Detecting and comparing non-coding RNAs

Giovanni Bussotti

TESI DOCTORAL UPF / ANY 2012

THESIS DIRECTOR

Dr. Cedric Notredame

THESIS DEPARTMENT

BIOINFORMATICS AND GENOMICS PROGRAMME AT CRG

(CENTER FOR GENOMIC REGULATION)

Vorrei dedicare non solo questa tesi ma più in generale questo periodo di quattro anni a tutti i familiari che hanno continuato ad appoggiarmi di la dal mare. Malgrado non sia possible recuperare il tempo passato, di certo bisogna rallegrarsi di quello presente e guardare con ottimismo al futuro.

In particolare la mia dedica va ai miei due bei nipoti Giacomo e Paolo, perche' possano avere sempre tanta curiosita'.

Il mio ricordo e la mia dedica vanno anche a nonno Torinto e ai pomeriggi passati insieme a raccogliere frutta.

« Sempre caro mi fu quest'ermo colle,
e questa siepe, che da tanta parte
dell'ultimo orizzonte il guardo esclude.
Ma sedendo e mirando, interminati
spazi di là da quella, e sovrumani
silenzi, e profondissima quiete
io nel pensier mi fingo, ove per poco
il cor non si spaura. E come il vento
odo stormir tra queste piante, io quello
infinito silenzio a questa voce
vo comparando: e mi sovvien l'eterno,
e le morte stagioni, e la presente
e viva, e il suon di lei. Così tra questa
immensità s'annega il pensier mio:
e il naufragar m'è dolce in questo mare. »

Giacomo Leopardi, L'infinito

Acknowledgments

During these unforgettable years in Barcelona I was very fortunate to be part of an extraordinary research group in an outstanding institute, the CRG. It would be impossible for me to say how thankful I am for all the help and support I have got, both from the scientific and the non-scientific personnel. I have taken advantage so many times of the interactions with people at CRG that it would just be impossible to acknowledge each and every single person. Still, I would like to take the opportunity to express gratitude to some people in particular.

Above all I wish to give credit to my supervisor Dr. Cedric Notredame. Working under his guidance has been not just a unique opportunity to learn science from one of the most brilliant researchers I ever met, but also a pleasant human experience. I acknowledge him for all the efforts in managing our group wisely, and creating the most favourable working conditions ever. In these years I really enjoyed going to work every single day, and this is mostly merit of Cedric.

I wish to acknowledge the members of my thesis committee, Dr. Roderic Guigó, Dr. Juan Valcárcel and Dr. Eduardo Eyras for their availability and help all along these years.

I would like to say special thanks to Roderic Guigó. During these years I had the chance to collaborate with him and his group in several projects. This has been a unique opportunity for me to get in touch with cutting-edge researchers all across the world and, even if only a little, to participate in the prestigious ENCODE project.

I would like to say thanks to all the people in Cedric's and Roderic's groups. Special thanks go to Ionas Erb and Carsten Kemena, colleagues and friends in these years. Especially Ionas followed my PhD progresses, and lastly participated in the revision of this thesis.

Finally, I would love to say thanks to Cedric Notredame, Roderic Guigó, Gian Gaetano Tartaglia and Anna Tramontano for providing references in the search of a post-doc position.

Abstract

In recent years there has been a growing interest in the field of non-coding RNA. This surge is a direct consequence of the discovery of a huge number of new non-coding genes, and of the finding that many of these transcripts are involved in key cellular functions. In this context, accurately detecting and comparing RNA sequences becomes extremely important. Aligning nucleotide sequences is one of the main requisite when searching for homologous genes. Accurate alignments reveal evolutionary relationships, conserved regions and more generally, any biologically relevant pattern. Comparing RNA molecules is, however, a challenging task. The nucleotide alphabet is simpler and therefore less informative than that of proteins. Moreover for many non-coding RNAs, evolution is likely to be mostly constrained at the structure level and not on the sequence level. This results in a very poor sequence conservation impeding the comparison of these molecules. These difficulties define a context where new methods are urgently needed in order to exploit experimental results at their full potential.

These are the issues I have tried to address in my PhD. I have started by developing a novel algorithm able to reveal the homology relationship of distantly related ncRNA genes, and then I have applied the approach thus defined in combination with other sophisticated data mining tools to discover novel non-coding genes and generate genome-wide ncRNA predictions.

Resumen

En los últimos años el interés en el campo de los ARN no codificantes ha crecido mucho a causa del enorme aumento de la cantidad de secuencias no codificantes disponibles y a que muchos de estos transcriptos han dado muestra de ser importantes en varias funciones celulares. En este contexto, es fundamental el desarrollo de métodos para la correcta detección y comparativa de secuencias de ARN. Alinear nucleótidos es uno de los enfoques principales para buscar genes homólogos, identificar relaciones evolutivas, regiones conservadas y en general, patrones biológicos importantes. Sin embargo, comparar moléculas de ARN es una tarea difícil. Esto es debido a que el alfabeto de nucleótidos es más simple y por ello menos informativo que el de las proteínas. Además es probable que para muchos ARN la evolución haya mantenido la estructura en mayor grado que la secuencia, y esto hace que las secuencias sean poco conservadas y difícilmente comparables. Por lo tanto, hacen falta nuevos métodos capaces de utilizar otras fuentes de información para generar mejores alineamientos de ARN. En esta tesis doctoral se ha intentado dar respuesta exactamente a estas temáticas. Por un lado desarrollado un nuevo algoritmo para detectar relaciones de homología entre genes de ARN no codificantes evolutivamente lejanos. Por otro lado se ha hecho minería de datos mediante el uso de datos ya disponibles para descubrir nuevos genes y generar perfiles de ARN no codificantes en todo el genoma.

Preface

This work focuses on the comparative genomics of non-coding RNAs in the context of new sequencing technologies. Within this vast subject, this work aims at dealing with two extremely important research aspects nowadays: the development of new methods to align RNAs and the analysis of high-throughput data. Regarding the methodological aspect, this work introduces BlastR, a new in-silico tool able to reveal the homologous relationships between distantly related non-coding genes in a fast and reliable way. This tool is able to deal with poor sequence conservation by taking into account additional information sources and is less computationally demanding than state of the art methods. The data analysis part of this work is centred mainly on investigating the conservation of long non-coding RNAs using a combination of techniques. The unprecedented amount of expression data returned by next generation sequencing technologies allowed the detection of thousands of new and uncharacterized non-coding genes. Despite the fact that just a few dozens were functionally characterized, many of these genes are likely to be key regulators of diverse cellular processes and probably involved in important biological functions.

Keywords

non-coding RNA, long non-coding RNA, alignment, comparative biology, Blast, homology search.

Index

Abbreviations

CAGE	Cap Analysis of Gene Expression
CPU	Central Processing Unit
ChIP-seq	Chromatin ImmunoPrecipitation - sequencing
dUTP	deoxyUridine TriPhosphate
E-value	Expect value
EST	Expressed Sequence Tags
FACS	Flow-Assisted Cell Sorting
LQ-DGE	Low-Quantity Digital Gene Expression
MFE	Minimum Free Energy
MPSS	Massively Parallel Signature Sequencing
MSA	Multiple Sequence Alignment
NGS	Next Generation Sequencing
NMR	Nuclear Magnetic Resonance
NP-complete	Nondeterministic Polynomial-time complete
ORF	Open Reading Frame
PEAT	Paired-End Analysis of TSSs
PETs	Paired-End diTags
Pol II	RNA polymerase II
RACE	Rapid Amplification of cDNA Ends
RNA-seq	high-throughput cDNA sequencing
SAGE	Serial Analysis of Gene Expression
SCFG	Stochastic Context Free Grammar
TSS	Transcription Start Site
cDNA	complementary DNA
lincRNA	large intervening ncRNA
lncRNA	long non coding RNA
mRNA	messenger RNA
ncRNA	non coding RNA
poly(A)	polyAdenylated
rRNA	ribosomal RNA
smsDGE	single-molecule sequencing Digital Gene Expression

CHAPTER 1: Introduction

In recent years, the non-coding RNA (ncRNA) field has rapidly expanded (Figure 1) with a fast increase in the number of newly identified and biologically relevant ncRNAs. Just a decade ago, the number of known ncRNAs was restricted to a small amount of housekeeping genes (including ribosomal RNAs, transfer RNAs and spliceosomal RNAs) and an even more limited collection of regulatory RNAs, such as lin-4 in *Caenorhabditis elegans* (Lee et al., 1993) and Xist in mammals (Brown et al., 1992). Since then, the number of novel ncRNAs has increased dramatically and much more is known about their function, biogenesis, length, structural and sequence features. New and ever more sophisticated high-throughput technologies, such as tiling arrays, 454 and Solexa sequencing have been applied to comprehensively profile the transcriptome of various organisms.

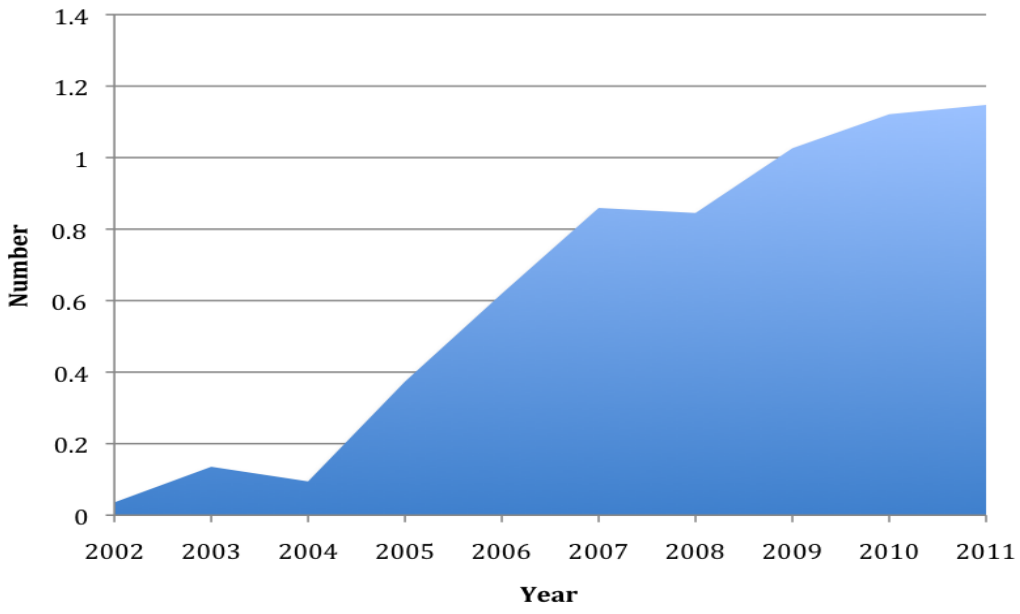

Figure 1 - *Number of publications in PubMed found using the keyword "ncRNA". The x-axis represents the timeline, the y-axis the number of times the word "ncRNA" matches a publication in PubMed normalized by the total number of publications in that year (expressed as one part per ten thousand).*

