Identification of long non-coding RNAs in insects genomes

Fabrice Legeai, Thomas Derrien

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Insect Science

Identification of long non-coding RNAs in insects genomes

Fabrice Legeai[1,2] and Thomas Derrien[3]

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 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, BP35327, 35653 Le Rheu cedex, France
[2] IRISA/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

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 endogeneous 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 lncRNome [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 lncRNAdb [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 RNACentral stores a total of 40,182 lncRNAs 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 lncRNAs could be broadly classified into main biological mechanisms such as development (lncRNA bithorax), behavior (sphinx and NB-1) or neural

Table 1

Description of the known fruitfly and honeybee lncRNAs which have been experimentally validated and annotated in specific databases (Rfam, lncRNAdb or FlyBase).

Gene	Species	Function	Mode
Gene regulation			
hsr-ω	*Drosophila melanogaster*	Long non-coding RNAs produced by the hsw-ω gene are actively expressed in nuclei, forming spots called perinuclear omega-speckles, in response to heat shock stress. These speckles are involved in the redistribution and sequestration of multiple processing proteins, in particular heterogeneous nuclear ribonucleoproteins (hnRNPs), HP1 or polII, which strongly affect multiple cellular networks subsequent to a stress [22,23]	*Trans*
Epigenetics control of genes regulation			
rox1/2	*Drosophila melanogaster*	In X0 male *Drosophila*, the transcription of genes located on the X chromosome is increased relative to the level of XX females by a mechanism known as dosage compensation. This mechanism is connected to the acetylation of histone H4 at lysine 16 induced by a protein complex which involves two long non-coding RNAs roX1 and roX2 [24,25]	*Trans*
Development			
bithorax	*Drosophila melanogaster*	The bithorax complex (BX-C) plays a key role in *Drosophila* development and covers a region larger than 300 kb that includes only four protein-coding genes. Non-coding genes have already been shown to be concomitantly expressed from the BX-C domain to regulate in *cis* the BX-C proteins in specific abdominal segments [26,27]	*Cis*
Lnccov1/2	*Apis mellifera*	These two transcripts lack evidence of functional ORFs and are differentially expressed in queen and worker ovariole transcriptomes at the embryonic stage. Temporal expression shows that lnccov1 might be involved in the autophagic cell death of ovarioles during worker embryogenesis, and fluorescent *in situ* hybridization (FISH) indicates perinuclear localization in omega speckle-like structures [28]	Unknown
Behavior			
yar	*Drosophila melanogaster*	yar is a long non-coding RNA, highly expressed during embryogenesis and located in a neural gene cluster between yellow (y) and achaete (ac) [29]. With the help of mutants, the function of this long non-coding RNA has been recently refined as a regulator of y and ac transcription, affecting as well the sleep behavior of the *Drosophila* [30]	*Cis*
sphinx	*Drosophila melanogaster*	sphinx is a lncRNA involved in the regulation of male courtship behavior. Sequence variations between close *Drosophila* species advocate for a functional role and a rapid adaptation. The mutagenesis of sphinx in *Drosophila* reveals the emergence of male-male courtship behavior, probably by the disruption of some sensory circuits, which is supported by its specific expression in chemosensory organs [31]	Unknown
Nb-1	*Apis mellifera*	Nb-1 is a 700 nt transcript, whose longest ORF encodes a putative 32 amino acid without any sequence conservation. This lncRNA is expressed in the honeybee brain with variation according to the age of the colony workers testifying to its putative role in polyethism [32]	Unknown
Neural expression			
Ks-1	*Apis mellifera*	Sawata *et al*. identified a 17knt transcript that is expressed restrictively in the mushroom body of Kenyon cells in the honeybee brain and which accumulates in the nucleus. The transcript exhibits seven putative ORFs longer than 67 amino acids without any conservation in a related species (Apis cerana) nor similarity with known proteins [33]	Unknown
AncR-1	*Apis mellifera*	AncR-1 is preferentially expressed in the brain, in sexual tissues and in some secretory organs and accumulates in nuclei. It contains multiple alternate isoforms, which are derived from a 6.9 kbp genomic locus [34]	Unknown
Kakusei	*Apis mellifera*	The kakusei is a ~7000 nt long non-coding RNA with multiple constitutive and inducible variants, the expression of which is transiently up-regulated by neural activity. It is localized exclusively in neural nuclei in discrete nuclear compartments. This gene may play specific roles in RNA metabolism in the honeybee brains, irrespective of behavioral experience [35]	Unknown

expression (kakusei). In these species, the mode of action of the lncRNAs involves either broad trans-regulation by epigenetic control of chromatin (re)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 *Acyrthosiphon pisum*, *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 metazoa 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

