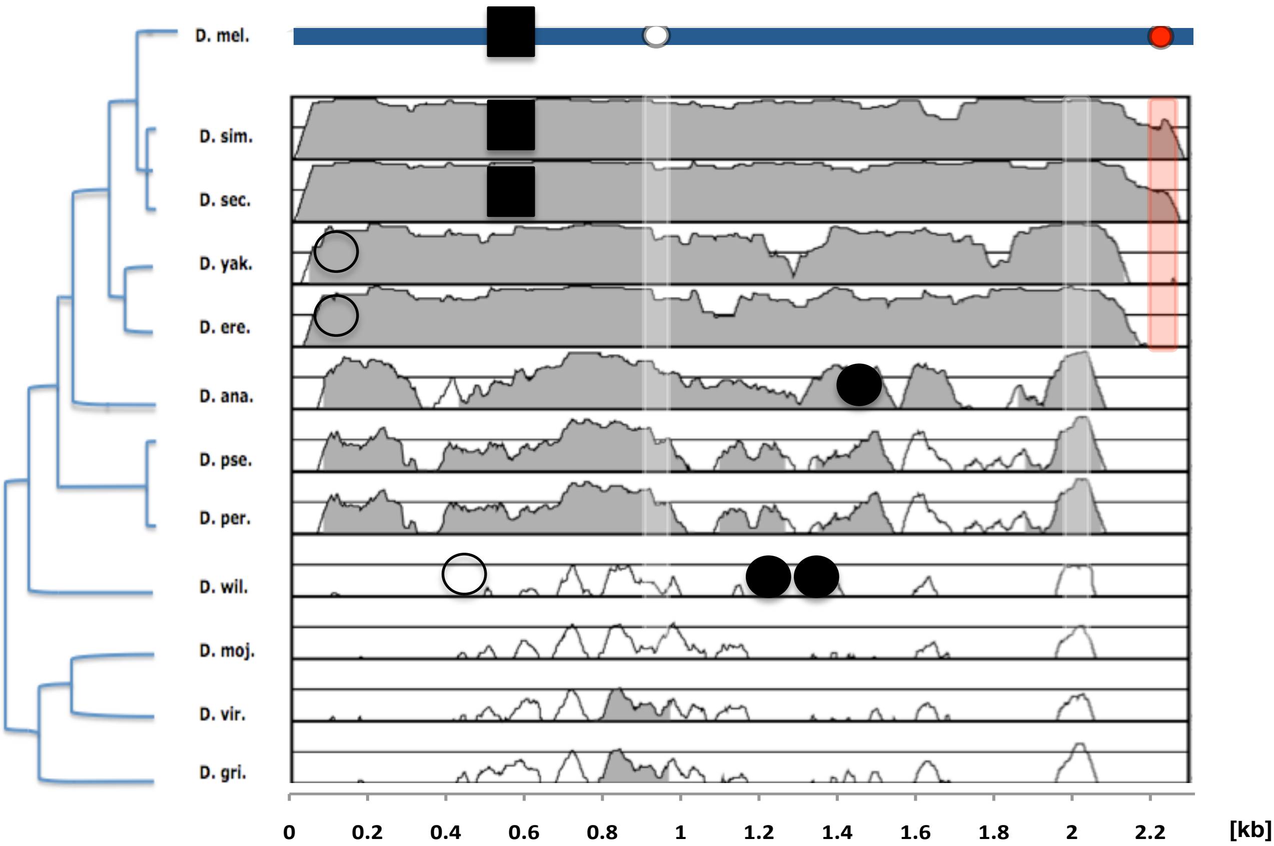
Gene	microRNA	Sites	Score
Seq1	dme-miR-317	5	-12.64
Seq1	dme-miR-313	7	-12.01
Seq1	dme-miR-954	2	-11.16
Seq1	dme-miR-92a	7	-10.51
Seq1	dme-miR-315	17	-10.04
Seq1	dme-let-7	5	-9.62
Seq1	dme-miR-966	4	-9.1
Seq1	dme-miR-309	3	-8.92
Seq1	dme-miR-iab-4-3p	1	-8.82
Seq1	dme-miR-375	12	-8.7
Seq1	dme-miR-973	4	-8.58
Seq1	dme-miR-92b	7	-8.51
Seq1	dme-miR-184	5	-8.45
Seq1	dme-miR-998	2	-8.23
Seq1	dme-miR-968	11	-8.22
Seq1	dme-miR-289	16	-7.6
Seq1	dme-miR-1008	4	-7.23
Seq1	dme-miR-210	2	-7.21
Seq1	dme-miR-960	2	-6.86
Seq1	dme-miR-280	12	-6.67
Seq1	dme-miR-312	7	-6.53
Seq1	dme-miR-993	3	-6.43
Seq1	dme-miR-287	4	-6.35
Seq1	dme-miR-190	1	-6.3
Seq1	dme-miR-iab-4as-5p	15	-6.2
Seq1	dme-miR-974	3	-5.77
Seq1	dme-miR-987	9	-5.76
Seq1	dme-miR-9a	8	-5.76

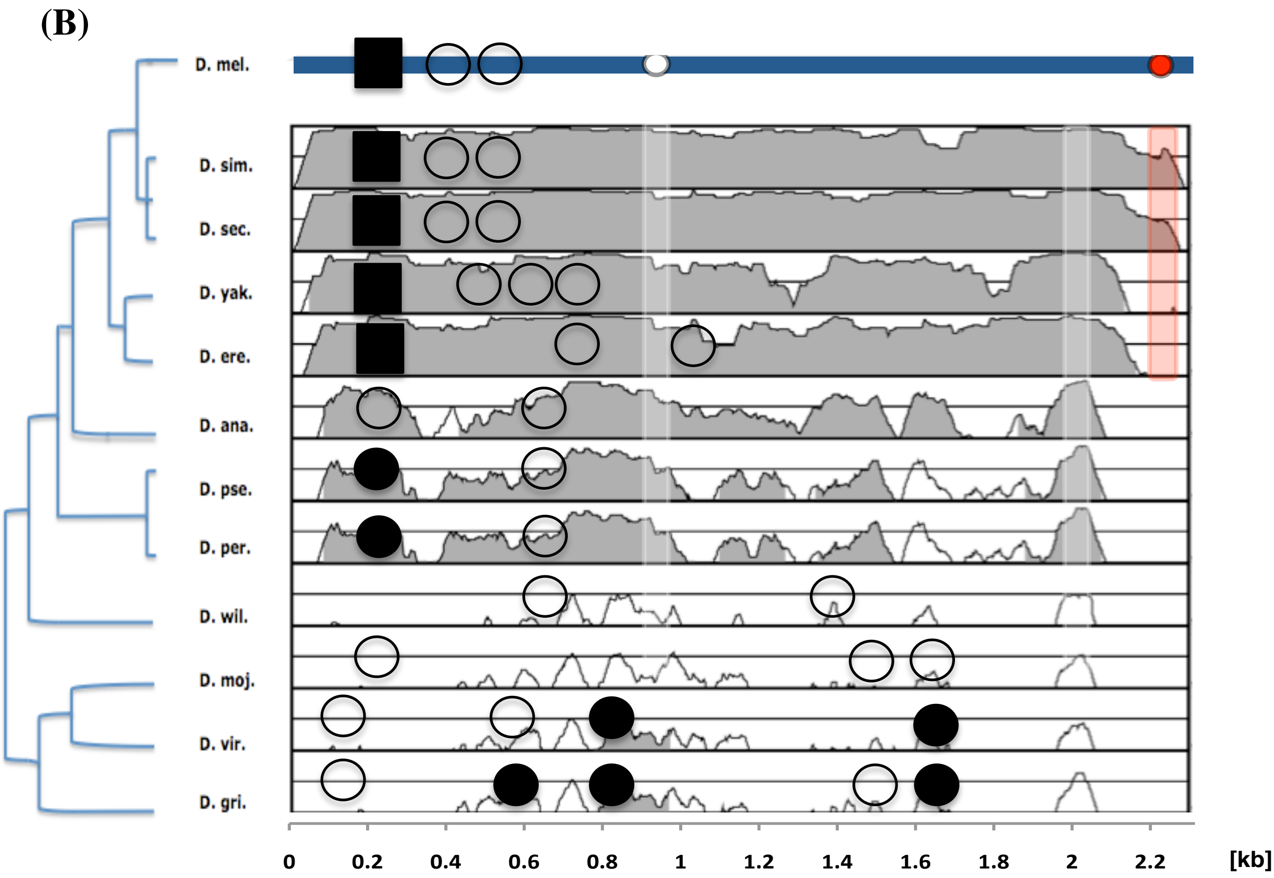
(B)

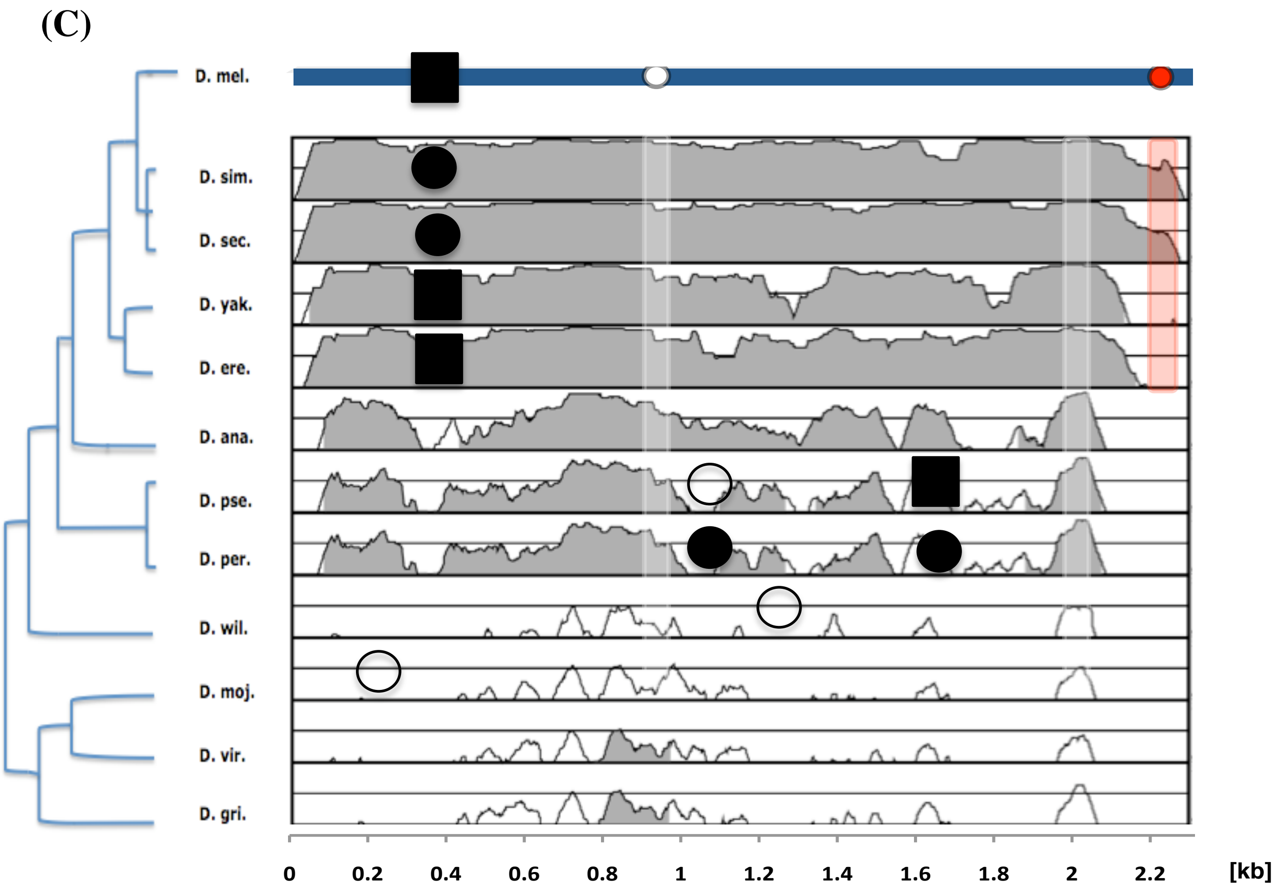
microRNA	Position	Seed	dGduplex	dGopen	ddG
dme-miR-317	1261	8:1:0	-25.7	-13.05	-12.64
dme-miR-313	919	8:1:1	-19.4	-7.38	-12.01
dme-miR-954	2089	8:1:1	-17.4	-6.23	-11.16
dme-miR-92a	919	8:1:1	-17.9	-7.38	-10.51
dme-miR-315	205	8:1:1	-20.1	-10.05	-10.04
dme-let-7	1878	8:1:1	-15.2	-5.57	-9.62
dme-miR-966	290	8:1:1	-18.9	-9.79	-9.1
dme-miR-309	2091	8:1:0	-15.8	-6.87	-8.92
dme-miR-iab-4-3p	475	8:1:0	-14.2	-5.37	-8.82
dme-miR-92b	919	8:1:1	-15.9	-7.38	-8.51
dme-miR-184	1975	8:0:1	-21.7	-13.25	-8.44
dme-miR-973	202	8:1:0	-20.79	-12.39	-8.39
dme-miR-375	342	8:1:0	-15.2	-6.86	-8.33
dme-miR-998	1331	8:1:0	-22.4	-14.16	-8.23
dme-miR-968	405	8:1:1	-19.6	-11.37	-8.22
dme-miR-289	1639	8:0:0	-10.29	-2.7	-7.58
dme-miR-375	317	8:1:0	-12.5	-5.19	-7.3
dme-miR-1008	1258	8:0:1	-20.4	-13.21	-7.18
dme-miR-210	2059	8:1:1	-14.42	-7.37	-7.04
dme-miR-960	434	8:1:0	-13.22	-6.35	-6.86
dme-miR-973	2087	8:1:0	-12.7	-5.86	-6.83
dme-miR-993	1256	8:1:0	-19.5	-13.06	-6.43
dme-miR-312	919	8:1:1	-13.8	-7.38	-6.41
dme-miR-317	410	8:1:1	-14.75	-8.4	-6.34
dme-miR-190	637	8:0:1	-16	-9.69	-6.3
dme-miR-iab-4as-5p	1672	8:0:0	-14.5	-8.3	-6.19