This wealth of data has allowed the identification of thousands of novel short ncRNAs, including PIWI interacting RNAs (Farazi et al., 2008) and small nucleolar RNAs (Bachellerie et al., 2002) and has resulted in the compilation or the update of many publicly available databases (Barrett et al., 2005; Parkinson et al., 2005; Griffiths-Jones et al., 2006; Fraser et al., 2011; Tuda et al., 2011; Mamidala et al., 2012). Furthermore, high-throughput approaches also revealed a massive transcription of long ncRNAs (lncRNAs) (Clark et al., 2011), operationally defined as RNA longer than 200 base pairs that do not template protein synthesis. In the human genome, for instance, the GENCODE consortium annotated 9640 lncRNA loci representing 15512 transcripts (Harrow et al., 2012). These discoveries were very timely in a context of growing concerns for the lack of a significant correlation between the number of protein coding genes and the commonly accepted concept of "organism complexity" (Mattick, 2001; Mattick and Gagen, 2001). It was proposed that alternative splicing and ncRNAs could be accountable for complex gene regulation architectures, meaning that the "Central Dogma" of genetic programming enunciated by Francis Crick in 1958 (RNA is transcribed from DNA and translated into protein) (Crick, 1958) had to be slightly altered, and at least in higher eukaryotes is inadequate (Mattick, 2001; Mattick and Gagen, 2001). The biological role of most of these novel long untranslated molecules is still a controversial issue. Some authors have even raised doubts on whether these transcripts are functional at all (Wang et al., 2004). The lack of shared discernible features is making it hard to define lncRNA classes, thus impeding any function prediction (Dinger et al., 2009). However mounting experimental evidences have shown that lncRNAs are implicated in a variety of biological processes (Mattick, 2009) and linked to various diseases including cancer (Wapinski and Chang, 2011). The functional roles of lncRNA transcripts have been uncovered in signalling sensors (Wang et al., 2011), embryonic stem cell differentiation (Dinger et al., 2008), brain function (Satterlee et al., 2007; Mercer et al., 2008), subcellular compartmentalization and chromatin remodelling (Kaikkonen et al., 2011). Among others, some examples include the X chromosome inactivation by Xist, the silencing of autosomal imprinted genes

accomplished by Air, the nuclear trafficking regulated by NRON and muscle differentiation controlled by linc-MD1 (Brown et al., 1992; Braidotti et al., 2004; Willingham et al., 2005; Cesana et al., 2011). See table 1 in (Mattick, 2009) and (Rinn and Chang, 2012) for more examples and lncRNAdb (Amaral et al., 2011) for the central repository of known lncRNAs in eukaryotes. lncRNAs are expressed, some are spliced, they are often conserved across vertebrates, and their expression is frequently tissue- and/or cell-specific and localized to specific subcellular compartments (Ravasi et al., 2006; Dinger et al., 2008; Mercer et al., 2008). It has been shown that lncRNAs can act both in cis (Wang et al., 2008b; Orom et al., 2010) and in trans (Rinn et al., 2007), some acting as precursors for short ncRNAs (Rodriguez et al., 2004; Kapranov et al., 2007b; Ogawa et al., 2008), while others are acting independently as long transcripts. lncRNAs can be compared and classified according to their similarity. So far, about half of reported human lncRNA have shown to be significantly conserved across mammals (Derrien et al., 2012b). These levels suggest some key cellular function, even though only a small fraction of these transcripts have so far been functionally characterized. Such functional analysis remains, however, very superficial and we are still in need of a precise molecular mechanism explicating the way this new class of transcripts acts.

Our low level of understanding can be in part attributed to the difficulty when working experimentally with lncRNAs: detection is difficult for a combination of biological and technical aspects. The first relates to the low levels of non-coding genetic expression. Once excluded the ribosomal RNA (rRNA) fraction that normally represents over 90% of total RNA, protein-coding mRNAs constitute by far the most abundant transcript component in cells. In (Ravasi et al., 2006; Guttman et al., 2010; Cabili et al., 2011) the authors report how lncRNAs are on average 3 to 10 fold less expressed than mRNAs. Besides the complicated task of capturing weaker expression signals, many lncRNAs have pronounced tissue/stage specificity (Cabili et al., 2011; Kutter et al., 2012). In other words, lncRNA genes can easily be left undetected unless the correct cell type and condition are considered. One more complication for ncRNA discovery has been the

difficulty of sequencing deep enough, a hurdle only recently overcome by Next Generation Sequencing (NGS). Only a few years ago, the amount of available sequences and the sequencing ability were both limiting factors, the genomes were not assembled, and there was virtually no notion of transcriptome complexity. Most of the classical low-throughput approaches, such as RT-PCR and northern blotting, have been successfully used to analyze the expression of small numbers of genes, but they were not adequate to address the "pervasive transcription" nature of the genomes (ENCODE Project Consortium, 2007; Clark et al., 2011).

A major obstacle in ncRNA detection is the difficulty to do informative sequence comparisons. Standard primary sequence alignment procedures are hampered at the very start by the low complexity of the nucleic alphabet, making difficult to produce statistically meaningful RNA alignment. The ribonucleic acid chemistry relies on just four different residues: two purines and two pyrimidines. Consequently, RNA gene sequences do not have strong statistical signals, unlike protein coding genes. For instance two RNA sequences must share an identity of at least ~60% to be considered significant in homology relationships prediction (Capriotti and Marti-Renom, 2010). Below this level, common ancestry is hard to infer with enough certainty. By comparison, this threshold is around ~20-35% for proteins (Rost, 1999). Furthermore, ncRNA appear to be evolving quickly (Pang et al., 2006) or under the influence of very specific evolutionary constraints (Pang et al., 2006). It was proposed that most ncRNAs evolve at higher mutation rates, with the maintenance of secondary structures being the main source of selection (Bernhart and Hofacker, 2009; Sun et al., 2012). This assumption makes sense from an evolutionary standpoint. As ncRNAs will be left untranslated, the nucleotide sequence itself is not restrained to keep the codon reading frame. Of course, exceptions exist. Specific ncRNAs types can hold functional sequences and act via their primary sequence (i.e. miRNAs). Previous report have shown that at least some miRNA genes are well conserved across species (Bentwich et al., 2005; Berezikov et al., 2006; Guerra-Assuncao and Enright, 2012), reinforcing the idea that sequences encoding a function evolve under purifying selection. Aside from

these specific and relatively rare examples, it seems that for most known ncRNAs, evolution is limited by structural constraints (Missal et al., 2005; Lindgreen et al., 2007) that induce a characteristic pattern of compensatory mutations (Figure 2). Such compensations occur when a mutation is affecting a nucleotide pairing to another in a structured domain. If the mutation breaks the base pairing so that the functionality of such a domain is compromised, the matching nucleotide is favoured to mutate in turn, i.e. is co-varying to restore the base pairing and keep the structure unchanged.

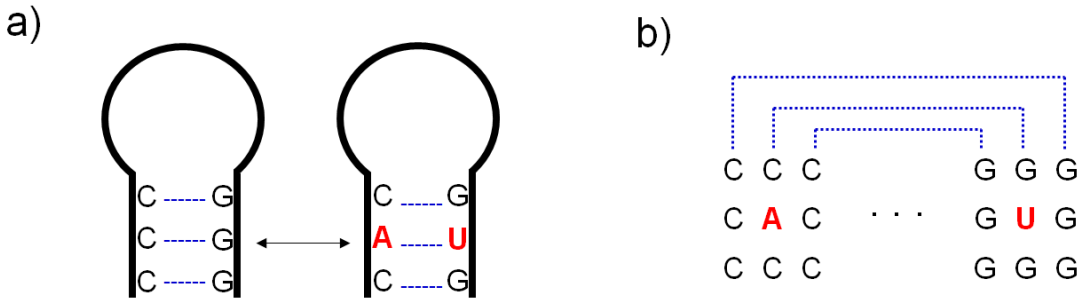

Figure 2 - *RNA mutations are tightly linked to the RNA structure conservation. a) Example where the mutation of a C into an A is compensated by the change G - U. The two positions are not independent, but communicating one with the other to maintain the structure unvaried. b) Same hairpin as shown in figure a). The presence of the compensatory mutation is highlighted by the multiple sequence comparison.*

For most aligners this is a circumstance hard to account for when using standard alignment procedures that postulate positional independence and seek only to maximize identity. As a consequence ncRNA sequences are much harder to align than proteins, a limitation that affects our ability to accurately detect and classify them. The difficulty in comparing ncRNAs calls for other information sources that alignment algorithms can use. More than ever, the issue of accurately comparing and aligning ncRNAs is of critical importance. This is precisely the problem discussed in the following section, where I will review older and more recent methodologies able to make the best of available RNA information.

Comparing non coding RNAs

As discussed, generating meaningful ncRNA alignments is a challenging task, and at least in some cases, the best accuracy could be achieved exploiting the RNA structural information. However, in many situations using such information is complicated. In spite of the development of aligners that take into account the RNA secondary structure information, one major issue is the poor availability of high quality structures. The problem is at least in part due to the difficulties encountered at experimental level in crystallization. Getting crystals from RNA molecules is complicated because of their chemical specificity. The accumulation of crystals is prevented by the high RNA flexibility. RNAs have flexible structures adopting inter-domain movements and with respect to proteins have weaker tertiary interactions (Ravindran et al., 2011). The polyanionic charge of phosphate backbone makes the nucleotide sequence swing much more than in proteins and this makes the packaging of crystals much harder to achieve. As a consequence, the crystals are either hard to grow, or poorly informative. Even when trying to resolve RNA molecules in solution through NMR, the resonance assignment is more difficult for RNA than for proteins (Furtig et al., 2003). RNAs have only 4 chemically similar nucleosides instead of the 20 different side chains found in proteins (Tzakos et al., 2006). Thus, the chemical shift dispersion is narrower in RNA than in proteins, resulting in less informative spectra (Tzakos et al., 2006). Because of these limitations, RNA structure is usually predicted (Zuker and Stiegler, 1981; Zuker, 2003). RNA secondary structure inference amounts to the computation of base-pairings that shape the in vivo molecule structure. The prediction can be done given the mere succession of bases of a single sequence. Another possibility is including other sources of statistical information to constrain a structure prediction, for instance an alignment of structurally homologous RNA sequences. Regarding single sequence RNA secondary structure predictions, there are two main groups of approaches: empirical free-energy parameters (Mathews et al., 2007) and knowledge based (Dowell and Eddy, 2004; Dima et al., 2005; Do et al., 2006). The first considers a biophysical model to calculate the structure whose folding has the minimum Gibbs free energy (ΔG). In this approach,

(Xia et al., 1998; Mathews et al., 1999; Mathews et al., 2004; Lu et al., 2006; Lu et al., 2009) the nearest stable folding is employed to compute the conformational stability of the Minimum Free Energy (MFE) structure. The energy parameters needed in this approach were assessed on a set of optical melting experiments on model systems (Mathews et al., 1999; Mathews et al., 2004; Lu et al., 2006). The two most popular implementations of the MFE structure prediction algorithm are mfold (Zuker and Stiegler, 1981) and RNAfold (Hofacker et al., 1994) packages. The latter implements McCaskill's algorithm (McCaskill, 1990), an approach to calculate the probability of a certain secondary structure in the whole thermodynamic ensemble. This approach is based on the partition function, which sums all Boltzmann weighted free energies of each secondary structure that is possible given an RNA sequence. In this model, the confidence estimate in a particular base pair *i,j* is given by the sum of the probabilities of all structures containing that base pair *i,j* divided by the sum over all structures (Durbin et al., 1998). Knowledge based approaches are rely on probabilistic models, where statistical learning procedures are used instead of empirical measurement of thermodynamic parameters. The Stochastic Context Free Grammar (SCFG) model (Dowell and Eddy, 2004) represents one popular example of such probabilistic models. The parameters used by the SCFG models are estimated on the set of RNAs with known structures (e.g. rRNA).