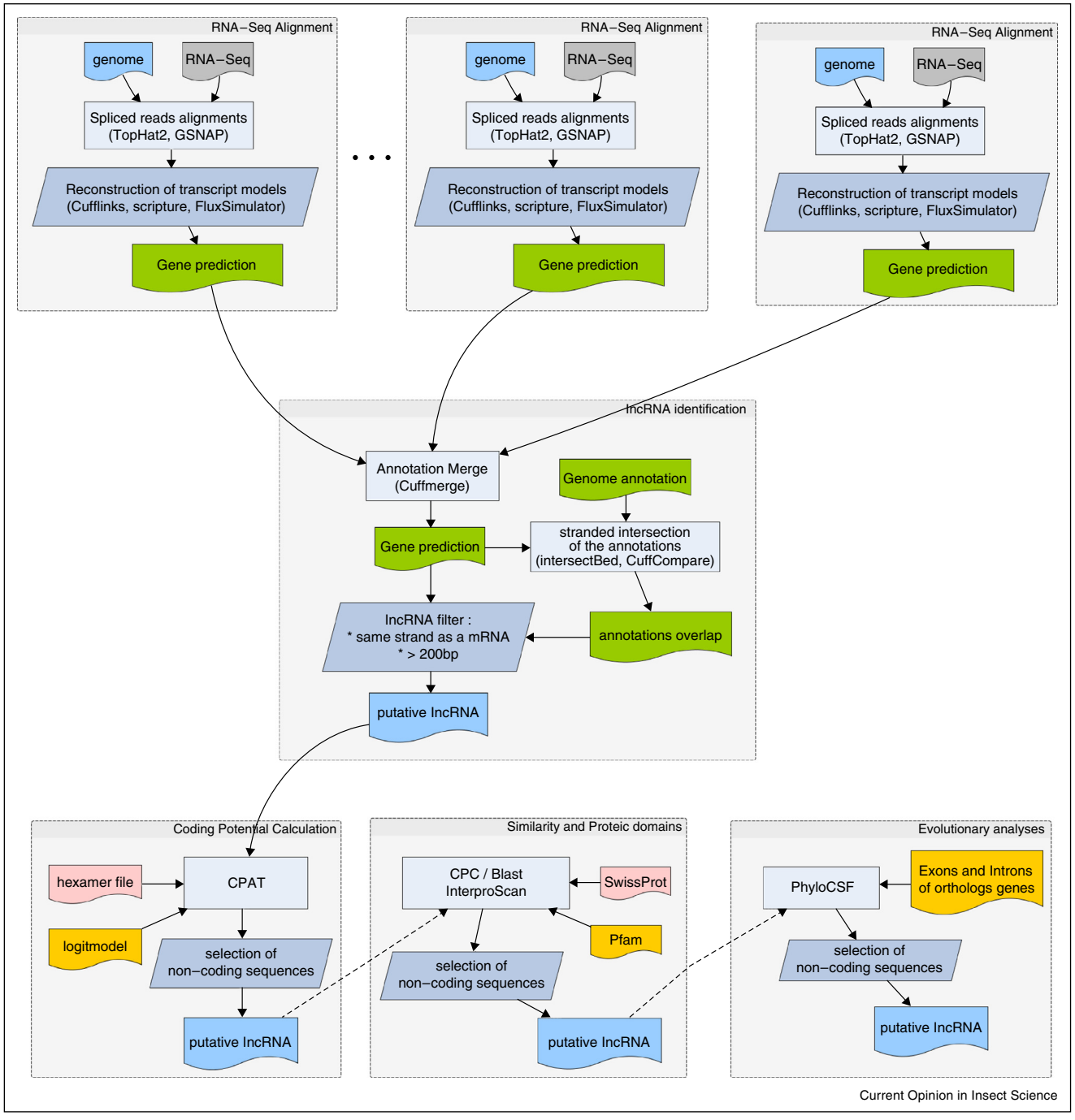
Whole transcriptome sequencing (RNASeq) represents the method of choice to discover new transcripts and also to quantify all RNAs in a variety of organisms, cell types

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).