SUPPLEMENTARY FIGURE 3

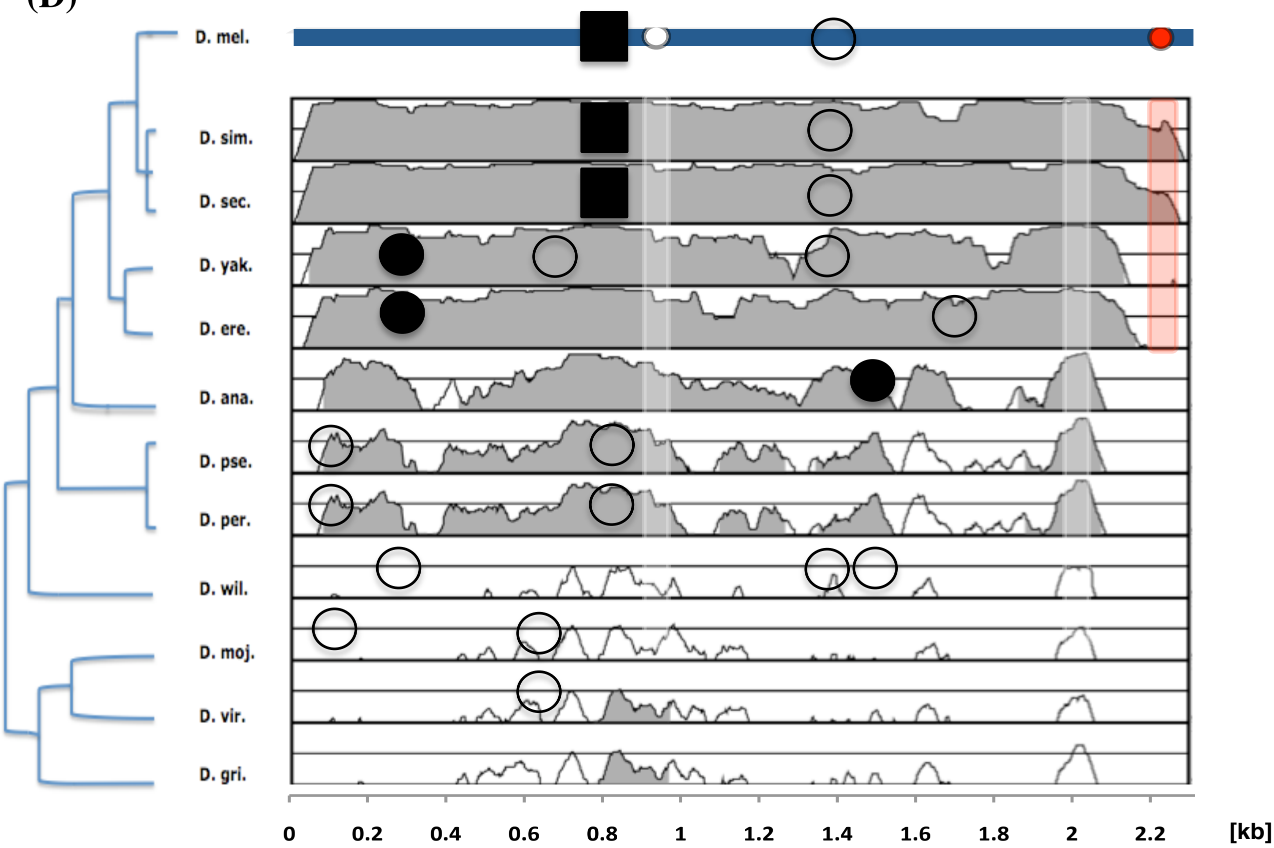
(A)



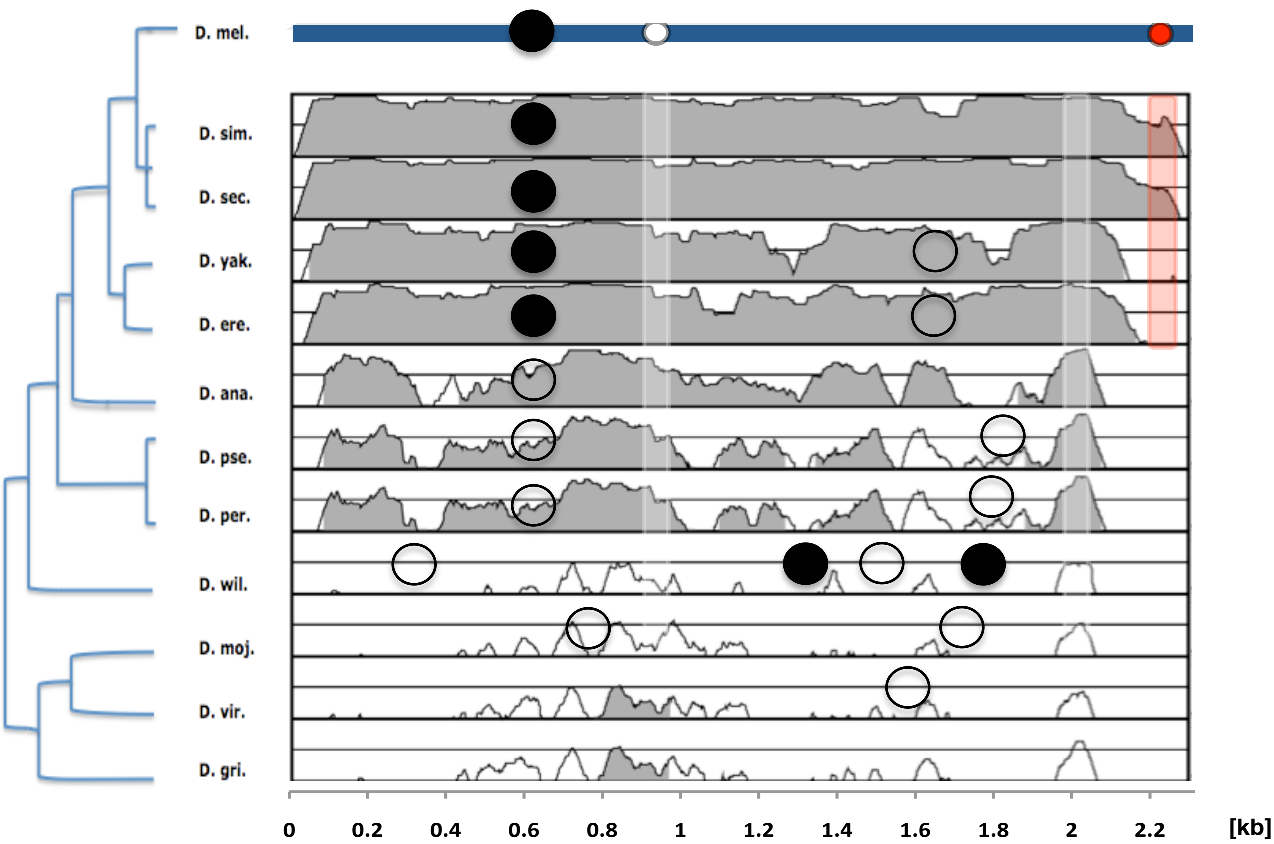

(B)
D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]