Prediction accuracy is the main limit of both MFE and knowledge based methods (Deigan et al., 2009) (see the example in figure 3). The percentage of known base pairs predicted correctly by the secondary structure prediction methods ranges from 40 to 75% (Doshi et al., 2004; Dowell and Eddy, 2004; Dima et al., 2005; Do et al., 2006). This low figure may result from three confounding factors. First of all, the folding in vivo can be influenced by RNA chaperons (Herschlag, 1995), RNA editing (Brennicke et al., 1999), and even by the transcriptional process itself (Pan and Sosnick, 2006). At present, there is no software able to accounts for these effects. Secondly, looking for a single structure can sometimes be inadequate. There are cases, such as the ribo-switches, (Mandal and Breaker, 2004; Soukup and Soukup, 2004) where multiple

functional structures can be derived from the same sequence. Standard predictors are not well suited to deal with such situations that require dedicated tools able to identify potential conformational switches (Bengert and Dandekar, 2004; Voss et al., 2004). Thirdly, RNAs might contain pseudo-knots, which are ignored by most tools due to reasons of computational complexity (Gardner and Giegerich, 2004).

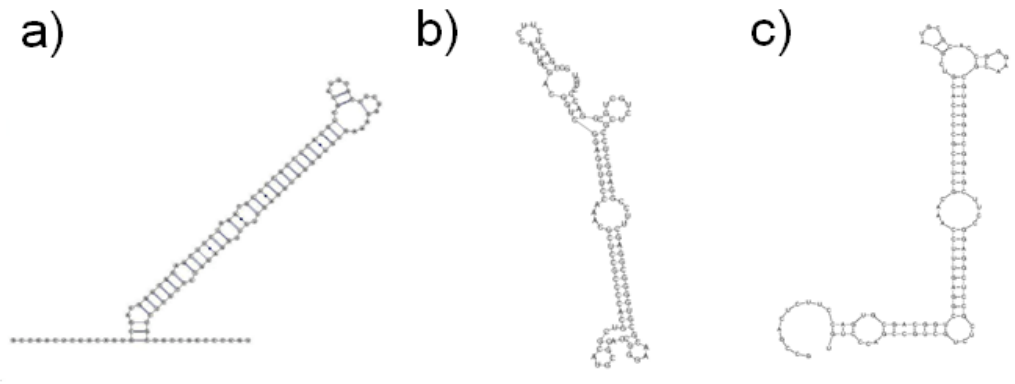

Figure 3 - *Reliability of RNA secondary structure predictions. In this example the human mir-3180 (Rfam accession id RF02010; AJ323057.1/363-249) was folded using different approaches yielding different output structures. a) Secondary structure of the family as estimated by Rfam release 10.1. b) RNAfold web server prediction based on Vienna RNA package version 2.0.0. (Hofacker, 2003) C) mfold (Zuker, 2003) web server prediction, running mfold version 4.6.*

The best secondary structure prediction accuracy can be achieved using comparative methods (Gardner and Giegerich, 2004). These apply to a set of structurally homologous RNA sequences being aligned. For some of these computation tools, the output will be the prediction of each individual homologous structure, while in other situations the result will be a unique consensus structure. The consensus structure does not exist in vivo, but rather it is a model that captures the common, relevant structural aspects conserved within the family. Due to the close relationship between sequence and structure, structure prediction and sequence alignment problem can be described as interdependent problems (Lindgreen et al., 2007). As theorized by Sankoff (Sankoff, 1985), the most suitable approach should involve the simultaneous alignment and

folding of the considered sequences. Unfortunately, a strict application of this approach would be computationally prohibitive and the lack of an appropriate heuristic solution is well reflected by the wealth of available alternative solutions. The web server WAR (Torarinsson and Lindgreen, 2008) is a good example. This tool allows the execution of a total of 14 different strategies to align and predict the secondary structure of multiple RNA sequences.

Over the years, so many methods have been described that some kind of classification is needed to catalogue them. Paul Gardner proposed three categories he refers to as "plans" (Gardner and Giegerich, 2004; Bremges et al., 2010). In plan A, one starts with the estimation of a multiple sequence alignment and then the aligned sequences are folded jointly (as a kind of consensus). The initial alignment can be done by any standard MSA aligner (e.g. ClustalW (Thompson et al., 1994), T-Coffee (Notredame et al., 2000)), and the folding of the aligned sequences can be performed by a plethora of tools (e.g. RNAalifold (Hofacker et al., 2002), PFOLD (Knudsen and Hein, 2003), ILM (Ruan et al., 2004), ConStruct (Luck et al., 1999)) optimizing a score based on compensated mutations and thermodynamics. However this strategy is effective just in a determined sequence similarity range. On the one hand, sequences too similar do not carry any covariance or purifying selection information and are not informative by an evolutionary standpoint. On the other hand, sequences need to be similar enough to be meaningfully aligned as below the "twilight zone" the sequence alignment tends to obscure the covariance signal (Bremges et al., 2010).

Plan B makes it possible to use evolutionary signal to help improving the reliability of structure predictions. This approach accounts for an underlying RNA substitution model where mutation probabilities are affected by structural dependencies. The maintenance of a 3D fold is a major evolutionary constraint influencing the acceptance of point mutations. From this perspective, a nucleotide located in the stem is not as free to mutate as a nucleotide located in a loop. Substitutions taking place in structured functional domains of RNAs can disrupt the wild-type conformation and seriously affect the molecular function. As a consequence, a nucleotide whose pairing has been

disrupted by the mutation of its mate, is more likely to mutate itself so as to recover the original structure and rescue the function. Back in 1985 Sankoff developed a dynamic programming algorithm able to take into account sequence and structure of an RNA molecule simultaneously (Sankoff, 1985). Unfortunately this algorithm is computationally expensive, with a running time equal to $O(N^{3m})$, where m is the number of sequences and N their length. This means that a pairwise comparison has the tremendous CPU cost of $O(N^6)$ which makes this algorithm inapplicable most of the times and calls for simplified heuristics. Several banded modifications of the Sankoff algorithm impose limits on the size or shape of substructures, like Dynalign (Mathews and Turner, 2002; Mathews, 2005), Foldalign (Gorodkin et al., 1997; Havgaard et al., 2005), Stemloc (Holmes, 2004, 2005), Consan (Dowell and Eddy, 2006). Another example is pmmulti (Hofacker et al., 2004), a Sankoff algorithm variant able to perform multiple secondary structure alignments by aligning consensus base pair probability matrices.

Plan C is used by programs such as R-Coffee (Wilm et al., 2008) or RNAcast (Reeder and Giegerich, 2005). In these methods each individual sequence is folded separately before running the alignment. Equivalent secondary structures between two RNAs can be used to enhance the alignment accuracy. For instance, let seq1 and seq2 be two RNA sequences, *x* and *y* be two nucleotides matching in a secondary structure in seq1, and *x'* and *y'* be two nucleotides matching in a secondary structure in seq2. If *x* aligns to *x'* then implicitly *y* should be driven to align to y'. For example, R-Coffee uses RNAplfold (Bernhart et al., 2006) to compute the secondary structure of the provided sequences. After that, R-Coffee computes the multiple sequence alignment having the best agreement between sequences and structures. This pre-folding approach is especially valuable when RNAs are too different to be meaningfully aligned merely by using an off-the-shelf multiple alignment tools (i.e. ClustalW (Thompson et al., 1994)). Plan C is particularly well suited to situations where experimental secondary structures are available.

Giving an exhaustive overview of the methods available for folding and aligning structured RNA sequences is well beyond the scope of this introduction. Over the last twenty years, more than 30 methods were described that deal with these problems, and that, on its own is an indication of the problem complexity that still remains open, more than 25 years after having been formally described by Sankoff. The situation is radically different when experimental 3D structure information is available. In that case the RNA alignment problem becomes similar to the protein structural alignment problem. This problem is nondeterministic polynomial-time complete (NP-complete) and it involves the alignment of two distance matrices. In most cases the problem can be solved in a rather satisfying way by using heuristics making the best of the geometric information contained in the PDB models. Examples of pairwise structural alignment methods for RNA are SARA (Capriotti and Marti-Renom, 2008), DIAL (Ferre et al., 2007) iPARTS (Wang et al., 2010), ARTS (Dror et al., 2005) and SARSA (Chang et al., 2008). Besides that, recently several 3D RNA structure database search programs were developed, such as LaJolla (Bauer et al., 2009) and FRASS (Kirillova et al., 2010).

ncRNA homologues detection

In the ncRNA field, a critical step is the collection of homologues to the genes of interest. Homologues can be used in several situations, like the detection of functional motifs, the inference of possible evolution steps or the design of wet lab experiments. For instance, the conservation across species of a certain ncRNA can be used to estimate how likely a gene is to be functionally important. Such information can be used to prioritize experiments, e.g knockdown experiments of the orthologous gene in a model organism. Over the last few years many different methods have been developed to approach the problem of RNA homology search. As shown in (Freyhult et al., 2007), the homology search methods can be grouped in three sets: sequence-based, profile HMM and structure-based methods. The first and most straightforward approach to look for homologues is by comparing the nucleotide sequences. Already in 1981 Smith and

Waterman developed a dynamic programming algorithm that allows for pairwise local alignment (Smith and Waterman, 1981). Nevertheless, this approach is CPU time demanding, and implementations of this method are unpractical for large-scale database and genome wide screenings. For this reason, alternative approaches such as FASTA (Lipman and Pearson, 1985) or Blast (Altschul et al., 1990) are frequently preferred. These are rapid searching heuristics able to boost computational speed at the cost of a reduced accuracy. In both Blast and FASTA, the underlying idea is to skip the time consuming comparison of entire query and target sequences, but rather to start identifying short conserved words in a first step called seeding. After that, more accurate time-consuming local alignments are performed.

