Identification of long non-coding RNAs in insects genomes
Fabrice Legeai, Thomas Derrien

To cite this version:
Fabrice Legeai, Thomas Derrien. Identification of long non-coding RNAs in insects genomes. Current Opinion in Insect Science, 2015, 7, pp.37 - 44. 10.1016/j.cois.2015.01.003. hal-01240461

HAL Id: hal-01240461
https://inria.hal.science/hal-01240461
Submitted on 28 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Identification of long non-coding RNAs in insects genomes

Fabrice Legeai¹,² and Thomas Derrien³

The development of high throughput sequencing technologies (HTS) has allowed researchers to better assess the complexity and diversity of the transcriptome. Among the many classes of non-coding RNAs (ncRNAs) identified in the last decade, long non-coding RNAs (lncRNAs) represent a diverse and numerous repertoire of important ncRNAs, reinforcing the view that they are of central importance to the cell machinery in all branches of life. Although lncRNAs have been involved in essential biological processes such as imprinting, gene regulation or dosage compensation especially in mammals, the repertoire of lncRNAs is poorly characterized for many non-model organisms. In this review, we first focus on what is known about experimentally validated lncRNAs in insects and then review bioinformatic methods to annotate lncRNAs in the genomes of hexapods.

Addresses
¹ INRA, UMR1349, Institute of Genetics, Environment and Plant Protection, Domaine de la Motte, BP35237, 35653 Le Rheu cedex, France
² IRIS/INRIA GenScale, Campus Beaulieu, 35000 Rennes, France
³ CNRS, UMR 6290, Institut de Génétique et Développement de Rennes, Université de Rennes 1, 2 Avenue du Pr. Léon Bernard, 35000 Rennes, France

Corresponding author: Legeai, Fabrice (fabrice.legeai@rennes.inra.fr)

Current Opinion in Insect Science 2015, 7:37–44
This review comes from a themed issue on Insect genomics
Edited by Susan Brown and Denis Tagu
For a complete overview see the Issue and the Editorial
Available online 13th January 2015
http://dx.doi.org/10.1016/j.cois.2015.01.003
2214-5745/© 2015 Published by Elsevier Inc.

Introduction

Whole transcriptome sequencing experiments or RNAseq has become very popular as a means to monitor the population of RNAs in cells and to provide a unique snapshot of all transcripts present at a specific time-point in a particular cell type or tissue [1,2]. Beyond the classical messenger RNAs (mRNAs) that will be translated into proteins, RNAseq has shed light on the multiple classes of non-coding RNAs (ncRNAs) players pervasively transcribed from the genome. Recently, particular attention has been paid to the class of long non-coding RNAs (lncRNAs) since they have been connected to various mechanisms such as cis and/or trans regulation of transcription, dosage compensation, imprinting and competing endogenous RNA (see for recent reviews [3⁺,4⁺,5⁺,6⁺]).

lncRNAs are arbitrarily defined as transcripts longer than 200 nucleotides that do not show any protein-coding capability and thus will not be translated into proteins [7]. Because of this vague definition, the catalog of lncRNAs represents a heterogeneous class of transcripts with mRNA-like characteristics, that is, transcribed by RNA polymerase II, 5' capped and often spliced [8]. Similar to the well-studied lncRNA XIST in mammals [9], pioneer work has been done in Drosophila melanogaster through the identification of the ROX1/2 RNA genes involved in dosage compensation [10]. More recently, the modENCODE project annotated thousands of lncRNAs via a deep exploration of the fruitfly transcriptomes [11] reinforcing the view that ncRNAs are essential components to link genotype to phenotype relationships. However, while the repertoire of annotated lncRNAs is regularly improved in phylogenetically distant species [12], many non-model organisms still do not benefit from the annotation of these functional elements of the genomes [13⁺]. In this review, we first compiled available biological information on functionally validated lncRNAs in insect genomes with particular emphasis on lncRNAs in fruitfly and honeybee species and then discussed the method to annotate lncRNAs using RNAseq.