(C)
D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

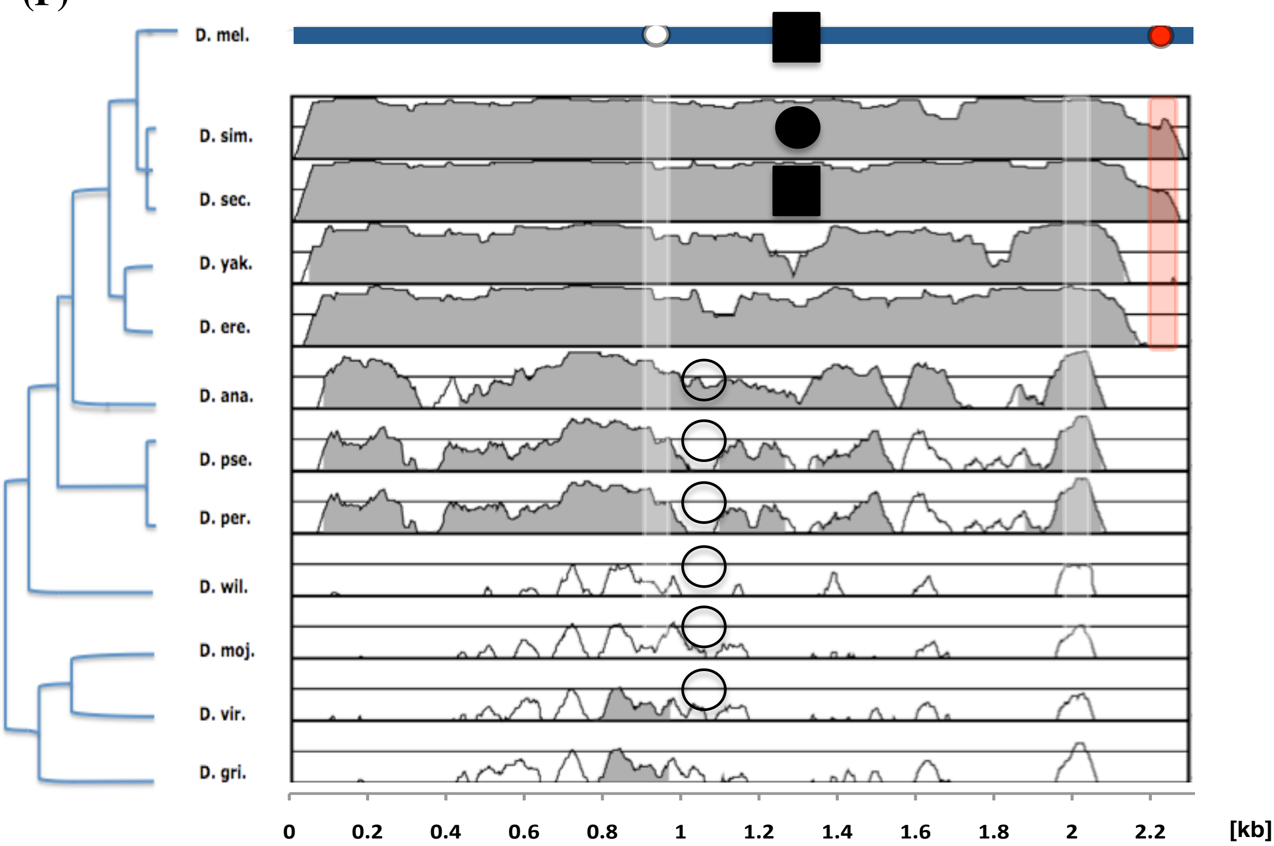
(D)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

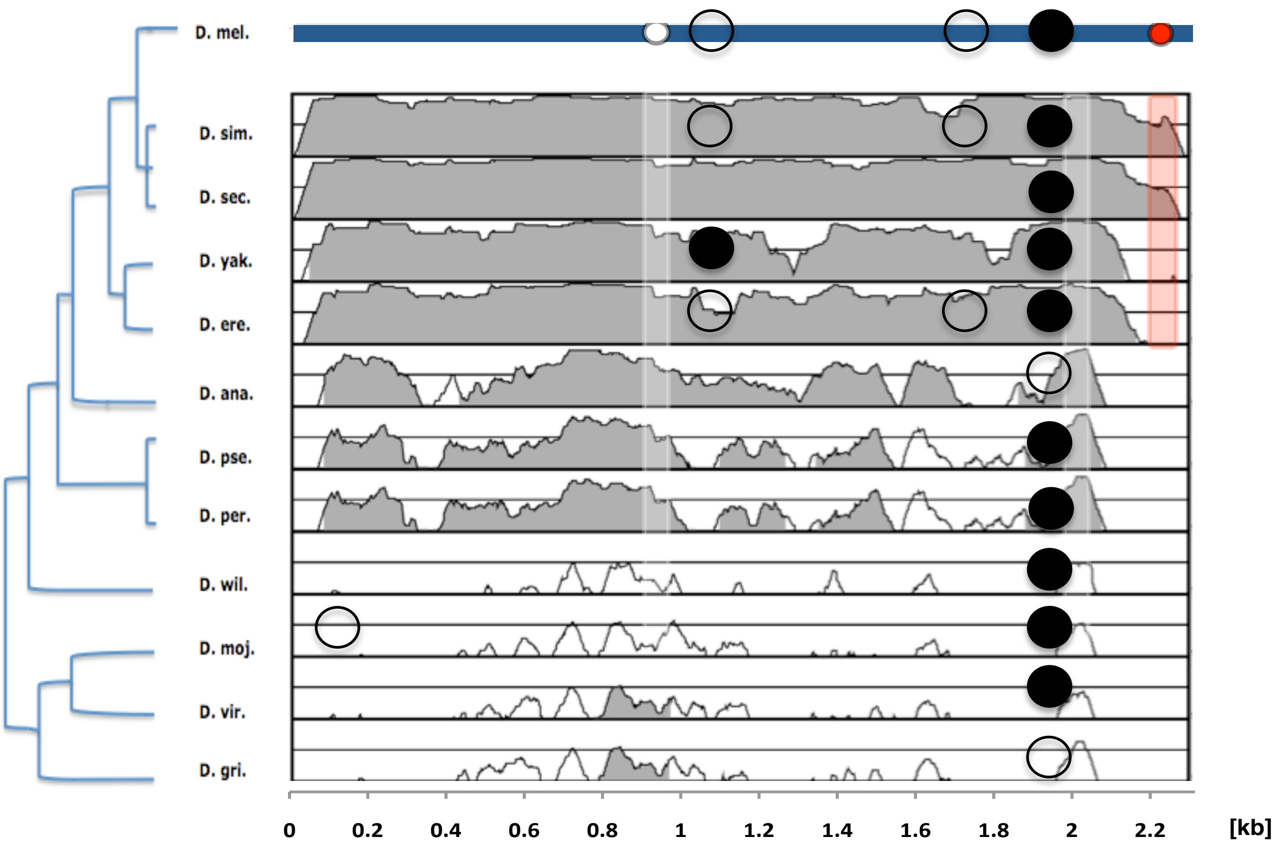
(E)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

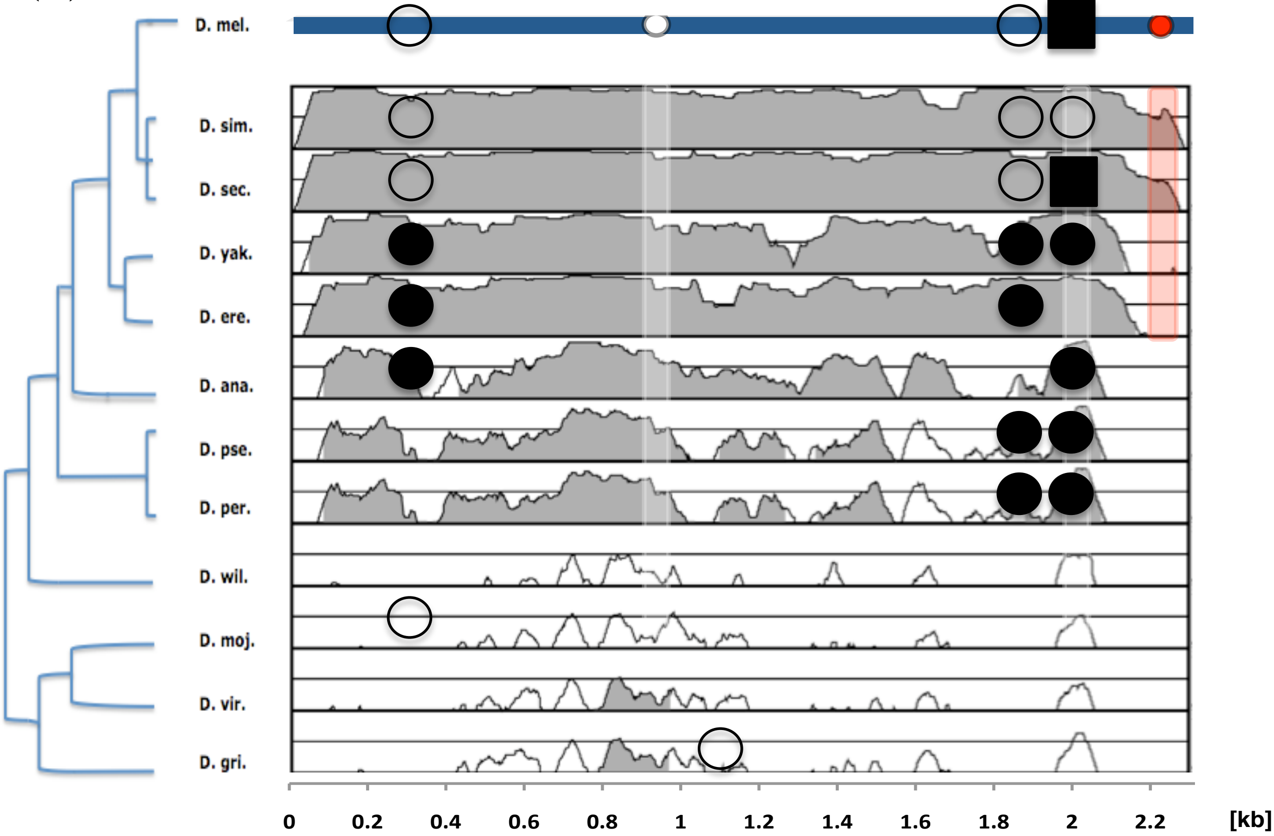
(F)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

(G)

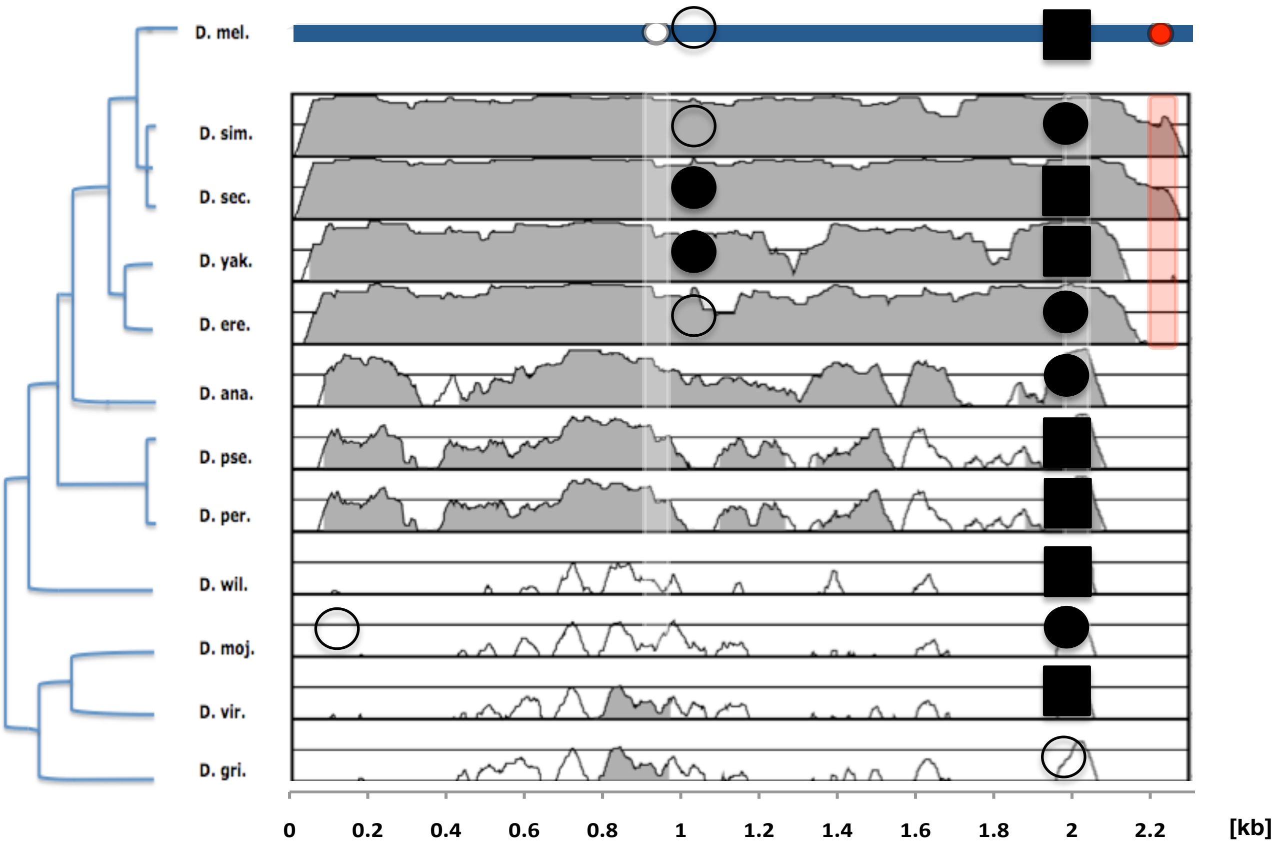

D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

(H)