Then the second class of approaches is based on profile HMMs. Profile HMMs are probabilistic models that are generated out of an input multiple sequence alignment where each position of the alignment is used to estimate nucleotide frequency. These models can be used to screen databases and look for homologs. Profile HMMs are heuristics having usually superior accuracy over methods based on single sequences (Eddy, 1998; Weinberg and Ruzzo, 2006). However, such models have a linear architecture best suitable for modelling linear protein sequences (as opposed to secondary structure relationships). A more appropriate modelling of an RNA alignment should also consider RNA base pair interactions. The best sensitivity can be attained when applying approaches taking into account at the same time sequence similarity and secondary structures, as the Sankoff algorithm does. Unfortunately, the Sankoff algorithm is computationally too demanding, hence the need for approximate heuristic implementations of this exact algorithm. Such approximations include banded Sankoff tools (Holmes, 2004; Havgaard et al., 2005; Mathews, 2005; Dowell and Eddy, 2006), genetic algorithm implementations such as RAGA (Notredame et al., 1997) and covariance models (CMs). CMs are the most commonly used methods to carry out efficient database screening, and can be described as special form of stochastic context free grammar (profile SCFGs). CMs were first introduced by Sean Eddy in (Eddy and Durbin, 1994) and implemented in Infernal (Eddy, 2002). This and other related

applications (Klein and Eddy, 2003; Weinberg and Ruzzo, 2006) belong to a class of broadly used homology search tools based on the automatic learning of statistical models (the CMs) estimated from an input multiple RNA alignment decorated with the consensus secondary structure. CMs are probabilistic models that can be derived unambiguously out of a structure-annotated sequence alignment and can be used in turn to query a target sequence database to find homologs. Conceptually CMs are similar to profile HMM but able to include RNA base-pairs interactions information. The modelling of such information results in a higher complexity and CMs are represented by a tree-like model architecture, where the tree shape directly mirrors the consensus RNA structure. Unlike HMM states that only allow the emission of matches and indels, CMs embed new states to handle paired/not-paired and directionality information. For instance, in the paired sites, deletions can involve either a single 5′ or 3′ nucleotide, or the complete base pair. The direction also matters for the insertions that can now concern either the 5′ or 3′ ends of a base pair. In order to permit multi-loops, the bifurcation states are incorporated as well. In spite of their superior accuracy, CM cannot be used in all situations and are restricted to the identification of unsplit genes. The mapping of queries composed by multiple exons is impossible due to the impossibility of aligning secondary structures to a target interrupted by introns whose position is unknown. Moreover CMs need to "learn" from a set of homologous transcripts, but the set of sequences required to train the model are not always available. There is some circularity in this problem where the CM is used to search homologs that are themselves needed to estimate the model. Another layer of complexity results from the need to assemble a multiple sequence alignment of homologous sequences needed to train the CM. In the CM the alignment will be used for a probabilistic description of the matches, mismatches, insertions and deletions. However, generating accurate RNA alignments is difficult. In Rfam (Griffiths-Jones et al., 2003) CMs parameters are trained on a high quality alignments (seed alignment) obtained from literature and/or manually curated. This input is used to estimate CMs, which are then passed to Infernal to do homology search. This CM/Infernal strategy is analogous to HMM/HMMER used

for Pfam (Finn et al., 2008). However, even considering later ameliorations (Nawrocki et al., 2009), Infernal is much more CPU expensive than HMMER and for this reason its use is not realistic for many real life large-scale screenings. An option for spotting promising sequence segments and accelerate the detection procedure is to include a pre-filtering step as done for the Rfam setup (Griffiths-Jones et al., 2005). This can be accomplished by means of *ad hoc* algorithms (Zhang et al., 2006), profile HMMs (Weinberg and Ruzzo, 2006) or Blast with relaxed expect values (E-values) to avoid losing sensitivity as achieved in Rfam (Gardner et al., 2009). A number of studies have been dedicated to the optimization of BlastN parameters for seeking RNA homologs. For instance, in (Freyhult et al., 2007; Roshan et al., 2008) the authors benchmarked the effectiveness of Blast and other popular homology search methods tuned for ncRNA screenings. In (Bussotti et al., 2011), we introduce BlastR, a method that both takes advantage of di-nucleotide conservation and BlastP as search engine to discover distantly related homologs. BlastR can be mounted on the top of CPU demanding algorithms to serve as a pre-filtering tool. One merit of this approach is that it does neither require profiles nor secondary structure information, but relies merely on the information encoded in the base sequences.

Together with sequence-based, profile HMM and structure-based methods, one possibility to get inter-species homologs involve the use of multiple genome alignments (Cabili et al., 2011). Once established reciprocity between blocks of genomes belonging to different organisms (i.e. syntenic regions), the coordinate transfer from one gene to its homolog is straightforward and just implies the projection of corresponding positions. This has been made possible thanks to the availability of genomic sequences (Lander et al., 2001; Venter et al., 2001; Aparicio et al., 2002; Waterston et al., 2002) and the development of alignment tools able to detect orthologous genomic regions, i.e. loci that proceeded from the same genomic position in the ancestral genome (Schwartz et al., 2003).

High-throughput technologies and genome-wide annotation of ncRNAs

Recent technological advances have allowed the collection of an unprecedented amount of RNA sequence data coming from a wide range of organisms and conditions. For many years the main strategy for transcript discovery had been the sequencing of cloned complementary DNA (cDNA) of expressed sequence tags (ESTs) (Boguski et al., 1994; Dias Neto et al., 2000; Gerhard et al., 2004). EST sequencing was successfully used for the generation of large-scale expression datasets (Boguski et al., 1993), and already in 1991 this approach was utilized for human gene discovery (Adams et al., 1991). Although it is widely acknowledged that ESTs represent a valuable resource to detect gene expression, they also came with severe limitations such as cost and sequencing requirements. Their dependence on bacterial cloning is an important source of bias that can lead to redundancy, under-representation or over-representation of host-selected transcripts (Bonaldo et al., 1996; Nagaraj et al., 2007; Mortazavi et al., 2008). More recently, DNA-Chip has made high throughput expression analysis much more practical, while the even more recent RNA-seq technology is promising transcriptomic analysis of unprecedented accuracy thanks to the application of NGS methods to transcriptome sequencing. Microarrays rely on a collection of nucleotide probe spots attached to a solid support. RNAs are labelled with fluorescent dyes, hybridized to the arrays, washed, and scanned with a laser (Malone and Oliver, 2011). Such arrays have been used for the investigation of known or predicted genes and have been until recently one of the most widespread technology for transcriptome exploration. Standard expression arrays are affected by several limitations including the hybridization and cross-hybridization artefacts (Eklund et al., 2006; Okoniewski and Miller, 2006; Casneuf et al., 2007), dye-based identification problems (Cox et al., 2004; Dombkowski et al., 2004; Rosenzweig et al., 2004; Dobbin et al., 2005; Martin-Magniette et al., 2005) and physical manufacturing restrictions, impeding the detection of splicing events and the discovery of unannotated genes (Mortazavi et al., 2008). A variant of traditional expression array is represented by tiling arrays. These are chips that use highly dense

and overlapping probes representing contiguous regions of genome. Several works relying on this technology and aiming at transcript discovery have been published (Bertone et al., 2004; Cheng et al., 2005; Royce et al., 2005; Kapranov et al., 2007b; Kapranov et al., 2007a). However tiling arrays require a substantial RNA quantity and have further limitations affecting the sensitivity, the specificity and splice detections (Mortazavi et al., 2008). For instance, as shown in (Wang et al., 2009), microarrays lack sensitivity for genes expressed either at low or very high levels and if compared with RNA-seq have much smaller dynamic range. As a consequence, microarrays are inadequate for the quantification of both the prevailing RNA classes, and the less abundant ones. For genes with medium levels of expression, RNA-seq and microarrays return comparable results (Marioni et al., 2008; Fu et al., 2009; Wang et al., 2009). Still, each approach presents very specific advantages and disadvantages. A thorough comparison of these two approaches lies outside the purpose of this text, for reference see (Marioni et al., 2008; Malone and Oliver, 2011; Ozsolak and Milos, 2011). Additional methods for high-throughput RNA discovery include the serial analysis of gene expression (SAGE) (Velculescu et al., 1995; Harbers and Carninci, 2005), several updated variants such as LongSAGE (Saha et al., 2002), RL-SAGE (Gowda et al., 2004), SuperSAGE (Matsumura et al., 2005) and analogous approaches like the massively parallel signature sequencing (MPSS) (Brenner et al., 2000). In general, SAGE-like methods consist in the cloning and then the sequencing of short tags (17-25 nucleotides) coming from RNA extract. The resulting tag sequences can be compared against the source genome or a reference RNA database to attain the digital count of transcript quantities. Two other protocols that can be used in combination with high-throughput sequencing are the paired-end ditags (PETs) (Ng et al., 2005) and the rapid amplification of cDNA ends (RACE) (Schaefer, 1995; Kapranov et al., 2005; Olivarius et al., 2009). Both approaches can be used to demarcate transcript boundaries, i.e. define start and end of a transcript. Such information is extremely valuable in situations where the first and last exons can be respectively 5’ and 3’ associated with other transcript isoforms, thus making it difficult to define gene boundaries. Similarly, the cap analysis

of gene expression (CAGE) (Shiraki et al., 2003) is a technique that allows high-throughout profiling of transcriptional starts points. Undoubtedly, high-throughput technologies opened the extraordinary possibility to get both qualitative and quantitative information on the whole transcripts mass produced by cells. This resulted in high-resolution views of RNA expression dynamics throughout different tissues and time points (Mathavan et al., 2005; Wang et al., 2008a; Graveley et al., 2011). RNA expression is the lowest measurable phenotypic trait as it represents the cell response to a particular environment or status. On a massive scale, the profiling of entire transcriptomes is a powerful resource to both evaluate the genetic expression in a steady state and also assess or quantify how various factors may perturb this normality. Examples include disease/health conditions (Demmer et al., 2008) and stress responses (Desikan et al., 2001; Halbeisen and Gerber, 2009).