Resources and known functions for lncRNAs in insects

At least 72 active databases are dedicated to collecting biological information about ncRNAs [14] including LNCipedia [15] and IncrNome [16], which are specifically devoted to catalog lncRNAs and describe their functions based on literature. Even if most of them are still limited to human or mouse organisms, more generalist databases such as NonCode [17] aim at gathering all ncRNA sequences produced from experimental protocols or derived from automatic computational scans, with emphasis on a dozen of model organisms, including Drosophila melanogaster. While the fourth version of NonCode contains more than 200 thousands human and mouse lncRNAs, it registers only 3193 lncRNAs from fruitfly, all of them being automatically predicted from RNAseq data (see below) without further functional description. On the other hand, the RNA family database
RFam [18] and IncRNAdb [19], databases, both accessible through the RNACentral repository [20*], comprise a set of experimentally verified ncRNAs which are also searched in other organisms using both sequence and structural similarity strategies. Interestingly, while RNA-Central stores a total of 40,182 IncRNAs only a few concern insect models.

Among them, much attention has been focused on two particular insect species: fruitfly and honeybee. The first because it benefits from outstanding resources for experimental investigation and validations, and the second because of its particular social behaviors and caste polyphenism, the latter processes mainly involve epigenetics [21**]. In Table 1, we briefly describe the few well-studied long ncRNAs for which functions have been characterized to date in fruitfly and honeybee. We show that these IncRNAs could be broadly classified into main biological mechanisms such as development (IncRNA bithorax), behavior (sphinx and NB-1) or neural

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description of the known fruitfly and honeybee IncRNAs which have been experimentally validated and annotated in specific databases (RFam, IncRNAdb or FlyBase).</td>
</tr>
<tr>
<td>Gene regulation</td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>hsr-(\omega)</td>
</tr>
<tr>
<td>Epigenetics control of genes regulation</td>
</tr>
<tr>
<td>rox1/2</td>
</tr>
<tr>
<td>Development</td>
</tr>
<tr>
<td>bithorax</td>
</tr>
<tr>
<td>Lnccov1/2</td>
</tr>
<tr>
<td>Behavior</td>
</tr>
<tr>
<td>yar</td>
</tr>
<tr>
<td>sphinx</td>
</tr>
<tr>
<td>Nb-1</td>
</tr>
<tr>
<td>Neural expression</td>
</tr>
<tr>
<td>Ks-1</td>
</tr>
<tr>
<td>AncR-1</td>
</tr>
<tr>
<td>Kakusei</td>
</tr>
</tbody>
</table>
expression (kakusei). In these species, the mode of action of the lncRNAs involves either broad trans-regulation by epigenetic control of chromatin (e.g.) organization or RNA sequestration in a nuclear compartment (such as omega-speckles), as well as neighboring cis-regulation of specific mRNA genes, both during development or following a stress.

In other insect model organisms, such as *Acrithosiphon pismum*, *Aedes gambiae*, *Anopheles gambiae*, *Danaus plexippus* or *Heliconius melpomene* no in-depth functional analysis of putative lncRNA have been published to date. But, interestingly in the red flour beetle *Tribolium castaneum*, numerous ncRNAs are expressed on the opposite strand of protein-coding genes localized in the Hox cluster [36]. Similarly, Li *et al.* recently observed that some ncRNAs of intermediate sizes are transcribed from the silk gland of *Bombyx mori* and may be involved in the repression of transcription by epigenetic modifications of histones [37]. Finally, in *Nasonia vitripennis*, where some individuals contain a paternally transmitted supernumerary chromosome (Paternal Sex Ratio, PSR), the paternal chromatin is modified during the first mitotic division via retention of histone H3 in a phosphorylated state. This first exhaustive transcriptome study in testis tissue revealed the presence of four putative ncRNAs that are specific to individuals having a PSR [38].