Reference	Species	Material	Structural annotation	Coding potential analysis	Number of identified lncRNA
Tupy [40]	*Drosophila melanogaster*	7972 cDNA	No intersection with existing annotations	KA/Ks, QRNA and no conserved codon structure	72
Hiller [41]	*Drosophila melanogaster*	Donor and acceptor sites identified by intronscan	Evolutionary signature of conserved introns structure between 15 insect genomes	N/A	129 (partial)
Young [42••]	*Drosophila melanogaster*	RNA-Seq from modENCODE project (30 developmental points)	Reads mapping (TopHat), annotation (Cufflinks), comparison with existing annotation (CuffCompare)	Coding potential calculation (CPC), and analysis of conservation (PhyloCSF)	1119
Brown [11]	*Drosophila melanogaster*	12.9 billions of pairs-ends single strand polyA+ RNASeq reads from 74 libraries (inc. distinct tissues, whole body libraries and cell lines)	Reads mapping (TopHat), annotation GRIT), comparison to Flybase annotation	Analysis of conservation (PhyloCSF)	1875 (3085 transcripts)
Padrón [43•]	*Anopheles gambiae*	RNA-Seq (223 millions reads)	Reads mapping (TopHat), annotation (Cufflinks), comparison with existing annotation (CuffCompare)	Coding potential calculation (CPAT)	9863
Jenkins [44]	*Anopheles gambiae*	More than 500 million alignable reads	Reads mapping (TopHat), annotation (Cufflinks), comparison with existing annotation (CuffCompare)	Size of the ORF and CDS, analysis of conservation (phyloCSF), no recognizable protein domains	633

Figure 1



Schematic view of standardized workflow. Steps 1–4 are designed to eliminate transcripts from known and potential protein coding genes. Step 5 is designed to identify known lncRNAs. (1) Align RNASeq reads using a splice-aware algorithm and predict genes. This step can be executed in parallel for each library. (2) Merge annotated transcripts from step 1 and compare to known protein coding genes. (3) Predict coding potential of ORFs in remaining genes. (4) Compare ORF predictions with known proteins and protein domains. (5) Align remaining transcript with those from closely related orthologs to identify signals of selection.

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, lncRNAs 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 lncRNAs (e.g. 11% of all lncRNAs in *Drosophila*) probably reflecting more relaxed chromatin structure [11]. In addition, many lncRNAs, 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 lncRNAs. For example, using these stranded protocols, the modENCODE project recently discovered 402 lncRNA loci (21% of all lncRNA 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 lncRNAs

When a high quality reference genome is available, one would favor a map-first then assemble 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 mappers 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 lncRNAs are multiexonic, the use of spliced-read aligners is critical to precisely discover novel exon junctions as for lncRNAs. In *D. melanogaster*, Young et al. used a more stringent mapping protocol by forcing read mapping onto exon junctions connecting mRNAs ends and known adjacent intergenic transcripts [42••] in order to exclude possible false positive intergenic lncRNAs [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 lncRNA), 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 lncRNAs 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 lncRNAs while the latter implies that lncRNAs 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 lncRNAs 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 lncRNAs 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 lncRNA functionality, the well-studied RNA rox genes are involved in dosage compensation similar to the Xist gene in mammals [9].

Many new insect lncRNAs 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 (prg 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 dChIRP [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 lncRNAtor [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

1. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57-63.

2. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F *et al.*: **Landscape of transcription in human cells**. *Nature* 2012, **488**:101-108.

3. •• Ulitsky I, Bartel DP: **lincRNAs: genomics, evolution, and mechanisms**. *Cell* 2013, **154**:26-46.

4. •• Fatica A, Bozzoni I: **Long non-coding RNAs: new players in cell differentiation and development**. *Nat Rev Genet* 2013, **15**:7-21.

5. •• Gardini A, Shiekhattar R.: **The many faces of long noncoding RNAs**. *FEBS J* 2014 http://dx.doi.org/10.1111/febs.13101.

6. •• Bonasio R, Shiekhattar R.: **Regulation of transcription by long noncoding RNAs**. *Annu Rev Genet* 2014, **48**:433-455.

These four articles are excellent recent reviews covering all the mechanisms involving long non-coding RNAs that has been described to date.

7. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermüller J, Hofacker IL *et al.*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription**. *Science* 2007, **316**:1484-1488.

8. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG *et al.*: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome Res* 2012, **22**:1775-1789.

9. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, Swift S, Rastan S: **The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus**. *Cell* 1992, **71**:515-526.

10. Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL: **roX1 RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system**. *Cell* 1997, **88**:445-457.

11. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM *et al.*: **Diversity and dynamics of the *Drosophila* transcriptome**. *Nature* 2014 http://dx.doi.org/10.1038/nature12962.

12. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H: **The evolution of lncRNA repertoires and expression patterns in tetrapods**. *Nature* 2014, **505**:635-640.

13. • Tagu D, Colbourne JK, Negre N: **Genomic data integration for ecological and evolutionary traits in non-model organisms**. *BMC Genomics* 2014, **15**:1-16.

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

14. Hoeppner MP, Barquist LE, Gardner PP: **An introduction to RNA databases**. *Methods Mol Biol* 2014, **1097**:107-123.

15. Volders P-J, Verheggen K, Menschaert G, Vandepoele K, Martens L, Vandesompele J, Mestdagh P: **An update on LNCipedia: a database for annotated human lncRNA sequences**. *Nucleic Acids Res* 2014 http://dx.doi.org/10.1093/nar/gku1060.

16. Bhartiya D, Pal K, Ghosh S, Kapoor S, Jalali S, Panwar B, Jain S, Sati S, Sengupta S, Sachidanandan C *et al.*: **lncRNome: a comprehensive knowledgebase of human long noncoding RNAs**. *Database (Oxford)* 2013, **2013**:bat034.

17. Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y: **NONCODEv4: exploring the world of long non-coding RNA genes**. *Nucleic Acids Res* 2014, **42**:D98-D103.

18. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A: **Rfam 11.0: 10 years of RNA families**. *Nucleic Acids Res* 2013, **41**:D226-D232.

19. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS: **lncRNAdb: a reference database for long noncoding RNAs**. *Nucleic Acids Res* 2011, **39**:D146-D151.

20. • The RNA Central Consortium: **RNAcentral: an international database of ncRNA sequences**. *Nucleic Acids Res* 2014 http://dx.doi.org/10.1093/nar/gku991.

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

21. •• Bonasio R: **Emerging topics in epigenetics: ants, brains, and noncoding RNAs**. *Ann N Y Acad Sci* 2012, **1260**:14-23.

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.

22. Lakhotia SC: **Forty years of the 93D puff of *Drosophila melanogaster***. *J Biosci* 2011, **36**:399-423.

23. Lakhotia SC, Mallik M, Singh AK, Ray M: **The large noncoding hsrω-n transcripts are essential for thermotolerance and remobilization of hnRNPs, HP1 and RNA polymerase II during recovery from heat shock in *Drosophila***. *Chromosoma* 2012, **121**:49-70.

24. Smith ER, Allis CD, Lucchesi JC: **Linking global histone acetylation to the transcription enhancement of X-chromosomal genes in *Drosophila* males**. *J Biol Chem* 2001, **276**:31483-31486.

25. Deng X, Meller VH: **roX RNAs are required for increased expression of X-linked genes in *Drosophila melanogaster* males**. *Genetics* 2006, **174**:1859-1866.

26. Lipshitz HD, Peattie DA, Hogness DS: **Novel transcripts from the Ultrabithorax domain of the bithorax complex**. *Genes Dev* 1987, **1**:307-322.

27. Pease B, Borges AC, Bender W: **Noncoding RNAs of the Ultrabithorax domain of the *Drosophila* bithorax complex**. *Genetics* 2013, **195**:1253-1264.

28. Humann FC, Hartfelder K: **Representational difference analysis (RDA) reveals differential expression of conserved as well as novel genes during caste-specific development of the honey bee (*Apis mellifera* L.) ovary**. *Insect Biochem Mol Biol* 2011, **41**:602-612.

29. Soshnev AA, Li X, Wehling MD, Geyer PK: **Context differences reveal insulator and activator functions of a Su(Hw) binding region**. *PLoS Genet* 2008, **4**:e1000159.

30. Soshnev AA, Ishimoto H, McAllister BF, Li X, Wehling MD, Kitamoto T, Geyer PK: **A conserved long noncoding RNA affects sleep behavior in *Drosophila***. *Genetics* 2011, **189**:455-468.

31. Chen Y, Dai H, Chen S, Zhang L, Long M: **Highly tissue specific expression of Sphinx supports its male courtship related role in *Drosophila melanogaster***. *PLoS ONE* 2011, **6**:e18853.

32. Tadano H, Yamazaki Y, Takeuchi H, Kubo T: **Age- and division-of-labour-dependent differential expression of a novel non-coding RNA, Nb-1, in the brain of worker honeybees, *Apis mellifera* L.**. *Insect Mol Biol* 2009, **18**:715-726.