Various groups and projects, such as RefSeq (Pruitt et al., 2009), GENCODE (Harrow et al., 2006; Harrow et al., 2012), HAVANA team (Havana team; Loveland et al., 2012) and Ensembl (Flicek et al., 2012) undertook the task to comprehensively annotate functional elements, including ncRNAs, of a number of species using experimental data. The RefSeq repository houses annotations resulting from automated analyses, collaboration and manual curation (Pruitt et al., 2009; Pruitt et al., 2012). The GENCODE pipeline combines HAVANA and Ensembl automatic annotations to annotate the human gene features generated in the context of the ENCODE project (Harrow et al., 2006; ENCODE Project Consortium, 2007, 2012; Harrow et al., 2012). The HAVANA team has the goal to provide manually curated annotations of transcripts aligned to human, mouse and zebrafish genomes. Ensembl runs an automatic *genebuild* process including *ab initio* gene predictions and release 64 supported a total of 61 species (Flicek et al., 2012). The Ensembl *genebuild* system is adapted to every species in the set according to the data that is available. For instance Ensembl imports and merges high quality HAVANA annotations exclusively for human and mouse. The annotations generated by these consortia are freely available through genome browsers, including UCSC (Kent et al., 2002), Ensembl (Stalker et al., 2004) and VEGA

(Wilming et al., 2008). As new genomic regions get annotated and new transcript sequences become publicly available, these gene sets continue to growth (Pruitt et al., 2009; Harrow et al., 2012; Pruitt et al., 2012). A recent publication (Harrow et al., 2012) indicated that in the last years the number of annotated protein coding and non-coding transcripts in GENCODE has dramatically increased. For instance, passing from GENCODE version 3c (July 2009) to version 7 (December 2010), the number of protein coding transcripts increased from 68880 to 76052, and the number of lncRNAs jumped from 10457 to 15512. The overall picture, however, remains blurred by inconsistent findings, suggesting that more analyses are still needed. For instance, recent estimates reported by the ENCODE project indicate that about the 62% of human genomic bases are expressed in long transcripts, while 5.5% only of the whole genome is found within the GENCODE annotated exons (ENCODE Project Consortium, 2012). This discrepancy can be in part explained by the fact that GENCODE catalogues transcripts using cDNA/EST alignments (Harrow et al., 2012) rather than RNA-seq short-reads. A classic low-throughput EST sequencing operated by the Sanger technology can identify mostly high abundant transcripts (Martin and Wang, 2011), while deep coverage RNA-seq experiments can reveal rare but potentially regulatory transcripts. Nonetheless, ESTs are longer than RNA-seq reads, and can provide more reliable transcriptional evidence (Rogers et al., 2012).

The full extent of RNA transcription landscape has remained largely unexplored till recently. The development of NGS platforms, including Roche/454, Illumina/Solexa and ABI/SOLiD, and their application to the study of the transcriptomes have improved our understanding of genome expression. The works that first tackled the study of mammalian transcriptomes and suggested the pervasive transcription of the genome were based on genome tiling array hybridization and cDNA sequencing (Kapranov et al., 2002; Carninci et al., 2005; Katayama et al., 2005). Then the progress in RNA-seq technology allowed the massive generation of both qualitative and quantitative expression data. The standard RNA-seq protocol can be divided in three main steps. The first step is the preparation of the RNA sample, involving both the RNA extraction and

the cDNA library generation. The Illumina Genome Analyzer platform protocol (Mortazavi et al., 2008) involves the poly(A) enrichment using oligo(dT) beads, RNA fragmentation and reverse transcription into cDNA primed by random hexamers (Hansen et al., 2010). The cDNA library set-up strongly depends on the research project, and either the poly(A)+ or poly(A)- fractions can be enriched. Most of the times is preferred to analyze the poly(A)+ fraction so as to bypass the sequencing of rRNA. Ribosomal RNA represents by far the most abundant RNA class in cells. To have sufficient sensitivity, detect other RNA species and measure gene expression variations, it is critical to sequence at enough depth and to clear the rRNA fraction. For this purpose, different strategies have been explored, including the use of poly(A) enriched RNA or the selective removal of rRNA (ribo-depletion) (Huang et al., 2011). The second step of a standard RNA-seq protocol is the sequencing experiment itself. The sequencer accepts the cDNA library in input, and returns in output millions of short reads. The last, final step, encompasses all the downstream bioinformatics analysis, including the mapping of the reads on a reference genome, assembling the reads into transcript models and estimating expression levels. Since it was first introduced, the RNA-seq technology evolved quickly and several specific applications and variations of the standard protocol are nowadays available. These include mapping the transcription start sites (TSSs) through CAGE derived methods (Kodzius et al., 2006) such as DeepCAGE (Valen et al., 2009), PEAT (Ni et al., 2010), nanoCAGE and CAGEscan (Plessy et al., 2010) and protocols for small RNA profiling (Rajagopalan et al., 2006; Ruby et al., 2006). RNA-seq stranded protocols can be used to discriminate transcripts expressed on the leading strand from transcripts expressed on the lagging strand. This issue is critical when dealing with transcripts that overlap in an anti-sense fashion, as often happens with lncRNAs and adjacent proteins (i.e. GENCODE 7 annotates 3214 antisense loci (Harrow et al., 2012)) or whenever anti-sense transcripts are produced by the genome (Katayama et al., 2005). Standard protocol to prepare RNA-seq libraries loose the read directionality information when passing from single stranded RNA to double strand cDNA. As a consequence, un-stranded reads cannot be used to quantify

the expression of genes overlapping on opposite strands. Most NGS platforms now support strand-specific RNA-seq protocols (Cloonan et al., 2008; Core et al., 2008; He et al., 2008; Lipson et al., 2009; Parkhomchuk et al., 2009; Mamanova et al., 2010; Ozsolak et al., 2010a; Ozsolak et al., 2010b). RNA-seq technology advances also contributed to the annotation of alternatively spliced transcripts thanks to the pair-end reads protocol, i.e. the reads come from two positions in a transcripts separated by an insert of controlled size. Pair-end reads make it possible to monitor how alternative splicing affects exon combinations without the need of pre-existing annotations (Trapnell et al., 2009; Ameur et al., 2010). Furthermore, pair-end reads can be used to detect gene fusion events. This application has been successfully used in the study of melanoma transcriptome (Berger et al., 2010) and other cancers (Palanisamy et al., 2010). Additionally, pair-end RNA-seq can facilitate the reads alignment to genomic repeats (Ozsolak and Milos, 2011). Thanks to protocol simplicity and to the cost effectiveness of RNA-seq technology (Wilhelm and Landry, 2009; Costa et al., 2010), an increasing number of transcriptomes have been released and published (Wang et al., 2009). Unfortunately, various RNA-seq protocols are biased in different ways when measuring transcript expression. This makes systematic comparisons across experiments a difficult process. Various bioinformatics patches have been proposed to alleviate this phenomenon (Oshlack and Wakefield, 2009; Roberts et al., 2011). A review from Helicos BioSciences Corporation about the state-of-art RNA-seq, the limitations, advances and future perspectives is provided in (Ozsolak and Milos, 2011).

A part of my thesis focused on data analysis of large numbers of putative lncRNAs detected via NGS technologies. To make the most out of the extraordinary possibilities NGS is offering, it is essential to understand the current limitations. One important point is that the reads returned by standard NGS platforms are usually short (35-500 base pairs (Metzker, 2010)) and as a consequence of that it becomes necessary to reassemble the full-length transcripts. The short RNAs (i.e. miRNA and piRNAs) represent an exception and there is no need to reassemble them, as they are small enough to be entirely covered by the read length. Unfortunately the process of reassembling

transcriptomes starting from short reads is difficult. Normally RNA-seq dataset are big (gigabases to terabases), and thus need to be handled by sufficient memories and by multi-CPU processors able to execute the algorithms in parallel. Although various short-read assemblers (Butler et al., 2008; Zerbino and Birney, 2008; Simpson et al., 2009) were successfully applied to genome assembly, these software cannot be easily used to reconstruct transcriptomes. Applying to transcriptomes tools normally designed for genome reconstruction leads to several complications. One main issue is that the DNA sequencing depth is supposed to be identical over the entire genome while the transcriptome sequencing depth is expected to fluctuate a lot. For this reason, DNA short-read assemblers could erroneously interpret highly abundant transcripts as repetitive genomic regions. Furthermore, when using genome short-read assemblers the read strand is not taken into account. On the contrary, when available, a transcriptome assembler should exploit the strand information to unravel possible antisense expressions on different strands. Finally, the transcript modelling is involved as transcript variants coming from the same gene can share exons and are difficult to resolve unambiguously (Martin and Wang, 2011).

It is possible to work out the transcriptome assembly following a reference-based approach, a *de novo* assembly or combinations of them (Martin and Wang, 2011). The first considers the initial mapping of the reads on a reference genome, and then the usage of transcript assemblers. To the end of labelling each read with the genomic location they come from, a new class of software, generally referred as read mapper, has recently shown up. In this context, the availability and the quality of the underlying reference genome are critical. Besides that, when dealing with massive amount of short-read data the CPU and the memory costs can be challenging, and several algorithms are being tailored to achieve best mapping efficiency (Li et al., 2008b; Lin et al., 2008; Langmead et al., 2009; Li and Durbin, 2009; Schatz, 2009; Ahmadi et al., 2012; Derrien et al., 2012a). Other important issues relate to the mapping of reads crossing exon-junction boundaries (Cloonan et al., 2009; Trapnell et al., 2009) and the uncertainty or lack of accuracy in read alignments. For most downstream applications, the accurate

positioning of the reads back to the source genome is crucial. To improve the mapping accuracy, the process can take into account the read quality information (Li et al., 2008a; Smith et al., 2008). The quality scores, introduced by the Phred algorithm (Ewing and Green, 1998; Ewing et al., 1998), indicate the reliability of each base call in each read. Since the bases with reduced quality scores have an increased possibility to be sequencing errors, a read mapper should either use less severe penalization for mismatches at positions with low base-call quality, or not align such positions at all. The information about the quality score is particularly relevant when mapping reads with bigger sizes. This is because it is recognized that the 3' extremity of longer reads are affected by sequencing errors at higher rates (Smith et al., 2008). Besides choosing a threshold on the amount of accepted mismatches, other important and sometimes arbitrary decisions regard the split mapping and multiple mapping reads. The first refers to reads that could not be aligned to the reference genome unless split in subparts. Such reads could either highlight the presence of an unreported exon-junction boundary, or be sequencing artefacts. The second indicate reads that align multiple times across the reference genome. This mapping uncertainty is caused by repeated elements and may results in flawed expression establishments. On the one hand, removing multiple mapping reads from the analysis would imply an underestimation of the expression of genes embedding repeats. On the other hand, considering multiple mapping reads would lead to artefactual expression measurements. Once mapped the reads, additional issues concern the application of transcript assemblers. Several bioinformatics tools have been developed with the purpose of reconstructing transcripts in their entire length, i.e. annotating exon-intron transcript structures. These methods include Cufflinks (Trapnell et al., 2010), Isolazo (Li et al., 2011) and Scripture (Guttman et al., 2010). In (Palmieri et al., 2012) the authors have shown that variations across transcript assemblers can be source of confusion, with low consistency across methods and a high number of false positives (Li et al., 2011; Rogers et al., 2012). Transcript assemblers seem to have a better agreement when reconstructing protein-coding transcripts (Cabili et al., 2011) with the agreement dropping dramatically when modelling large intervening ncRNAs

(lincRNAs). For instance, Cufflinks and Scripture share a bare 46% agreement for lincRNAs transcript models (Cabili et al., 2011). Such discrepancies are caused by the differences in how each assembler reconstruct lowly expressed transcripts (Cabili et al., 2011). In other words, about half of the isoforms estimated by a method in areas with low reads density do not correspond to isoforms called by the other method. This poor agreement between transcript assemblers highlights the need of further improvements, calling for the development of new algorithms to accurately represent low abundant transcripts.