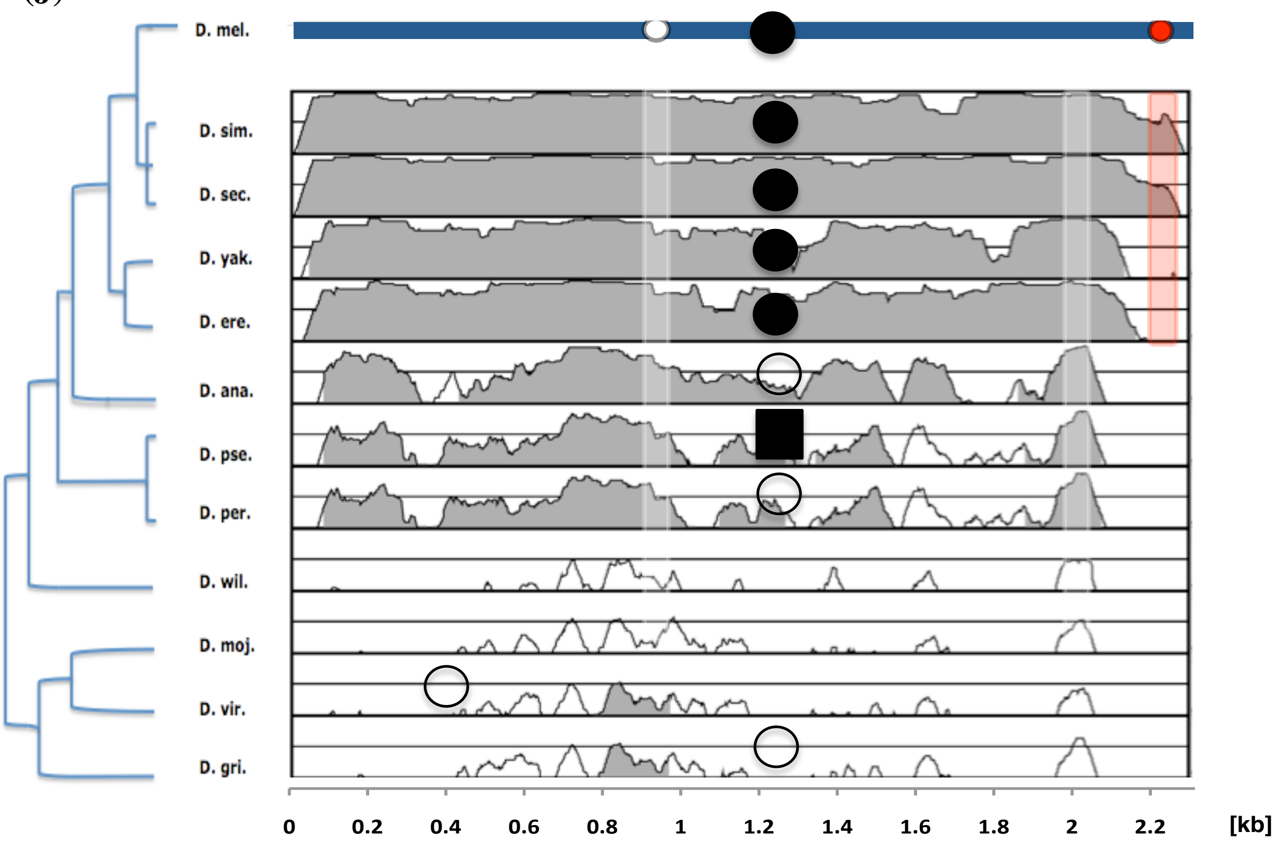
D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 [kb]

(I)

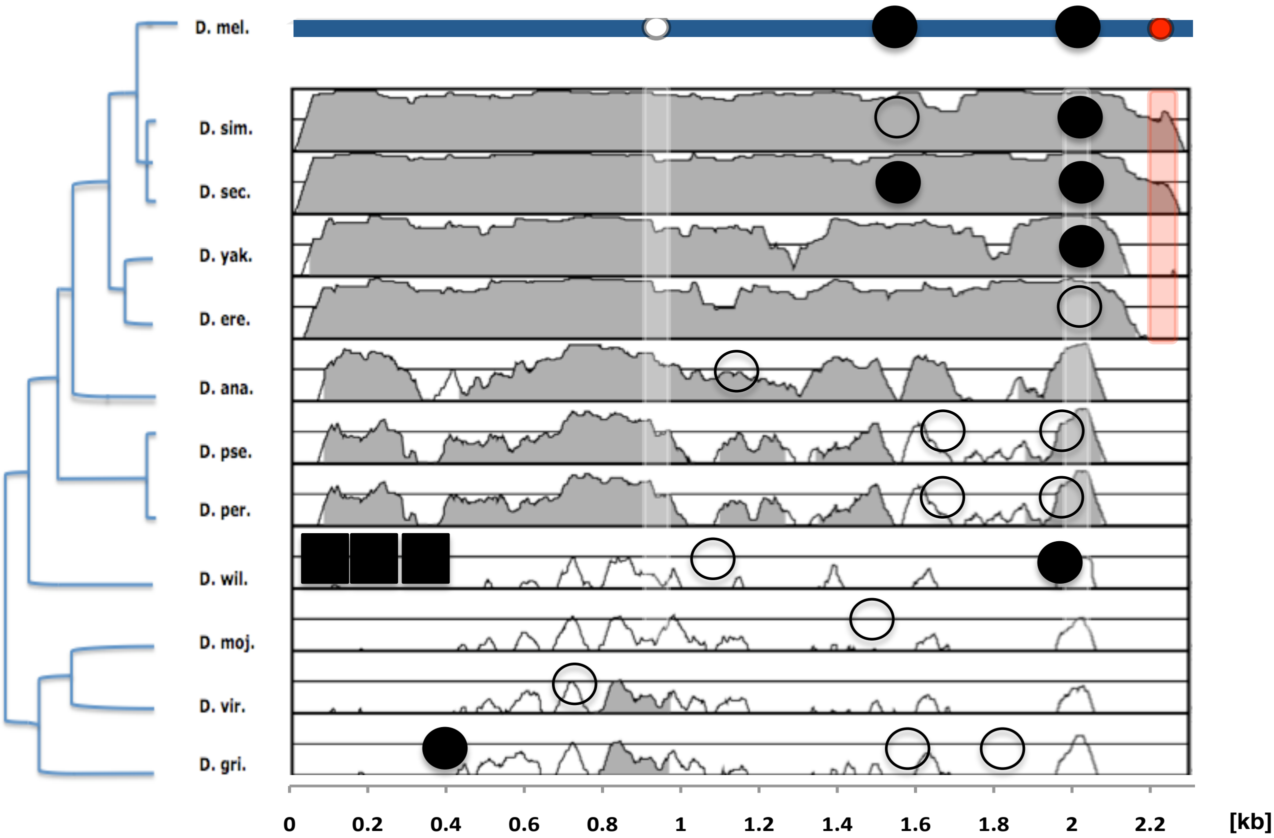


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8 2 2.2 [kb]

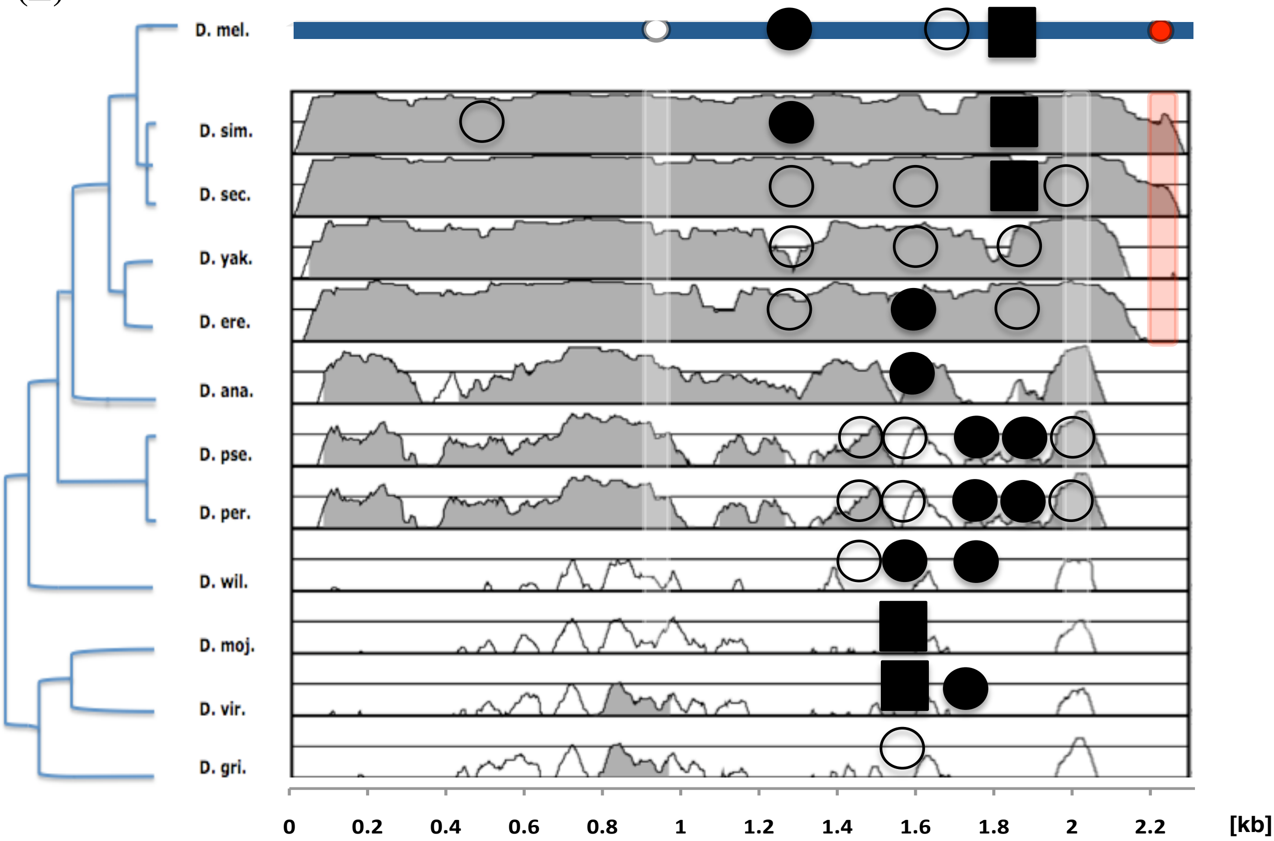
(J)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