33. Sawata M, Daisuke Y, Takeuchi H, Kamikouchi A, Kazuaki O, Kubo T: **Identification and punctate nuclear localization of a novel noncoding Identification and punctate nuclear localization of a novel noncoding RNA, Ks-1, from the honeybee brain**. *RNA* 2002:772-785.

34. Sawata M, Takeuchi H, Kubo T: **Identification and analysis of the minimal promoter activity of a novel noncoding nuclear RNA gene, AncR-1, from the honeybee (*Apis mellifera* L.)**. *RNA* 2004 http://dx.doi.org/10.1261/rna.5231504.use.

35. Kiya T, Kunieda T, Kubo T: **Inducible- and constitutive-type transcript variants of kakusei, a novel non-coding immediate early gene, in the honeybee brain**. *Insect Mol Biol* 2008, **17**:531-536.

36. Shippy TD, Ronshaugen M, Cande J, He J, Beeman RW, Levine M, Brown SJ, Denell RE: **Analysis of the *Tribolium homeotic* complex: insights into mechanisms constraining insect Hox clusters**. *Dev Genes Evol* 2008, **218**:127-139.

37. Li D-D, Liu Z-C, Huang L, Jiang Q-L, Zhang K, Qiao H-L, Jiao Z-J, Yao L-G, Liu R-Y, Kan Y-C: **The expression analysis of silk gland-enriched intermediate-size non-coding RNAs in silkworm *Bombyx mori***. *Insect Sci* 2014, **21**:429-438.

38. Akbari OS, Antoshechkin I, Hay BA, Ferree PM: **Transcriptome profiling of Nasonia vitripennis testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio**. *G3 (Bethesda, MD)* 2013, **3**:1597-1605.

39. Kapusta A, Feschotte C: **Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications**. *Trends Genet* 2014, **30**:439-452.

40. Tupy JL, Bailey AM, Dailey G, Evans-holm M, Siebel CW, Misra S, Celniker SE, Rubin GM: **Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster***. *Proc Natl Acad Sci USA* 2005, **102**:5495-5500.

41. Hiller M, Findeiss S, Lein S, Marz M, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G *et al.*: **Conserved introns reveal novel transcripts in *Drosophila melanogaster***. *Genome Res* 2009, **19**:1289-1300.

42. •• Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu J-L, Ponting CP: **Identification and properties of 1119 candidate lincRNA loci in the *Drosophila melanogaster* genome**. *Genome Biol Evol* 2012, **4**:427-442.

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

43. • Padrón A, Molina-Cruz A, Quinones M, Ribeiro JEM, Ramphul U, Rodrigues J, Shen K, Haile A, Ramirez JEL, Barillas-Mury C: **In depth annotation of the *Anopheles gambiae* mosquito midgut transcriptome**. *BMC Genomics* 2014, **15**:636.

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

44. Jenkins AM, Waterhouse RM, Kopin AS: **Long non-coding RNA discovery in *Anopheles gambiae* using deep RNA sequencing**. *bioRxiv* 2014 http://biorxiv.org/content/early/2014/07/26/007484.

45. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinform* 2010, **11**:94.

46. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW: **The antisense transcriptomes of human cells**. *Science* 2008, **322**:1855-1857.

47. Magistri M, Faghihi MA, St Laurent G, Wahlestedt C: **Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts**. *Trends Genet* 2012, **28**:389-396.

48. Johnsson P, Ackley A, Vidarsdottir L, Lui W-O, Corcoran M, Grandér D, Morris KV: **A pseudogene long-noncoding-RNA network regulates PTEN transcription and translation in human cells**. *Nat Struct Mol Biol* 2013, **20**:440-446.

49. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: **Comprehensive comparative analysis of strand-specific RNA sequencing methods**. *Nat Methods* 2010, **7**:709-715.

50. Marco-Sola S, Sammeth M, oacute RG, Ribeca P: **The GEM mapper: fast, accurate and versatile alignment by filtration**. *Nat Methods* 2012 http://dx.doi.org/10.1038/nmeth.2221.

51. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**:1105-1111.

52. Li H, Durbin R: **Fast and accurate short read alignment with Burrows–Wheeler transform**. *Bioinformatics* 2009, **25**:1754-1760.

53. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics* 2012, **29**:15-21.

54. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Alioto T, Behr J, Bertone P, Bohnert R, Campagna D *et al.*: **Systematic evaluation of spliced alignment programs for RNA-seq data**. *Nat Methods* 2013 http://dx.doi.org/10.1038/nmeth.2722.

55. Van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most "dark matter" transcripts are associated with known genes**. *PLoS Biol* 2010, **8**:e1000371.

56. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**:511-515.

57. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al.*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs**. *Nat Biotechnol* 2010, **28**:503-510.

58. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigo R, Sammeth M: **Modelling and simulating generic RNA-Seq experiments with the flux simulator**. *Nucleic Acids Res* 2012, **40**:10073-10083.

59. Steijger T, Abril JF, Engström PG, Kokocinski F, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J *et al.*: **Assessment of transcript reconstruction methods for RNA-seq**. *Nat Methods* 2013 http://dx.doi.org/10.1038/nmeth.2714.

60. Wang L, Park HJ, Dasari S, Wang S, Kocher JP, Li W: **CPAT: coding-potential assessment tool using an alignment-free logistic regression model**. *Nucleic Acids Res* 2013 http://dx.doi.org/10.1093/nar/gkt006.

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

61. Li A, Zhang J, Zhou Z: **PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme**. *BMC Bioinform* 2014, **15**:1-10.

62. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G: **CPC: assess the protein-coding potential of transcripts using**

sequence features and support vector machine. *Nucleic Acids Res* 2007, **35**:W345-W349.

63. Lin MF, Jungreis I, Kellis M: **PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions**. *Bioinformatics* 2011, **27**:i275-i282.

64. •• Haerty W, Ponting CP: **Mutations within lncRNAs 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 lncRNAs but not in human lncRNAs probably owing to the difference in effective population sizes between the two species.

65. Bánfai B, Jia H, Khatun J, Wood E, Risk B, Gundling WE, Kundaje A, Gunawardena HP, Yu Y, Xie L *et al.*: **Long noncoding RNAs are rarely translated in two human cell lines**. *Genome Res* 2012, **22**:1646-1657.

66. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling**. *Science* 2009, **324**:218-223.

67. Martin JA, Wang Z: **Next-generation transcriptome assembly**. *Nat Rev Genet* 2011, **12**:671-682.

68. Bao E, Jiang T, Girke T: **BRANCH: boosting RNA-Seq assemblies with partial or related genomic sequences**. *Bioinformatics* 2013, **29**:1250-1259.

69. Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F, Peterlongo P, Lacroix V: **KISSPLICE: de-novo calling alternative splicing events from RNA-seq data**. *BMC Bioinform* 2012, **13(Suppl. 6)**:S5.

70. Dinger ME, Pang KC, Mercer TR, Mattick JS: **Differentiating protein-coding and noncoding RNA: challenges and ambiguities**. *PLoS Comp Biol* 2008, **4**:e1000176.

71. Kageyama Y, Kondo T, Hashimoto Y: **Coding vs non-coding: translatability of short ORFs found in putative non-coding transcripts**. *Biochimie* 2011, **93**:1981-1986.

72. Cohen SM: **Everything old is new again: (linc)RNAs make proteins!** *EMBO J* 2014, **33**:937-939.

73. Chooniedass-Kothari S, Emberley E, Hamedani MK, Troup S, Wang X, Czosnek A, Hube F, Mutawe M, Watson PH, Leygue E: **The steroid receptor RNA activator is the first functional RNA encoding a protein**. *FEBS Lett* 2004, **566**:43-47.

74. Simon MD, Wang CI, Kharchenko PV, West JA, Chapman BA, Alekseyenko AA, Borowsky ML, Kuroda MI, Kingston RE: **The genomic binding sites of a noncoding RNA**. *Proc Natl Acad Sci* 2011, **108**:20497-20502.

75. •• Quinn JJ, Ilik IA, Qu K, Georgiev P, Chu C, Akhtar A, Chang HY: **Revealing long noncoding RNA architecture and functions using domain-specific chromatin isolation by RNA purification**. *Nat Biotechnol* 2014:32.

This article presents domain-specific chromatin isolation by RNA purification (dChIRP), a technique to analyze RNA–RNA, RNA–protein and RNA–chromatin interactions used to precisely characterize the chromosomal binding of roX1 and roX2.

76. • Park C, Yu N, Choi I, Kim W, Lee S: **lncRNAtor: a comprehensive resource for functional investigation of long non-coding RNAs**. *Bioinformatics* 2014, **30**:2480-2485.

A new database integrating functional information about lncRNAs from six species, including fruitfly.