Another possibility to assemble a transcriptome from short reads is the *de novo* assembly of the transcripts. This strategy does not require any reference genome and is therefore independent on the correct alignment of the reads to the splice sites. Examples of applications adopting this strategy are described in (Garg et al., 2011; Grabherr et al., 2011; Jager et al., 2011). Nevertheless the application of *de novo* assembly to complex transcriptomes (e.g. higher eukaryotes) is complicated by the dataset sizes and the dense network of alternatively spliced variants. Furthermore, *de novo* transcriptome assemblers need much deeper sequencing than reference-based assemblers and are largely affected by sequencing errors (Martin and Wang, 2011).

Once generated a transcriptome dataset, there are additional complications in the downstream analysis if trying to distinguish genuine ncRNAs from mRNAs. Nowadays this issue is getting more and more important as many researches expressly focus just on one of these two parts. The most straightforward procedure would be comparing the newly generated transcriptome against existing gene annotations. However in most cases annotations are far from being complete, and the great majority of genes they include are protein coding. As a consequence, in a normal RNA-seq experiment a substantial fraction of read contigs map outside of annotated exons (ENCODE Project Consortium, 2012). Previously unreported transcripts can be either classified as ncRNA or mRNA according to the protein coding potential they have. The in-silico assignment of a transcript to one of these two groups is not always simple and it might require dedicated expert analysis (Havana team). Some transcript isoforms might insert coding

exons and therefore could be partially translated, i.e. generating small peptides. There are further ambiguities for coding transcripts whose untranslated structured molecules are also functional as ncRNAs (Keiler, 2008) and for genes having both coding and non-coding isoforms (Novikova et al., 2012). A commonly used approach to predict the coding potential involves the codon substitution frequency (CSF) estimation (Clamp et al., 2007). This measure is based on an input multiple alignment of orthologous sequences. The CSF score deems a region to be coding depending on how the sequences of the multiple alignment evolved, i.e. showing distinctive mutation patterns, as are expected in coding and non-coding loci. A coding region is expected to embed prevalently conservative amino acid substitutions and synonymous codon substitutions, while showing low occurrence of nonsense and missense mutations. Although CSF has been successfully applied in various research projects (Clamp et al., 2007; Lin et al., 2007; Liao et al., 2011), the score is not always easy to estimate with the availability of trustworthy orthologues being the main limiting factor when dealing with new transcriptome datasets. Issues include scarcity or even the absence of orthologs, erroneous insert of pseudogenes in the set and absence of informative variations. For instance, as shown in (Derrien et al., 2012b) many putative human lncRNAs are not found in other species, and cannot be analysed using CSF. Besides that, primate specific lncRNAs rarely show sufficient changes to highlight a sense/non-sense mutations pattern. In addition to CSF, other strategies not relying on evolutionary signatures can be effectively used to predict if a transcript is going to be translated into protein or not. For example, there are dedicated Blast flavours including BlastX and RPS-Blast (Altschul et al., 1990; Marchler-Bauer et al., 2002) that can be used to identify transcripts whose translational product possesses a match in protein databases such as Pfam (Finn et al., 2008) and UniProt (UniProt, 2012). Unfortunately, bioinformatics predictions can easily return mistaken assignments when dealing with ncRNAs closely related to coding mRNAs, and result in some confusion when transferring annotation across species, or within a genome. Such observations may wrongly suggest pseudogenization events or a turnover between proteins and ncRNAs.

Over the last few years other approaches alternative or complementary to RNA-seq have been attempted to generate high-throughput ncRNA annotations. In 2009, Mitchell Guttman and co-workers published the first of a series of analysis that recently came out linking lncRNA detection to histone modifications (Guttman et al., 2009). In this work, the authors pioneered a chromatin-state based method to identify well-defined transcriptional units occurring between known protein-coding genes. Their analysis relied on the observation by (Mikkelsen et al., 2007) that promoters of genes expressed by the RNA polymerase II (Pol II) are signed by trimethylation of lysine 4 of histone H3 (H3K4me3) while the transcribed area is marked by trimethylation of lysine 36 of histone H3 (H3K36me3). Following this observation, the authors did chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq) (Mikkelsen et al., 2007) to generate profiles of chromatin states. This approach revealed thousands of mouse lincRNAs, corresponding to H3K4me3-H3K36me3 chromatin domains and lying outside of protein coding regions. The prediction reliability has been estimated by additional analysis showing that lincRNAs are more conserved than neutrally evolving sequences and that most of experimentally tested loci were found to be expressed (Guttman et al., 2009). An alternative strategy used for ncRNA detection involves a combination of different high-throughput data sources and their integration using bioinformatics (Lu et al., 2011). This approach, named incRNA, relies on a machine learning method and has been applied to the genome-wide identification of *Caenorhabditis elegans* ncRNAs. incRNA combines predicted and experimental data for a total of nine different information sources. These include the expression data coming from various developmental stages and conditions, as well as the GC content, the predictions of RNA secondary structure folding energy, the prediction of evolutionary conserved DNA sequence and secondary structure. The results show how the integration of multiple information sources ends in highly accurate predictions of novel ncRNA genes.

Over the past few years, a number of works reporting a massive quantity of novel ncRNA genes in various species has been published (Cabili et al., 2011; Esteve-Codina

et al., 2011; Kutter et al., 2012; Nam and Bartel, 2012). Such rapid growth has been possible thanks to the contribution and the parallel development of new and ever more sophisticated bioinformatics approaches. In this thesis, I report on two such projects in which I have been involved. The analysis of pig transcriptome and the analysis of GENCODE version 7 lncRNAs, described in chapters 3 and 4 respectively. In both works I curated the sections regarding the evolutionary analysis of lncRNAs. My analysis included the identification of putative lncRNA transcripts, the detection of evolutionarily conserved elements using PhastCons (Siepel et al., 2005), the homology-based annotation of novel lncRNA homologs, the prediction of lncRNA families and the detection of compensatory mutations. Nevertheless, such analyses remain superficial with uncertainties of different type and degree affecting most predictions. For example, the homology search pipeline described in chapters 3 and 4 is not sensitive enough to map fast evolving lncRNAs, hence our limit to play comprehensive evolutionary study. Our lncRNA predictions should be taken with care, not just because they are not experimentally verified, but also because they are far from representing the complete genome-wide lncRNA figure. Fortunately, many lncRNAs are constrained enough to be successfully mapped with our approach, and we succeeded at identifying thousands of novel lncRNA candidates. Thanks to the large amount of data considered in our analyses, we gathered enough evidences to infer general lncRNAs properties, to define families and detect structurally conserved homologs. Both projects reported on chapters 3 and 4 are good examples of how bioinformatics approaches, including ncRNA alignments, can be applied to analyse transcriptomic data. On the short run available transcription data is expected to increase very rapidly, and the necessity to accurately and quickly align ncRNAs is becoming more pressing than ever. In my thesis I tried to address exactly this issue by developing BlastR (cf. Chapter 2), an algorithm suited for accurate ncRNA detection at a moderate CPU cost.

CHAPTER 2: BlastR algorithm for ncRNA search

A major focus of my doctoral thesis has been the improvement of state-of-the-art methods for the homology detection of ncRNAs. To this end I participated and took the lead of the BlastR project. The underlying project idea was to apply the di-nucleotide conservation signal to off-the-shelf Blast packages and check whether this could be beneficial in terms of accuracy. The strategy we adopted consisted first in the generation of an *ad hoc* di-nucleotide substitution matrix estimated on accurate RNA alignments. Such matrix indicates how frequent it is that a certain di-nucleotide mutates into another one. We called this matrix BlosumR for the analogy with the standard BLOSUM (Henikoff and Henikoff, 1992). Next, we recoded both query and target databases into an amino acid like alphabet. This conversion allowed readapting tools like BlastP, i.e. applications normally dedicated to search proteins. Finally, we mounted the BlosumR matrix onto BlastP and verified its performance in detecting ncRNA homologs in a benchmark built on Rfam. The results are encouraging, with BlastR showing to be superior in terms of sensitivity and specificity with respect to competing algorithms. We further investigated the source of the improvement, and we found that di-nucleotides bring only little, although real, accuracy improvement. Remarkably, most of the improvement comes from the use of the BlastP algorithm. These results, the benchmark and the algorithm details are discussed in the paper, published last year in Nucleic Acids Research.

BlastR-fast and accurate database searches for non-coding RNAs.

Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, Beaudoing E, Bucher P, Notredame C.

Nucleic Acids Res. 2011 Sep 1;39(16):6886-95. Epub 2011 May 30.

CHAPTER 3: Analyzing the pig transcriptome

My PhD involved data analysis projects complementary to the methodological part. In this chapter I present a work carried out in collaboration with the Universitat Autònoma de Barcelona and the Institut für Populationsgenetik of Vienna. The aim of the work was to produce and analyze the transcriptome profile of two highly divergent pig breeds using RNA-seq. One result of the study was the identification of thousands of previously unreported pig genes. My contribution to the project had been to help defining the set of putative porcine lncRNA and then to investigate their conservation across mammals. This comparative biology analysis allowed the identification of human homologs located in unannotated regions of the genome, i.e. potentially novel human non-coding genes. This work was published last year in BMC Genomics, and is a typical example where the wealth of high-throughput data comes together new bioinformatics challenges.

Exploring the gonad transcriptome of two extreme male pigs with RNA-seq.
Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Pérez-Enciso M.
BMC Genomics. 2011 Nov 8;12:552.

CHAPTER 4: Analyzing the human lncRNA dataset

This chapter presents a recent publication in Genome Research. In this paper are provided extensive analysis and statistics on the largest available human lncRNA dataset. In the context of this work expressly focused on lncRNA characterization, I took care of the lncRNA conservation part. My targets were to search homologous genes across mammalian species, measuring the conservation of promoter, exonic and intronic regions, and detecting lncRNA families. The results we got indicate that a considerable fraction of lncRNAs seem to be primate specific, thus suggesting that these genes have a very high turnover if compared with proteins. Moreover we confirmed that exonic sequences seem to be more constrained than neutrally evolving sequences, but less than protein coding exons. On the other hand, lncRNAs promoters show a conservation level similar to the ones of proteins. Additionally, in this work we sought human lncRNA families by using blastClust (Altschul et al., 1990), a standard clustering algorithm. Some domains of the lncRNA clusters we detected embed degraded versions of repeat elements. Remarkably, compensatory mutation might arise from these regions, possibly to preserve functional secondary structures.

The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R.

Genome Res. 2012 Sep;22(9):1775-89.