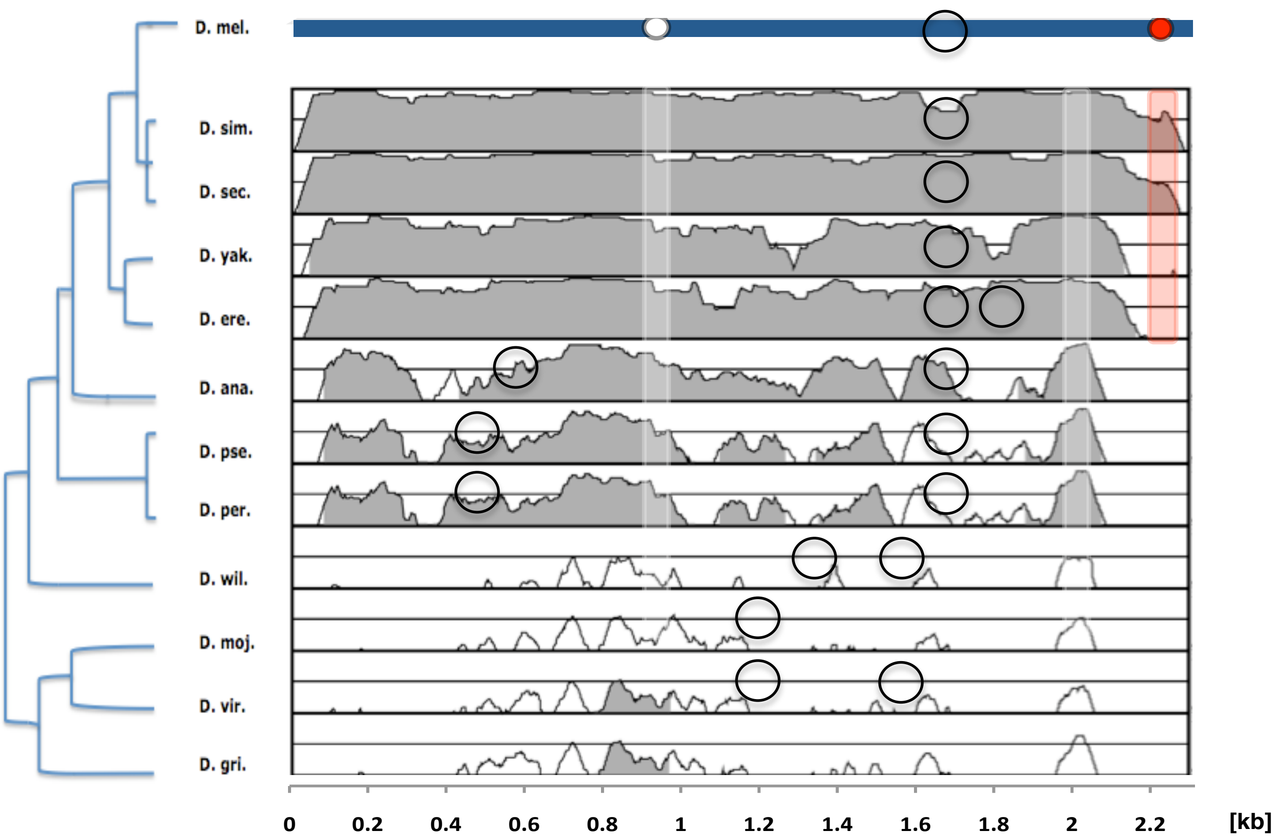
(K)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

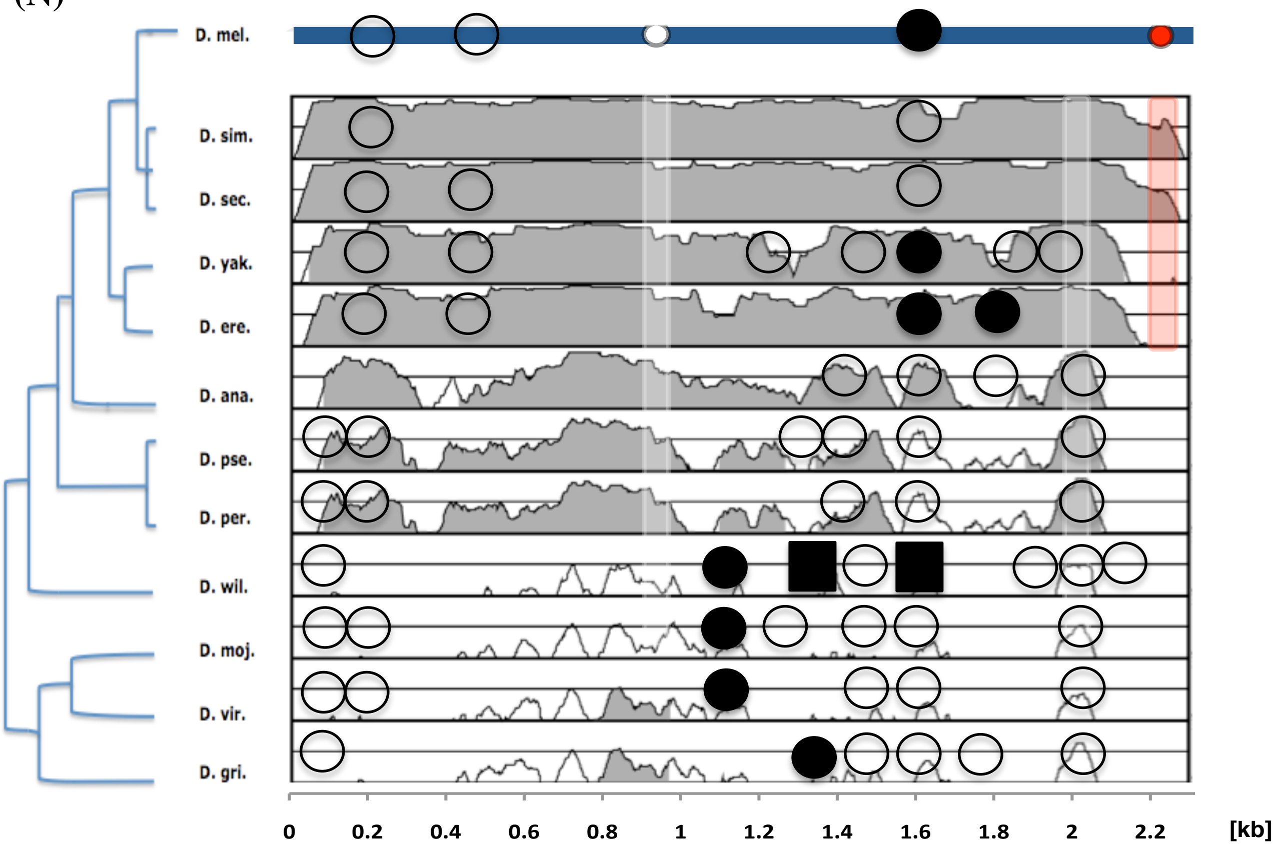
(L)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

(M)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

(N)


D. mel.
D. sim.
D. sec.
D. yak.
D. ere.
D. ana.
D. pse.
D. per.
D. wil.
D. moj.
D. vir.
D. gri.
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
[kb]

OXFORD JOURNALS
MOLECULAR BIOLOGY AND EVOLUTION

Evolution of Hox 3'UTR regulation by alternative polyadenylation and microRNA regulation within twelve Drosophila genomes

Journal:	*Molecular Biology and Evolution*
Manuscript ID:	Draft
Manuscript Type:	Letter
Date Submitted by the Author:	n/a
Complete List of Authors:	Patraquim, Pedro; University of Sussex, School of Life Sciences Alonso, Claudio; University of Sussex, School of Life Sciences
Key Words:	Development, Gene Regulation, miRNAs , Hox genes, Drosophila, RNA

Evolution of *Hox* 3'UTR regulation by alternative poly-adenylation and microRNA regulation within twelve *Drosophila* genomes

Pedro Patraquim and Claudio R. Alonso*

John Maynard Smith Building
School of Life Sciences
University of Sussex
Brighton BN1 9QG
United Kingdom

(*) To whom all correspondence should be addressed

Tel: +44 1273 876621 (Office)
+44 1273 877081 (Laboratory)
Email: c.alonso@sussex.ac.uk

ABSTRACT

The Hox genes encode a family of transcriptional regulators that operate differential developmental programs along the anteroposterior axis of animal bodies. Regulatory changes affecting Hox gene expression are believed to have been crucial for the evolution of animal body plans. In *Drosophila melanogaster*, Hox expression is post-transcriptionally regulated by microRNAs (miRNAs) acting on target sites located in Hox 3' untranslated regions (3'UTRs). Notably, recent work has shown that during development *Hox* genes produce mRNAs with variable 3'UTRs (short and long forms) in different tissues as a result of alternative polyadenylation; importantly, Hox short and long 3'UTRs contain very different target sites for miRNAs. Here we use a computational approach to explore the evolution of Hox 3'UTRs treated with especial regard to Hox miRNA regulation. Our work is focused on the twelve *Drosophila* species for which genomic sequences are available, and shows, first, that alternative polyadenylation of Hox transcripts is a feature shared by all Drosophilids tested in the study. Second, that the regulatory impact of miRNAs is evolving very fast within the *Drosophila* group, and, third, that in contrast to the low degree of conservation observed at the level of primary sequence Hox 3'UTR regions show very similar RNA topology indicating that RNA structure is under strong selective pressure. Finally, we also demonstrate that alternative polyadenylation leading to the formation of short and long Hox 3'UTRs can remodel the control regions seen by miRNAs by at least two mechanisms: by gradually adding target sites to a short 3'UTR form, as well as modifying the regulatory value of multiple miRNA target sites simultaneously through changes in RNA secondary structure.

INTRODUCTION

Hox genes encode a family of transcriptional regulators that pattern animal bodies along the anteroposterior axis (Lewis 1978)(McGinnis and Krumlauf 1992)(Alonso 2002). Evolutionary changes affecting Hox expression patterns and functions are thought to have mediated the evolution of animal body plans (Holland and Garcia-Fernàndez 1996; Pearson, Lemons, and McGinnis 2005). The molecular nature of such regulatory changes affecting *Hox* expression has yet not been completely resolved (Alonso and Wilkins 2005; Alonso 2008).

The regulation of mRNA levels in time and space seems to lie at the heart of the genetic programs controlling development. Such control of RNA expression levels relies on both transcriptional and posttranscriptional mechanisms (Alonso and Wilkins 2005; Davidson 2006; Alonso 2008). Current molecular models for gene expression indicate that information contained in mRNA 3'untranslated regions (3'UTRs) is read by the cell to determine patterns of mRNA decay, localisation and rates of protein translation (Moore 2005). Mechanistically these distinct outputs in RNA behaviour are determined by physical contacts between RNA-binding proteins and small RNAs, such as microRNAs (miRNAs); these RNA regulators are able to bind to specific *cis*-regulatory elements located in transcript 3'UTRs (Bartel 2004; Bartel and Chen 2004).

In *Drosophila melanogaster* Hox genes are regulated by miRNAs via miRNA target sequences located in Hox 3'UTRs (Ronshaugen et al. 2005; Bender 2008; Stark et al. 2008; Tyler et al. 2008). Intriguingly, recent work described that during development *Hox* genes produce mRNAs with variable 3'UTRs – i.e. short and long forms – in different tissues as a result of alternative poly-adenylation; notably, short and long 3'UTRs contain very different target sites for miRNAs converting each mRNA species into substantially different miRNA targets (Thomsen et al. 2010).