In contrast to short ncRNAs such as tRNAs or miRNAs, neither the sequence nor the structure of the lncRNA appear to be phylogenetically conserved throughout the metazoan kingdom [39] or even among the insects even when the biological processes involving lncRNA functions are similar between species. As a result, new specialized computational prediction protocols and tools are being developed to discriminate coding versus non-coding transcripts and to refine the functional annotation of lncRNAs (see next section for details). Thanks to the recent advances in sequencing technologies (RNA-Seq), the systematic identification of long ncRNAs has already been applied to few insects having high-quality genome assemblies in order to complete the repertoire of functional elements in their genomes (Table 2).

### Bioinformatics workflows for the systematic identification and annotation of lncRNAs

In order to annotate lncRNAs, a typical workflow (Figure 1) could be applied to the growing number of assembled insect genomes with particular attention on the following three key points.

#### RNASeq protocols

Whole transcriptome sequencing (RNASeq) represents the method of choice to discover new transcripts and also to quantify all RNAs in a variety of organisms, cell types

### Table 2

**Computational identification of long non-coding RNAs in fruitfly and mosquito genomes.** All the methods described here used a dedicated pipeline with different biological material as input (column material) and different protocols and scoring scheme (columns structural annotation and coding potential analysis).

<table>
<thead>
<tr>
<th>Reference</th>
<th>Species</th>
<th>Material</th>
<th>Structural annotation</th>
<th>Coding potential analysis</th>
<th>Number of identified lncRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tupy [40]</td>
<td><em>Drosophila melanogaster</em></td>
<td>7972 cDNA</td>
<td>No intersection with existing annotations</td>
<td>KA/Ks, ORF and no conserved codon structure</td>
<td>72</td>
</tr>
<tr>
<td>Hiller [41]</td>
<td><em>Drosophila melanogaster</em></td>
<td>Donor and acceptor sites identified by intronscan</td>
<td>Evolutionary signature of conserved introns structure between 15 insect genomes</td>
<td>N/A</td>
<td>129 (partial)</td>
</tr>
<tr>
<td>Young [42]</td>
<td><em>Drosophila melanogaster</em></td>
<td>RNA-Seq from modENCODE project (60 developmental points)</td>
<td>Reads mapping (TopHat), annotation (Cufflinks), comparison with existing annotation (CuffCompare)</td>
<td>Coding potential calculation (CPAT), and analysis of conservation (PhyloCSF)</td>
<td>1119</td>
</tr>
<tr>
<td>Brown [11]</td>
<td><em>Drosophila melanogaster</em></td>
<td>12.9 billions of paired-end single strand polyA+ RNASeq reads from 74 libraries (inc. distinct tissues, whole body libraries and cell lines)</td>
<td>Reads mapping (TopHat), annotation (GRIT), comparison to Flybase annotation</td>
<td>Analysis of conservation (PhyloCSF)</td>
<td>1875 (3085 transcripts)</td>
</tr>
<tr>
<td>Padrón [43]</td>
<td><em>Anopheles gambiae</em></td>
<td>RNA-Seq (223 millions reads)</td>
<td>Reads mapping (TopHat), annotation (Cufflinks), comparison with existing annotation (CuffCompare)</td>
<td>Coding potential calculation (CPAT)</td>
<td>9863</td>
</tr>
<tr>
<td>Jenkins [44]</td>
<td><em>Anopheles gambiae</em></td>
<td>More than 500 million alignable reads</td>
<td>Reads mapping (TopHat), annotation (Cufflinks), comparison with existing annotation (CuffCompare)</td>
<td>Size of the ORF and CDS, analysis of conservation (PhyloCSF), no recognizable protein domains</td>
<td>633</td>
</tr>
</tbody>
</table>
or tissues [1]. Even if RNAseq technology has become the ‘de facto’ standard for transcriptome profiling, it is worth noting that this technology undergoes rapid evolution in terms of library preparation, sequencing platforms and subsequent bioinformatics analyses leading to regular updates of guidelines and standards [45].
Given that first, IncRNAs tend to be rare compared to mRNAs and second, display both spatial and temporal expression patterns [8], the depth of sequencing, the number of different tissues/cell lines and time-points need to be considered in planning each experiment. For instance, it has been shown in human and fruitfly that testis tissue shows the highest number of tissue-specific IncRNAs (e.g. 11% of all IncRNAs in Drosophila) probably reflecting more relaxed chromatin structure [11]. In addition, many IncRNAs, are expressed antisense to protein-coding genes [46,47], which they often regulate [48]. It is thus recommended to favor stranded RNASeq protocols (also known as directional transcriptome sequencing) that keep track of the strand of origin of the transcript [49] if the purpose of the study includes the identification of antisense IncRNAs. For example, using these stranded protocols, the modENCODE project recently discovered 402 IncRNA loci (21% of all IncRNA loci) located antisense to mRNA transcripts of protein coding genes in D. melanogaster [11] while this proportion is slightly lower (~15%) in the human genome [8].