DISCUSSION

ncRNA functional characterization is a rapidly expanding research area. In the past few years, it has become clear that the majority of the transcripts in cells are more than mere intermediates between the hereditary information encoded in DNA and the mechanical operative component represented by proteins. Indeed, it appears that numerous transcripts may not be translated at all while still being involved in critical biological functions such as cell differentiation and chromatin remodelling. Taking together 15 human cell lines, the cumulative coverage of transcribed regions is ~62% and ~75% of the whole human genome for processed and primary transcripts, respectively (Djebali et al., 2012). This "pervasive transcription" is strikingly high, especially when considering that a mere 3% of the human genome codes for protein coding exons. (ENCODE Project Consortium, 2012). Numerous novel, previously uncharacterized RNA species have been recently detected. A sizeable fraction of them is defined as lncRNA, i.e. molecules longer than 200 nucleotides that do not show any coding potential. Some of these molecules are spliced, capped, differentially expressed in tissues/cells or developmental stages and tend to be more conserved across species than would result from neutral evolution. For these reasons and because of the increasing number of transcripts whose function was experimentally validated, it is believed that many of these new ncRNAs belong to an important, almost unexplored class of regulatory elements. Thanks to ongoing improvements in sequencing technologies it has become possible to collect a significant amount of these uncharacterized transcripts. The latest generation of sequencer make it possible to do large scale sequencing of entire transcriptomes. This technique, known as RNA-seq has already had a dramatic impact on our perception of the human transcription landscape (Wang et al., 2008a; Djebali et al., 2012). Similar studies have been carried out in a number of genetic model organisms including rodents (Mortazavi et al., 2008; Kutter et al., 2012), plants (Eveland et al., 2008), insects (Graveley et al., 2011), worms (Hillier et al., 2009) and yeasts (Nagalakshmi et al., 2008). In (Shendure, 2008) the author argues that RNA-seq

represents the most promising technology for transcriptome research. The RNA-seq main strength is the potentially unlimited dynamic range it offers, returning better sensitivities than microarrays without the need of a priori speculations regarding the genomic loci being transcribed (Morozova et al., 2009). If the pace of scientific progress is maintained and if costs keep dropping, one can reasonably expect this technology to rapidly become a key component of personalized medicine, especially when considering the new venues of development that are currently being considered (Auer and Doerge, 2010; Malone and Oliver, 2011).

Nowadays there exist new directions and emerging applications of NGS that are relevant for ncRNA research and worth commenting. Two promising RNA-seq developments regard the direct RNA sequencing and the study of tiny RNA quantities. The first technique is about the direct sequencing of RNA samples without the need to retro-transcribe it into cDNA. Skipping the conversion to cDNA has many advantages. These include avoiding the generation of spurious second-strand cDNAs, the template switching, the nucleotide composition bias caused by random hexamer priming, the cDNA synthesis in primer-independent manner, and prevention of the use of the reverse transcriptase which is known to have lower fidelity than other polymerases (Roberts et al., 1989; Chen and Patton, 2001; Hansen et al., 2010; Ozsolak and Milos, 2011). Furthermore, direct RNA sequencing does not involve any amplification steps, thus it is not affected by PCR amplification biases (Oyola et al., 2012). Direct RNA sequencing is a technology still under development, and facing many challenges like the generation of deeper sequencing data, and lowering the frequency of sequencing errors (Ozsolak and Milos, 2011). The other emerging technology applied to RNA-seq is the profiling of low-quantity RNA samples. This technology is especially suited when exploring the transcriptome of specific groups of cells. Typical biological samples include body fluids and tissues made of a combination of different cell populations. To choose the cell type of interest and then isolate the RNA one can use different tools, including the flow-assisted cell sorting (FACS), the laser-capture micro dissection (Simone et al., 1998), the serial dilution, specialized micro fluidic apparatus (Marcy et al., 2007) and

micromanipulation (Ozsolak and Milos, 2011). Even so, in spite of the possibility to extract small RNA quantities from pure cell populations, one major obstacle to transcriptome profiling is the impossibility to perform high-throughput sequencing of tiny RNA amounts (e.g. picograms). This limit is critical whenever the amount of material is a limiting factor (forensics, stem cell biology, biopsy analysis in cancer). A wealth of approaches meant to address this issue has recently been reported. These include both sequencing based approaches such as nanoCAGE (Plessy et al., 2010), smsDGE (Lipson et al., 2009) and LQ-DGE (Ozsolak et al., 2010b), and hybridization approaches like the NanoString nCounter (Geiss et al., 2008) and the Fluidigim systems (Byrne et al., 2009; Helzer et al., 2009). Unfortunately, none of these technologies is yet mature enough for large-scale analysis and further progresses remain needed to achieve comprehensive high-quality transcriptome analyses equally informative across the entire transcriptome dynamic range. Another promising NGS application, named RNA CaptureSeq and recently reported on Nature Biotechnology (Mercer et al., 2011), is able to reach unprecedented sequencing depth. RNA CaptureSeq is inspired from exome sequencing techniques and relies on the use of tiling arrays in order to enrich the population of RNAs one wants to sequence. This enrichment step allows a sequencing depth that would be impossible when dealing with the full transcriptome. Although RNA CaptureSeq is not suited to generate full transcriptome profile, it can be used to target specific genomic sites and detect transcript isoforms expressed at very low abundance. As shown in (Mercer et al., 2011) RNA CaptureSeq can be used to fuel the detection of ncRNAs that are missed by genome-wide standard RNA sequencing.

From a functional perspective, a lot remains to be done for the characterization of ncRNA analysis. Comparative studies offer a very efficient way of prioritizing analysis. They can be used to predict function by homology, assess phylogenetic relationships, detect functional motifs or classify related molecules in order to identify families. A main challenge when tackling ncRNA comparisons results from the remarkable variability of traits and functions. Considering sizes only, ncRNA molecules can be as short as a miRNA (~22nt) and up to ~17kb long in the case of Xist (Brown et al., 1992).

Another source of difficulty when comparing ncRNAs is that most of these genes have poorly conserved sequences. Such diversity challenges our ability to compare, classify and search with conventional alignment tools. In addition ncRNA genes have no equivalent of codon bias and Open Reading Frames (ORFs) that help powering the statistical component of machine learning approaches when doing protein prediction (Rivas and Eddy, 2000). The strongest signal contained by RNA sequences are usually evolutionarily conserved secondary structures. Many efficient algorithms exist that are able to predict potential structures using MFE or SCFG computations. Unfortunately, these predictions ignore the contribution of the environment and are not always accurate enough to significantly improve alignment accuracy and homology modelling. Emerging technologies allowing the high-throughput generation of experimentally derived secondary structures (Kertesz et al., 2010) will hopefully help addressing this problem. Unfortunately, taking into account secondary structures while comparing sequences is a challenging procedure, too intensive from a computing point of view to be practical in the most common circumstances (Dowell and Eddy, 2006). This makes it is difficult to compare mono-exonic genes while taking the secondary structure into account, and totally impossible when the transcripts are multi-exonic (i.e. the secondary structures are interrupted by introns). This obvious need for new strategies for fast and accurate comparison has motivated our BlastR project. BlastR is a special adaptation of BlastP (Altschul et al., 1990) that takes advantage of the evolutionary signal contained in di-nucleotides. Starting from high quality non-coding RNA alignments we captured the frequencies at which di-nucleotides mutate into others and used this information to set a log odd matrix that BlastP can use as substitution scoring scheme. Although the di-nucleotide signal is weak, it does exist and improves the search accuracy by a small amount. This study also shows that BlastP seems to be a more sensitive, albeit slower, algorithm than BlastN (Altschul et al., 1990) when searching ncRNAs. This result opens to the possibility to readapt tools developed for proteins to ncRNA research. Such options are realistic when doing genome wide search of sequences with no known secondary structure. The collection of a set of homologues, diverse enough to show co-

variation, is a compulsory pre-requisite for the estimation of the consensus structures needed by CM based methods, like infernal. A method like BlastR would be ideal in combination with accurate and computationally demanding model-based algorithms. For instance, BlastR could be used to gather the homologs needed to train the models, and then to filter the target database so as to reduce the search space.

We have shown (Esteve-Codina et al., 2011; Derrien et al., 2012b) that Blast can be effectively used for lncRNA homolog prediction, in combination with splicing informed heuristics such as exonerate (Slater and Birney, 2005) or GeneWise (Birney et al., 2004). This strategy is not new, and similar approaches have already been used for the discovery of protein coding homologs (Eyras et al., 2005; Mariotti and Guigo, 2010; Vieira and Rozas, 2011). As one would expect, homology based RNA searches are severely limited by our capacity to align distant homologues. For instance, when searching the human lncRNA complement against mammalian genomes (Derrien et al., 2012b) or when using an estimated pig complement (Esteve-Codina et al., 2011), we only managed to find, beyond primates, less than 50% of the query genes across cow, mouse or dog. This result may reflect a high turnover, but the conservation/disappearance patterns, poorly correlated to phylogenetic history, are most likely indicative of a limited detection capacity. Other confounding factors include misassembled or partially sequenced genomes. Additional analysis would be needed to validate the Blast/exonerate mapping approach. At this stage, it is therefore impossible, without further experiments, to establish whether the lncRNA queries that failed the mapping are really absent in the target species or undetected. Another issue with homology based analysis is its tendency to miss edge exons, e.g. the first or the last one. Such bias results from the exonerate step, where exons at transcripts extremities might be excluded or just partially included in the alignment. Those exons will then appear to be lost or truncated in the output transcript model. In this context, high quality templates, as the GENCODE queries used in (Derrien et al., 2012b), offer better chances to return precise annotations. Our Blast/exonerate mapping procedure has

returned thousands of putative new lncRNA genes in (Esteve-Codina et al., 2011) and (Derrien et al., 2012b). These annotations could be used to further investigate lncRNA evolution or to train models (i.e. CMs). We also found that a sizeable fraction of the human lncRNAs seems to be primate specific (Derrien et al., 2012b). Our result is in agreement with a recently published study (Kutter et al., 2012) where the authors identified lncRNAs expressed in rodents' adult liver, and then compared the expression of the orthologous genomic regions. In the paper it is shown that loss of lncRNA transcription among rodents is associated with loss of sequence constraints and that many lncRNA genes seems to be species or lineage specific. Another application of our Blast/exonerate mapping analysis is the possibility to identify novel human lncRNA genes candidates by using non-human templates as query (Esteve-Codina et al., 2011). As shown in figure 2 of the paper, there are 131 pig lncRNAs mapping to unannotated regions of the human genome. This result suggests that although human is probably one of the most extensively annotated higher-eukaryote, extra improvements might be achieved using data gathered in other non-model organisms. In (Derrien et al., 2012b) we also extended the lncRNA conservation study to a multiple genome alignments strategy based on PhastCons conservation scores. The analysis, reported in figure 4 of the paper, is in agreement with previous reports (Guttman et al., 2009; Orom et al., 2010), and confirms that lncRNAs sequences are less constrained than those of protein coding genes. Remarkably, we show that the distribution of lncRNA exons conservation is bimodal, with a fraction substantially approximate to ancestral repeats, and another group appreciably shifted toward the protein coding set. This indicates that some lncRNA are under a selection as strong as proteins and suggests that a sizeable fraction of lncRNA genes are probably functional. The lncRNA portion having a mutation rate almost indistinguishable from repeats suggests that at least some lncRNAs (close to a third) might be transcriptional noise.