Here we use publicly available genome sequences from twelve *Drosophila* species to investigate the evolution of 3'UTR regions in the *Drosophila Hox* genes *Ultrabithorax* (*Ubx*), *abdominal-A* (*abd-a*), *abdominal-b* (*abd-b*) and *antennapedia* (*antp*) searching for variation affecting primary sequence and secondary structure. We focus our analysis on the distribution of discrete *cis*-

regulatory modules including poly-adenylation signals and miRNA target sites, and RNA structural features affecting local and global topology of Hox 3'UTRs predicted to impact the recognition of Hox mRNAs by RNA regulators.

RESULTS AND DISCUSSION

Evolution of alternative polyadenylation of *hox* genes: conservation and plasticity

To assess evolutionary constraints on *hox* 3'UTR sequence size, we examined the conservation of Hox poly-adenylation signal sites (PASs) for *Ubx, abd-a, abd-b* and *antp* within the twelve Drosophilids for which genomic sequences are available. Sequences were retrieved from the UCSC genome browser and aligned using *mVISTA-LAGAN* software (Brudno et al. 2003). Our analysis showed that poly-adenylation signals leading to the production of (at least) two alternative transcripts of distinct length are conserved throughout the group; however, the exact position of the poly-adenylation signals within each mRNA transcript shows some variation from species to species indicating certain level of plasticity in the mechanism of alternative poly-adenylation (FIGURE 1a). Interestingly we find that transcript total length was approximately maintained within the group: in those species where the proximal 3'UTR is shorter, the distal 3'UTR is extended (FIGURE 1b). These observations indicate that the basic alternative poly-adenylation patterns found in *D. melanogaster* Hox sequences are conserved throughout the group suggesting that Hox alternative poly-adenylation is likely to be a feature present in the common ancestor of the group.

miRNA regulation shows distinct and dynamic evolutionary profiles across Drosophilids.

Two main factors determine the likelihood of a given miRNA to regulate a target mRNA via 3'UTR sequences: primary sequence composition at target sites and local RNA topology (Kertesz et al. 2007; Long et al. 2007; Li et al. 2010). Therefore the combined computation of sequence composition and RNA accessibility in target 3'UTRs allows for an accurate calculation of the regulatory impact of miRNAs on mRNA targets (Kertesz et al. 2007; Long et al. 2007). To investigate the evolution of miRNA regulation on the 3'UTRs of Hox genes we foces on *Ubx,* in which miRNA regulation is better understood, and submitted the *Ubx* 3'UTR sequence for all Droophilids to the prediction tool PITA (Kertesz et al. 2007) which computes target sequence and RNA topology simultaneously. PITA represents predicted regulatory strength in the form of a energy-based score termed ΔΔG, ascribed to a given miRNA-mRNA interaction; this value is calculated by subtracting the free energy lost by opening locally paired RNA structures to the free energy gained by the formation of a specific miRNA-mRNA duplex. To maximise the stringency of our analysis we focused on those miRNAs which were known to be temporally co-expressed with Ubx in *D. melanogaster* and for which miRNA seeds are known to be ultraconserved within the 12 Drosophilid genomes (Ruby et al. 2007). This approach identified fourteen miRNAs which we used for further analysis.

Regulatory analysis of each one of the fourteen miRNAs within the twelve Drosophilid phylogeny shows that the evolution of miRNA regulation varies as per miRNA: 6 miRNAs show a marked tendency to increase their regulatory impact on Hox transcripts during evolution (Figure 2a), 1 miRNA shows a decrease in its regulatory impact (Figure 2b), and 3 display no significant regulatory changes within the group (Figure 2c). The remaining 3 miRNA species show no obvious evolutionary pattern. Interestingly, different miRNAs from the iab4/iab8 locus belong to different categories: for instance, miR-iab4-3p shows a markedly positive evolutionary trend while miR-iab4-5p shows a constant profile across all Drosophilids. We also explored how the evolution of individual target-sites relates to net miRNA-dependent regulatory effects on *Ubx*. For this we took PITA outputs for all predicted targets for the fourteen miRNAs and divided the resulting 317 miRNA target sites into three broad categories: *strong* ($\Delta\Delta G$<-8), *mild* (-8<$\Delta\Delta G$<-4) or *weak* sites ($\Delta\Delta G$>-2). We then related these values to the Drosophilid phylogeny. This analysis first revealed that the targeting of a given *Ubx* mRNA by miRNAs seems to depend on a predominant *'core'* site, responsible for most of the net regulatory value, followed by *'auxiliary'* miRNA target sites which when present, make a small contribution to net mRNA target regulation. Secondly, 76.7% of the miRNA target-sites remained within the same 3'UTR isoform (long or short) in which the *core* target-site lies in *D. melanogaster* (Figure 3a) (χ^2=7.679>6.635). Third, *core* sites tended to be less evolutionarily volatile, as only 3 out of 38 *core* sites identified are not conserved in Ubx 3'UTRs, in contrast with the 59% conservation level observed for *auxiliary* sites. Fourth, we superimposed individual target-site evolution to the positive, negative, and neutral net miRNA-mRNA patterns described above, and found that neutral behaviour or *stasis* is linked to a higher level of conservation of auxiliary (*mild* and *weak*) sites (χ^2 for α=0.1 3.563>2.706), while the converse is true for *positive* trends which show a higher level of conservation of *core* miRNA sites (χ^2 for α=0.10 was 2.966>2.706). (We were unable to relate miRNA site regulatory strength to negative evolutionary trends, given that this type of pattern was only represented by one case in our sample – see above). These results suggest a directional, gradual, and quantitative model for the evolution of miRNA regulation of target mRNAs. This analysis also shows that mRNA target regulation is dominated by *core* miRNA target sites and that most auxiliary sites tend to remain in the same 3'UTR location suggesting that miRNA regulation is isoform-specific for mRNAs with alternative 3'UTRs.

RNA accessibility in Hox 3'UTRs is ultraconserved despite significant change at the primary-sequence level

The conservation of primary 3'UTR sequences within the Drosophila group is limited to small stretches of sequence (see Figure 1A). However, given that miRNA regulation of target genes relies on both, primary sequence as well as on RNA secondary structure we decided to test to what extent 3'UTR secondary structure had evolved within the Drosophila group. For this we first divided the *D. melanogaster* Hox long 3'UTRs into 200bp windows and then used *mVISTA-LAGAN* primary sequence alignments (Figure 1a) to define homologous regions for each *D. melanogaster* window in all species. We then calculated average accessibility values for each window in all 12 Drosophilids analysed using one of the intermediate outputs of PITA (ΔG_{open}) (Figure 4). Strikingly, this experiment revealed that patterns of RNA secondary structure within Hox 3'UTRs are extremely conserved (Figure 4a) suggesting that the maintenance of a particular RNA topology is likely to be under strong selective pressure. As a negative control, we used exactly the same approach to analyse an untranscribed intergenic region of the same length (confirmed by RNAseq data available at *www.flybase.org*), and found out that for this region there is very little conservation of accessibility patterns among the twelve Drosophilids (Figure 4b). We further validated these differences by looking at the profiles of variance in accessibility values between these two cases (Figure 4C). These results suggest that target RNA secondary stucture may play a very significant role for the evolution of miRNA regulation. We also looked at the putative role of alternative poly-adenylation as an effector of an RNA secondary structure switch. For this we focused on *Ubx*, for which miRNA regulation is known in higher detail, using the *RNAFold* software (Vienna package) (Hofacker 2003) to fold entire Ubx mRNA sequences including only the *proximal* (Figure 5a) or the *full* 3'UTR (Figure 5b). We observed that sequences located at the end of the proximal 3'UTR are predicted to change in secondary structure when the distal tract is also present in the molecule (Figure 5c). We further confirmed this by comparing accessibility values for both long and short 3'UTRs in the previously mentioned alternatively poly-adenylated Hox genes, delimitating an "unstable" region starting around 80bp upstream of the first poly-adenylation signal. We compared *Ubx:short* vs. *Ubx:long* PITA miRNA target-site predictions for this region, and found that the regulatory strength of miRNA sites for miR-312, miR-313, miR-92a and miR-92b (all of which are located within the proximal 3'UTR segment) is predicted to decrease when the long 3'UTR is also present in the 3'UTR (Figure 5d). These results point to a previously unknown post-transcriptional mechanism, by which the addition of a nucleotide stretch to the 3'-UTR of a transcript changes the structure of the constitutive 3'UTR. This implies that alternative poly-adenylation can remodel the regulatory landscape of the mRNA molecule by at least two mechanisms: by gradually adding target sites to a given 3'UTR form, and by a *non-linear* mechanism modifying the regulatory value of multiple miRNA target sites simultaneously through changes in RNA secondary structure.

REFERENCES

Alonso, C. R. 2002. Hox proteins: sculpting body parts by activating localized cell death. Curr Biol **12**:R776-778.

Alonso, C. R. 2008. The molecular biology underlying developmental evolution *in* A. a. F. Minelli, G., ed. Key Themes in Evolutionary Developmental Biology. Cambridge University Press.

Alonso, C. R., and A. S. Wilkins. 2005. The molecular elements that underlie developmental evolution. Nat Rev Genet **6**:709-715.

Bartel, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell **116**:281-297.

Bartel, D. P., and C.-Z. Chen. 2004. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet **5**:396-400.

Bender, W. 2008. MicroRNAs in the Drosophila bithorax complex. Genes & Development **22**:14-19.

Brudno, M., C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, N. C. S. Program, E. D. Green, A. Sidow, and S. Batzoglou. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Research **13**:721-731.

Davidson, E. H. 2006. The Regulatory Genome. Academic Press.

Hofacker, I. L. 2003. Vienna RNA secondary structure server. Nucleic Acids Res **31**:3429-3431.

Holland, P. W., and J. Garcia-Fernàndez. 1996. Hox genes and chordate evolution. Dev Biol **173**:382-395.

Kertesz, M., N. Iovino, U. Unnerstall, U. Gaul, and E. Segal. 2007. The role of site accessibility in microRNA target recognition. Nat Genet **39**:1278-1284.

Lewis, E. B. 1978. A gene complex controlling segmentation in Drosophila. Nature **276**:565-570.

Li, X., G. Quon, H. D. Lipshitz, and Q. Morris. 2010. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA **16**:1096-1107.

Long, D., R. Lee, P. Williams, C. Y. Chan, V. Ambros, and Y. Ding. 2007. Potent effect of target structure on microRNA function. Nat Struct Mol Biol **14**:287-294.

McGinnis, W., and R. Krumlauf. 1992. Homeobox genes and axial patterning. Cell **68**:283-302.

Moore, M. J. 2005. From birth to death: the complex lives of eukaryotic mRNAs. Science **309**:1514-1518.

Pearson, J. C., D. Lemons, and W. McGinnis. 2005. Modulating Hox gene functions during animal body patterning. Nat Rev Genet **6**:893-904.

Ronshaugen, M., F. Biemar, J. Piel, M. Levine, and E. C. Lai. 2005. The Drosophila microRNA iab-4 causes a dominant homeotic transformation of halteres to wings. Genes & Development **19**:2947-2952.

Ruby, J. G., A. Stark, W. K. Johnston, M. Kellis, D. P. Bartel, and E. C. Lai. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. Genome Research **17**:1850-1864.

Stark, A., N. Bushati, C. H. Jan, P. Kheradpour, E. Hodges, J. Brennecke, D. P. Bartel, S. M. Cohen, and M. Kellis. 2008. A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. Genes & Development **22**:8-13.

Thomsen, S., G. Azzam, R. Kaschula, L. S. Williams, and C. R. Alonso. 2010. Developmental RNA processing of 3'UTRs in Hox mRNAs as a context-dependent mechanism modulating visibility to microRNAs. Development (Cambridge, England).