**Mapping and reconstructing IncRNAs**

When a high quality reference genome is available, one would favor a map-first then assembly strategy. To this end, the common but still critical point in RNASeq analysis consists of mapping the millions of sequenced reads onto a reference genome. Fortunately, several software solutions or mapping programs have been developed in the last few years that efficiently and rapidly align reads on a reference genome [50–53]. Regarding RNASeq reads, the task is even more challenging since the mapper should also handle spliced-read alignments, that is, reads mapping over exon/intron junctions (see for benchmark [54]). Since many of the IncRNAs are multiexonic, the use of spliced-read aligners is critical to precisely discover novel exon junctions as for IncRNAs. In D. melanogaster, Young et al. used a more stringent mapping protocol by forcing reads mapping onto exon junctions connecting mRNAs ends and known adjacent intergenic transcripts [42**] in order to exclude possible false positive intergenic IncRNAs [55]. The resulting mapping file is then used by graph-based approaches such as cufflinks [56], scripture [57] or FluxSimulator [58] to reconstruct all transcript models. Despite intensive works in this area, these methods are far from being perfect in terms of accuracy, sensitivity and specificity but seem to behave better for smaller genomes (nematode and fruitfly) compared to mammalian genomes [59].

**Measuring the coding potential**

Once a set of transcripts has been assembled, it is important to estimate the protein-coding potential of the sequences. After having removed transcripts whose size is below a certain cutoff (often 200 nt for IncRNA), a first step could be the filtering of transcripts which overlap mRNAs exons in the sense orientation as they more likely correspond to novel isoforms of a protein-coding gene. Subsequently, two kinds of methods are available to classify coding versus non-coding transcripts using computational programs and/or biochemical experiments. Computationally, the coding potential could be determined by measuring the intrinsic properties of the sequences which correspond (non-exhaustively) to first, the length of the Open Reading Frame (ORF), second, the coverage of the ORF compared to the length of the transcripts, third, the bias in k-mer frequencies between coding and non-coding sequences and fourth, the presence of protein-coding specific motifs [60,61]. All these features are often integrated into machine learning algorithms such as a support vector machine (SVM) or random forest (RF), which are trained with known sets of protein-coding and non-coding transcripts. Additionally, some other tools also require the alignment of the candidate IncRNAs sequences with protein databases (such as PFAM or Swiss-Prot) to search for evidence of translatability [62] or with multiple genomes in order to specifically tag the selective pressure acting on mRNAs [63]. However, the former method suffers from the inherent lack of protein sequences (especially for insect proteomes) and may therefore lead to false positive annotation of insect IncRNAs while the latter implies that IncRNAs are evolutionary conserved which may not be the case with respect to both the phylogenetic distances [39] and effective population sizes [64**]. Finally, further work is needed to develop programs that simulate non-coding sequences in the absence of non-coding training sets.

At the experimental level, evidence for protein-coding capability can be directly obtained by mass spectrometry data [65] although these experiments may not be available for all insect species. Another complementary technique, ribosome profiling, was developed by Ingolia et al. and utilizes high throughput sequencing to map RNA regions associated with translating ribosomes [66]. While some IncRNAs have been shown to be associated with ribosomes suggesting that they are in fact wrongly annotated, it is still unclear whether the resulting short peptides are really functional since they are also reported in the 5’UTR of protein-coding transcripts.

**Conclusion and future directions**

Despite growing evidence that IncRNAs are key players in mammalian cells, only a few of them have been experimentally validated in insects, mostly in D. melanogaster. As a proof of concept for IncRNA functionality, the well-studied RNA rox genes are involved in dosage compensation similar to the Xist gene in mammals [9].

Many new insect IncRNAs will be discovered in the next few years owing to both the availability of cheaper RNASeq protocols and the development of dedicated bioinformatics programs. For instance, when a reference genome is not available or when the quality of the genome assembly is relatively poor (which may be the
case for several non-model organisms), de novo or genome-independent assembly approaches [67–69] have to be envisaged to better identify the lncRNAs repertoire.

Moreover, differentiating non-coding from coding transcripts remains a challenging task [70] and can be hampered by biological artifices. Indeed, at least two Drosophila lncRNAs (pro and polished rice) have been recently re-classified as coding for small peptides [71,72]. Furthermore, bioinformatics tools are still missing that distinguish bifunctional RNAs such as the steroid receptor activator gene (SRA) [73] harboring two different functions, one at the RNA level and another when translated into proteins.

Finally, it is obvious that new experimental methods have to be implemented to understand the function of these intriguing RNAs, such as the recent and promising technologies CHART [74] or ChIRP [75], that can be applied to identify the DNA binding sites of lncRNAs. In parallel, computational approaches are required to unveil functions of lncRNAs on a large-scale perspective. As illustrated by the recent implementation of IncRNAtor [76], a new database integrating functional information about lncRNAs from six species, including fruitfly. These attempts have to be extended to non-model organisms in order to shed light on the many components of the genome (coding and non-coding) that are responsible for phenotypic traits.

Acknowledgements
The authors wish to thank Sue Brown and Denis Tagu for their comments and corrections on the manuscript.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:
● of special interest
◆ of outstanding interest


An interesting claim for improving data and bioinformatics resources to increase our knowledge on non-model organisms.


This database is one of the most comprehensive resources for IncRNA gene annotation in multiple organisms including many insects.


A relevant analysis of what are the future directions to better understand the impact of epigenetics in the context of polyphenism and polyethism of eusocial insects.


23. Lakhotia SC, Mallik M, Singh AK, Ray M: The large noncoding hsnu-r transcripts are essential for thermotolerance and remodeling of mRNPs, HP1 and RNA polymerase II during recovery from heat shock in Drosophila. Chromosoma 2012, 121:49–70.


Identification of long non-coding RNAs in insects genomes


This paper describes a stringent methodology to annotate and fully characterize one of the first catalogs of long intergenic ncRNAs in fruitfly.


A first application of systematic prediction of lncRNAs in an insect species apart from Drosophila melanogaster.


This paper describes a fast and effective program called CPAT which is able to discriminate between coding and non-coding sequences using a logistic regression model built on a set of known coding and non-coding transcripts.


64. Haerty W, Ponting CP: Mutations within IncRNAs are effectively selected against in fruitfly but not in human. Genome Biol 2013, 14:R49.
By comparing sequence polymorphism in Drosophila and human populations, this paper demonstrates the deleterious effect of mutations within fruitfly IncRNAs but not in human IncRNAs probably owing to the difference in effective population sizes between the two species.
This article presents domain-specific chromatin isolation by RNA purification (dChIP), a technique to analyze RNA–RNA, RNA–protein and RNA–chromatin interactions used to precisely characterize the chromosomal binding of roX1 and roX2.
A new database integrating functional information about IncRNAs from six species, including fruitfly.