Tyler, D. M., K. Okamura, W.-J. Chung, J. W. Hagen, E. Berezikov, G. J. Hannon, and E. C. Lai. 2008. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. Genes & Development **22**:26-36.

FIGURE LEGENDS

Figure 1

Alternative polyadenilation is conserved within the Drosophilids (A) The *Ubx* gene in *Drosophila melanogaster* produces two alternatively poly-adenylated mRNA forms: *Ubx* short 3'UTR and Ubx long 3'UTR (see top diagram). Multiple-alignments for Drosophilid *Ubx* 3'UTR primary sequences using the VISTA-LAGAN software. *Ubx Drosophila melanogaster* 3'UTR sequences (represented by a blue bar – see top) is used as a baseline sequence – see top rectangle. *Drosophila melanogaster* poly-adenilation signals (PAS) are shown: PAS1 in white and PAS2 in red; a putative additional PAS is shown in black with an ultraconserved canonical sequence (AATAAA). Sequence homology is represented on the vertical axis of each aligned sequence, with a minimal value of 50% and a maximum value of 100% (gray regions correspond to segments with 70% or more of sequence similarity). **(B)** The ratio of distal/proximal 3'UTR length is generally greater in species more closely related to *Drosophila melanogaster*, with the first poly-adenylation signal site receding approximately 350bp across this time-window.

Figure 2

Quantitative evolution of miRNA regulation of Ubx. The figure shows the regulatory evolution of seven miRNAs illustrating **(A)** a positive (increase in predicted regulatory effects) evolutionary trend, **(B)** a negative (decrease in predicted regulatory effects) trend, and **(C)** a *stasis* trend (no significant change in predicted regulatory effects). Notice that miRNAs produced from the iab-4/iab-8 locus – which have been experimentally shown to target *Ubx* mRNAs – have distinct evolutionary trends regarding the targeting of *Ubx 3'UTR:long* mRNAs.

Figure 3

Evolution of individual miRNA target-sites within Drosophilid Ubx 3'UTR sequences. **(A)** The diagram shows the evolution of miRNA target sites for miRNA-iab-4-3p (green) and miRNA-iab-4-5p (red) within Ubx 3'UTR sequences. miRNA target sites for each miRNA species are depicted according to their regulatory strength (full circle, strong sites; empty circle, mild site, empty squares, weak sites). Notice the recent acquisition of strong conserved sites for miR-iab4-3p within the melanogaster subgroup (including *D. mel. D. sim, D. sec., D. yak,* and *D. ere.*), pointing to a likely recent regulatory novelty. (**B)** Diagram describing the evolution of let-7 target sites within the Drosophila group. Notice the general most miRNA target sites are located in the long 3'UTR form of Ubx indicating a trend towards maintaining let-7 target sites witin one of two *Ubx* isoforms only. This illustrates a general trend observed for the majority of miRNA sites analysed, which tend to remain within one specific mRNA isoform (see text for further details).

Figure 4

RNA accessibility is conserved across Hox 3'UTRs, unlike primary sequence

(A) RNA accessibility alignments for *Ubx* 3'UTRs. Homologous regions were ascribed to 200 bp homology windows using *Drosophila melanogaster* as a baseline. A measure of RNA accessibility (ΔG_{open}) is plotted vs. Ubx 3'UTR length. Low ΔG_{open} values indicate low accessibility. Despite significant divergence at the level of primary sequence, the accessibility of homologous regions of the *Ubx* 3'UTR remains generally constant. (**B)** ΔG_{open} values for a control sequence extracted from an untranscribed intergenic region in *D. mel.* (Ch3R:12604500-12607000); note the high level of variation in ΔG_{open} values observed in this case. **(C)** Variance analysis of *Ubx* and the control intergenic segment**;** while Ubx variance in accessibility values remains fairly unchanged across the Ubx 3'UTR, variance for the control segment shows distinct peaks revealing lack of conservation in secondary structure predictions.

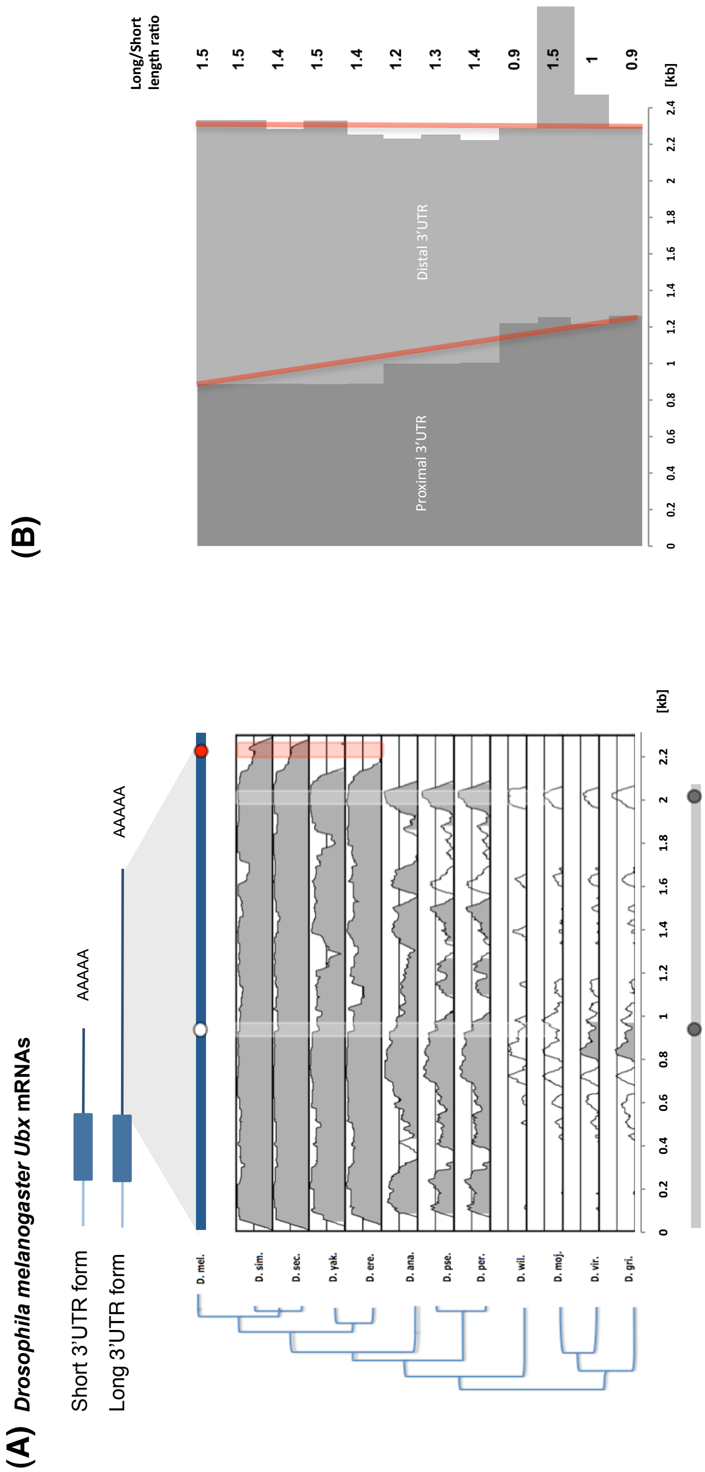
Figure 5

RNAFold secondary structure predictions of *hox* gene mRNAs

Anatomy of *Ubx* mRNA transcripts in regards to RNA secondary structure. Ubx mRNA regions are represented in different colours: 5'UTR sequences (pale blue), coding sequences (grey), proximal 3'UTR (green), and distal 3'UTR (dark blue). Although the broad secondary structure of the proximal 3'UTR sequences is maintained in both, (**A**) short and (**B**) long 3'UTR tails, the configuration of the RNA molecules in the absence of the distal 3'UTR segment is significantly different from the one obtained with the inclusion of distal 3'UTR sequences in a subregion of 150 ribonucleotides immediately upstream of the first PAS. **(C)** Magnification of the 150 nucleotide subregion within the proximal Ubx 3'UTR which changes its conformation according the alternative poly-adenylation pattern used during the transcript RNA processing. (**D**) The conformational change affecting the 150mer mentioned above is predicted to affect the effectiveness of miRNA target-sites in the region by changing their accessibility values. Note that these miRNAs have a similar seed sequence, hence the same target position (35 bp upstream of the first poly-adenylation signal).

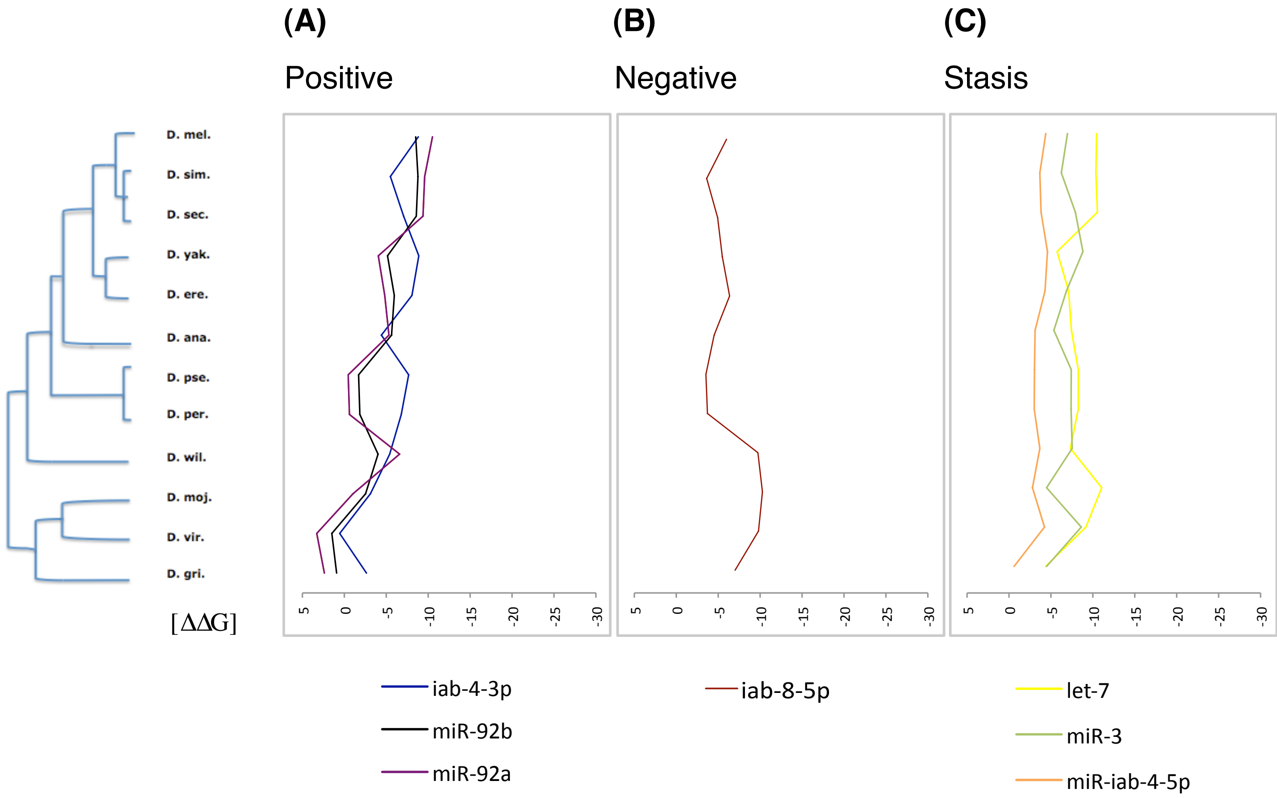
Patraquim & Alonso

Figure 1



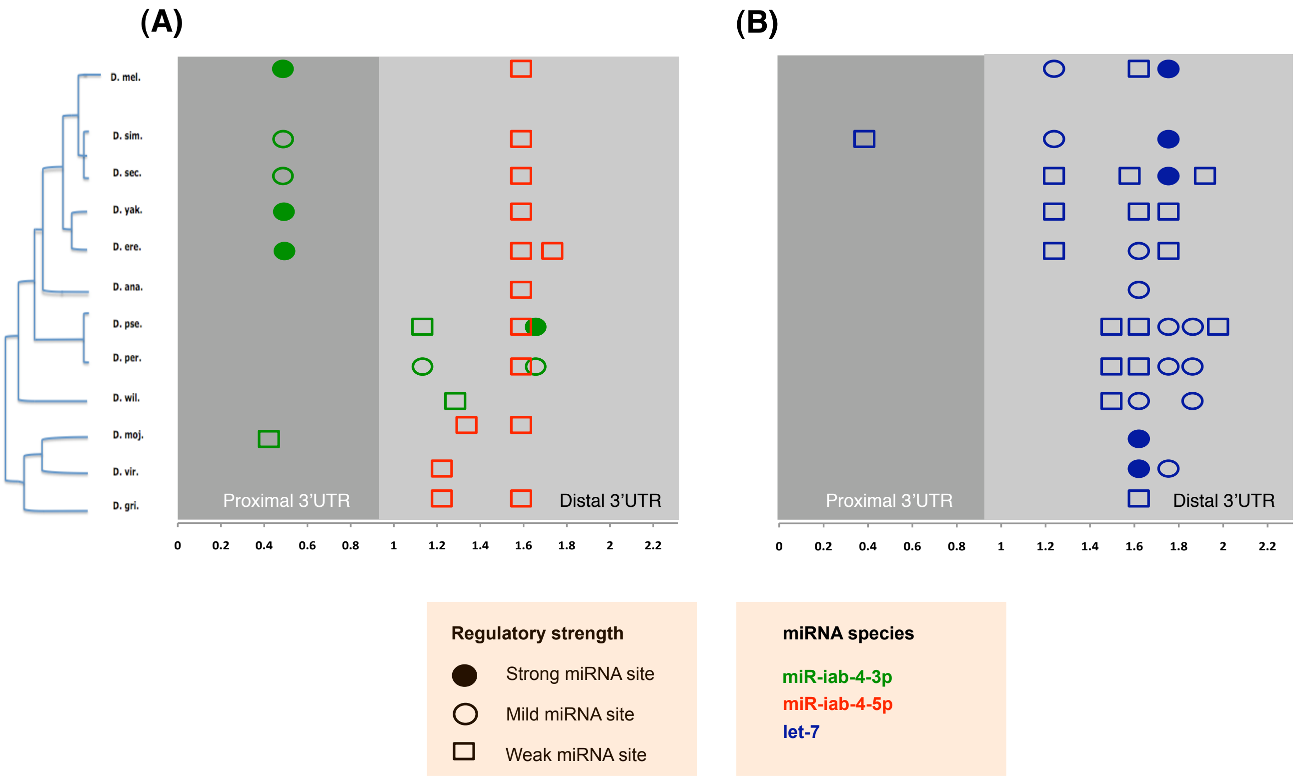
Patraquim & Alonso

Figure 2



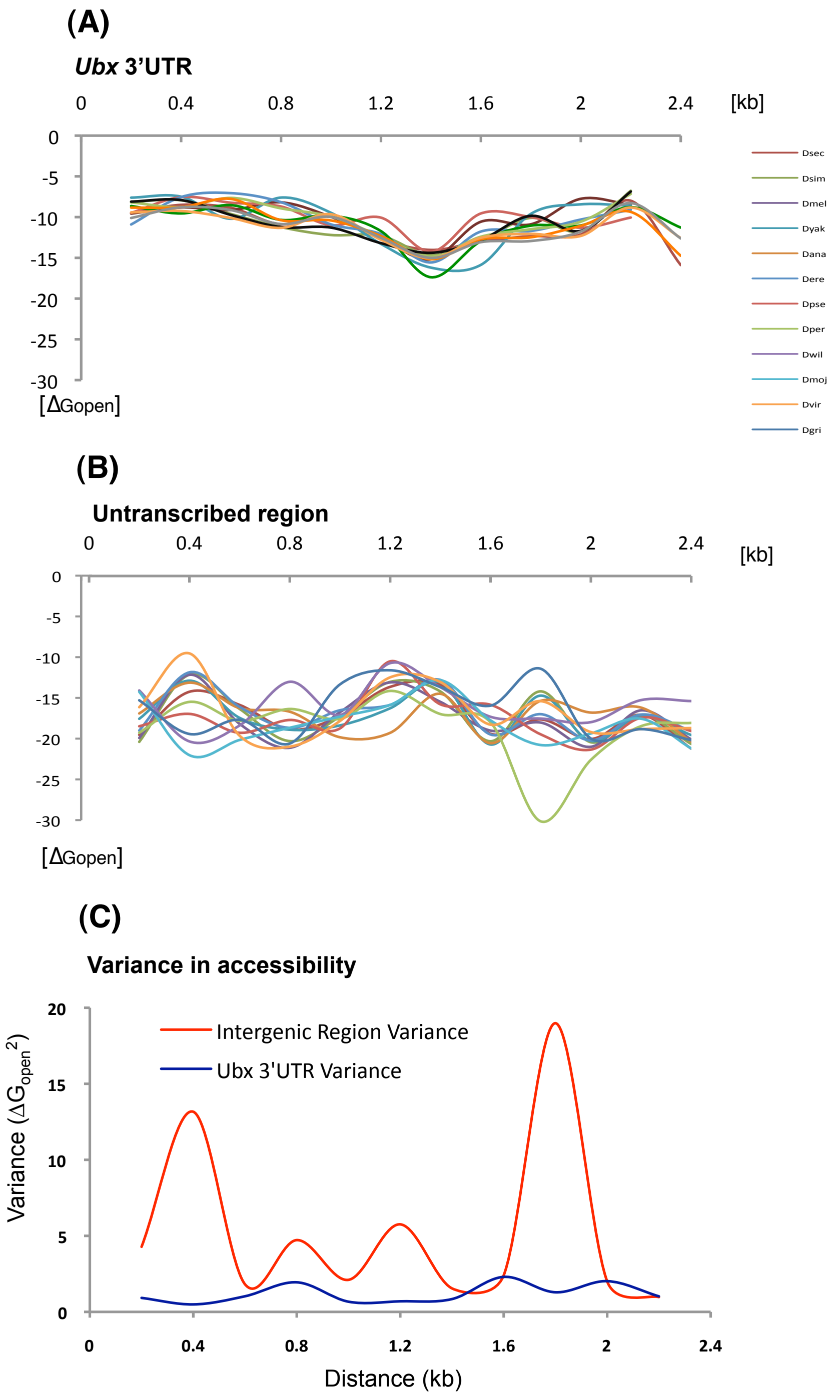
Patraquim & Alonso

Figure 3



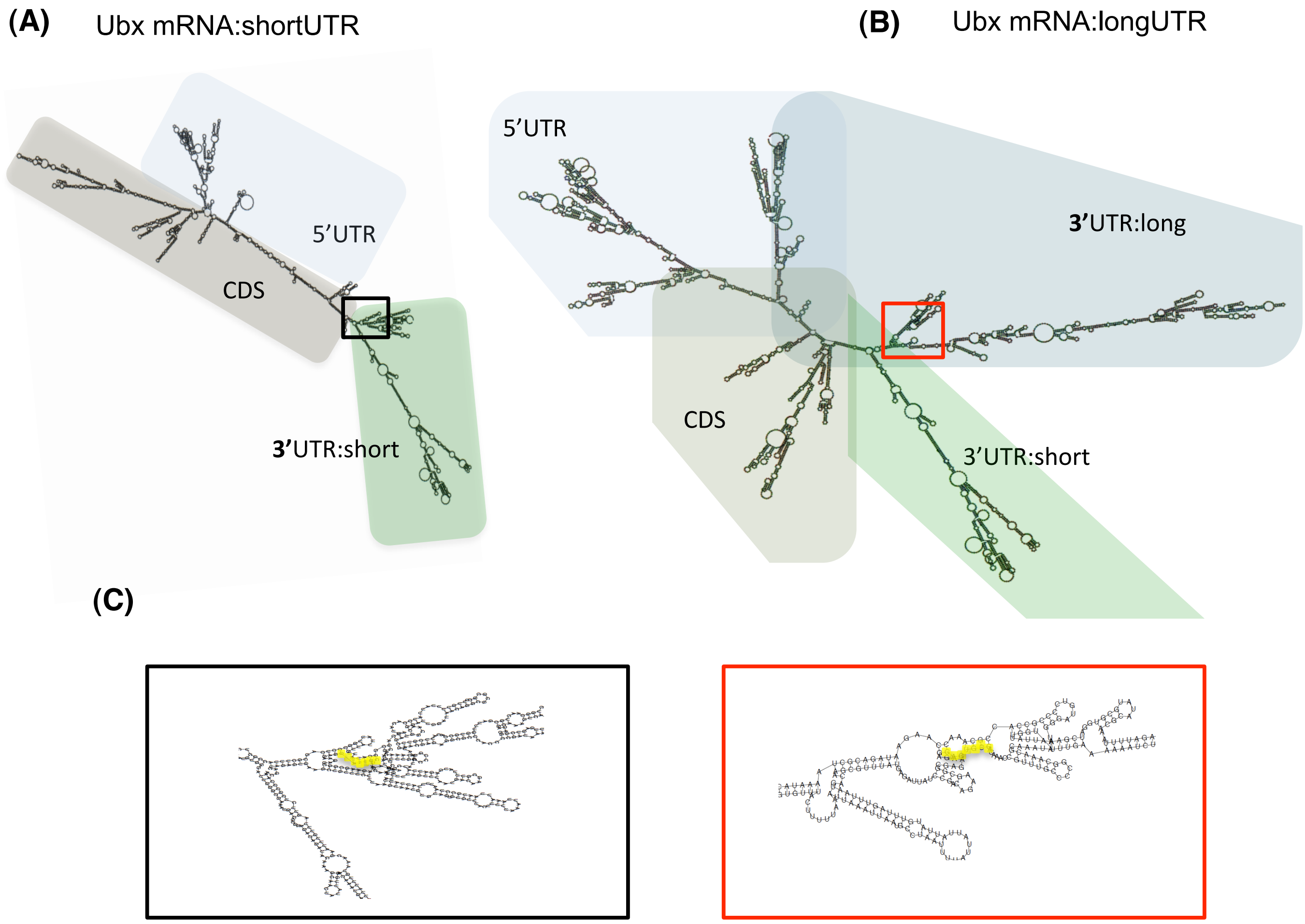
Patraquim & Alonso

Figure 4



Patraquim & Alonso

Figure 5



(D)

microRNA	***Position***	***ΔΔG***	***microRNA***	***Position***	***ΔΔG***
miR-313	**919**	-12.01	**miR-313**	**919**	**-10.37**
miR-92a	919	-10.51	miR-92a	919	-8.87
miR-92b	919	-8.51	miR-92b	919	-6.87
miR-312	**919**	-6.41	**miR-312**	**919**	**-4.77**