As shown in this thesis, despite the difficulties encountered when comparing ncRNAs, the homology search of ncRNAs can be operatively used to detect new genes. New and ever more sophisticated algorithms will help addressing the challenges brought by new

technologies. The ultimate goal is the creation of thorough transcriptome annotations and unbiased expression profiling of each individual transcript. It is still early to tell, but if they live up to their promises, the discovery of this new large class of RNAs may well define one of the turning points of modern biology. During my thesis I had the opportunity to get into this world and move ahead a few steps, and these have been so far extremely encouraging.

CONCLUSION

From chapter 2: BlastR algorithm for ncRNA search

- The proposed approach, named BlastR, aims to improve ncRNA homology search accuracy at a reasonable computation cost. The algorithm takes advantage of di-nucleotides conservation information in combination with BlastP, a tool normally dedicated to search proteins. BlastR shows increased specificity and sensitivity when scanning ncRNA databases.
- In BlastR the accuracy improvement comes at the cost of a reduced speed. The accuracy and the slower runtime could be both effects of the alignment of higher numbers of high scoring segment pairs (HSPs). However we demonstrated that the HSP number does not explain alone the CPU costs of tested Blast approaches. BlastN and BlastP are pretty different algorithms, based on different parameterizations. We argue that is the combination of algorithm features that makes BlastP a slower but more accurate tool for ncRNA search.
- In the paper we show how the di-nucleotide conservation signal, although small, is strong enough to improve BlastR sensitivity. Remarkably, the BlastP algorithm itself used in combination with BlastN-like substitution scores is substantially more accurate than BlastN. The best mix of accuracy and efficiency is reached using BlastR, i.e. di-nucleotides on a BlastP engine.
- BlastR accepts as input simple RNA sequences, and can therefore be used to search the ncRNA homologues needed to train CMs, like the infernal ones. Moreover, since BlastR is orders of magnitude faster than infernal, it could be used with relaxed parameters to pre-filter the target databases and reduce infernal's search space.
- Although in our benchmark BlastR proved to outperform other Blast alternatives, a thorough genome-wide analysis is still missing. We plan to assess the quality of different homology search algorithms and different

parameterizations by overlapping whole-genomes mapping with RNA-seq expression data.

From chapter 3: Analyzing the pig transcriptome

- RNA-seq was used to characterize the poly(A) RNA fraction of two phenotypically extreme pigs.
- Thanks to protein coding potential and comparative genomics analysis we were able to annotate 2047 new putative porcine lncRNAs.
- We found 469 pig lncRNAs having significant homology with human sequences and out of them 131 mapping to unannotated regions of the genome. This result indicates that homology search allows the prediction of new human lncRNA genes using non-human genes as template.

From chapter 4: Analyzing the human lncRNA dataset

- This work reports the analyses of 14880 human lncRNAs from GENCODE. In agreement with previous reports, we show that lncRNAs have stronger purifying selection than neutrally evolving sequences but are less constrained than protein coding genes.
- We show that lncRNAs promoters have levels of conservation close to protein coding genes. Such evolutionary information could be used to improve lncRNA homology screenings.
- A Blast/exonerate screening was used to map the 14880 human lncRNAs against 18 mammal species. This analysis showed that about 30% of lncRNAs seems to be primate specific, and less than 1% seems to be specific to the human lineage. About 1% is detected in all the considered species, and this might represent a set of lncRNAs undertaking essential biological functions.

- By using an all-against-all clustering algorithm we were able to identify 194 human lncRNAs families, although a great part of them includes just two members. Many clusters embed domains corresponding to degraded versions of repeat elements. The detection of compensatory mutations within repeat domains suggests that the secondary structure of these modules is maintained across the families. This observation suggests that repeat elements within lncRNAs are not neutrally decaying, but are rather re-used as functional domains.

BIBLIOGRAFY

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651-1656.

Ahmadi A, Behm A, Honnalli N, Li C, Weng L, Xie X (2012) Hobbes: optimized gram-based methods for efficient read alignment. Nucleic Acids Res 40:e41.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403-410.

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS (2011) lncRNAdb: a reference database for long noncoding RNAs. Nucleic Acids Res 39:D146-151.

Ameur A, Wetterbom A, Feuk L, Gyllensten U (2010) Global and unbiased detection of splice junctions from RNA-seq data. Genome Biol 11:R34.

Aparicio S et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 297:1301-1310.

Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405-416.

Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. Biochimie 84:775-790.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 33:D562-566.

Bauer Rp, Rother Kp, Moor Pp, Reinert Kp, Steinke Tp, Bujnicki Jp, Preissner Rp (2009) Fast Structural Alignment of Biomolecules Using a Hash Table, N-Grams and String Descriptors\par. Algorithms\par 2\par:692\par-709\par.

Bengert P, Dandekar T (2004) Riboswitch finder--a tool for identification of riboswitch RNAs. Nucleic Acids Res 32:W154-159.

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z (2005) Identification of hundreds of conserved and nonconserved human microRNAs. Nat Genet 37:766-770.

Berezikov E, van Tetering G, Verheul M, van de Belt J, van Laake L, Vos J, Verloop R, van de Wetering M, Guryev V, Takada S, van Zonneveld AJ, Mano H, Plasterk R, Cuppen E (2006) Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res 16:1289-1298.

Berger MF et al. (2010) Integrative analysis of the melanoma transcriptome. Genome Res 20:413-427.

Bernhart SH, Hofacker IL (2009) From consensus structure prediction to RNA gene finding. Brief Funct Genomic Proteomic 8:461-471.

Bernhart SH, Hofacker IL, Stadler PF (2006) Local RNA base pairing probabilities in large sequences. Bioinformatics 22:614-615.

Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M (2004) Global identification of human transcribed sequences with genome tiling arrays. Science 306:2242-2246.

Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14:988-995.

Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST--database for "expressed sequence tags". Nat Genet 4:332-333.

Boguski MS, Tolstoshev CM, Bassett DE, Jr. (1994) Gene discovery in dbEST. Science 265:1993-1994.

Bonaldo MF, Lennon G, Soares MB (1996) Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res 6:791-806.

Braidotti G, Baubec T, Pauler F, Seidl C, Smrzka O, Stricker S, Yotova I, Barlow DP (2004) The Air noncoding RNA: an imprinted cis-silencing transcript. Cold Spring Harb Symp Quant Biol 69:55-66.

Bremges A, Schirmer S, Giegerich R (2010) Fine-tuning structural RNA alignments in the twilight zone. BMC Bioinformatics 11:222.

Brenner S et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630-634.

Brennicke A, Marchfelder A, Binder S (1999) RNA editing. FEMS Microbiol Rev 23:297-316.

Brown CJ, Hendrich BD, Rupert JL, Lafreniere RG, Xing Y, Lawrence J, Willard HF (1992) The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. Cell 71:527-542.

Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, Beaudoing E, Bucher P, Notredame C (2011) BlastR--fast and accurate database searches for non-coding RNAs. Nucleic Acids Res 39:6886-6895.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Res 18:810-820.

Byrne JA, Nguyen HN, Reijo Pera RA (2009) Enhanced generation of induced pluripotent stem cells from a subpopulation of human fibroblasts. PLoS One 4:e7118.

Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25:1915-1927.

Capriotti E, Marti-Renom MA (2008) RNA structure alignment by a unit-vector approach. Bioinformatics 24:i112-118.

Capriotti E, Marti-Renom MA (2010) Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. BMC Bioinformatics 11:322.

Carninci P et al. (2005) The transcriptional landscape of the mammalian genome. Science 309:1559-1563.
Casneuf T, Van de Peer Y, Huber W (2007) In situ analysis of cross-hybridisation on microarrays and the inference of expression correlation. BMC Bioinformatics 8:461.
Cesana M, Cacchiarelli D, Legnini I, Santini T, Sthandier O, Chinappi M, Tramontano A, Bozzoni I (2011) A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. Cell 147:358-369.
Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES (2007) Distinguishing protein-coding and noncoding genes in the human genome. Proc Natl Acad Sci U S A 104:19428-19433.
Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, Ponting CP, Stadler PF, Morris KV, Morillon A, Rozowsky JS, Gerstein MB, Wahlestedt C, Hayashizaki Y, Carninci P, Gingeras TR, Mattick JS (2011) The reality of pervasive transcription. PLoS Biol 9:e1000625; discussion e1001102.
Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DT, Kolle G, Grimmond SM (2009) RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data. Bioinformatics 25:2615-2616.
Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat Methods 5:613-619.
Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845-1848.
Costa V, Angelini C, De Feis I, Ciccodicola A (2010) Uncovering the complexity of transcriptomes with RNA-Seq. J Biomed Biotechnol 2010:853916.
Cox WG, Beaudet MP, Agnew JY, Ruth JL (2004) Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. Anal Biochem 331:243-254.
Crick FH (1958) On protein synthesis. Symp Soc Exp Biol 12:138-163.
Chang YF, Huang YL, Lu CL (2008) SARSA: a web tool for structural alignment of RNA using a structural alphabet. Nucleic Acids Res 36:W19-24.
Chen D, Patton JT (2001) Reverse transcriptase adds nontemplated nucleotides to cDNAs during 5'-RACE and primer extension. Biotechniques 30:574-580, 582.
Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149-1154.
Deigan KE, Li TW, Mathews DH, Weeks KM (2009) Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci U S A 106:97-102.

Demmer RT, Behle JH, Wolf DL, Handfield M, Kebschull M, Celenti R, Pavlidis P, Papapanou PN (2008) Transcriptomes in healthy and diseased gingival tissues. J Periodontol 79:2112-2124.
Derrien T, Estelle J, Marco Sola S, Knowles DG, Raineri E, Guigo R, Ribeca P (2012a) Fast computation and applications of genome mappability. PLoS One 7:e30377.
Derrien T et al. (2012b) The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res 22:1775-1789.
Desikan R, S AH-M, Hancock JT, Neill SJ (2001) Regulation of the Arabidopsis transcriptome by oxidative stress. Plant Physiol 127:159-172.
Dias Neto E et al. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci U S A 97:3491-3496.
Dima RI, Hyeon C, Thirumalai D (2005) Extracting stacking interaction parameters for RNA from the data set of native structures. J Mol Biol 347:53-69.
Dinger ME, Pang KC, Mercer TR, Crowe ML, Grimmond SM, Mattick JS (2009) NRED: a database of long noncoding RNA expression. Nucleic Acids Res 37:D122-126.
Dinger ME, Amaral PP, Mercer TR, Pang KC, Bruce SJ, Gardiner BB, Askarian-Amiri ME, Ru K, Solda G, Simons C, Sunkin SM, Crowe ML, Grimmond SM, Perkins AC, Mattick JS (2008) Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. Genome Res 18:1433-1445.
Djebali S et al. (2012) Landscape of transcription in human cells. Nature 489:101-108.
Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics 22:e90-98.
Dobbin KK, Kawasaki ES, Petersen DW, Simon RM (2005) Characterizing dye bias in microarray experiments. Bioinformatics 21:2430-2437.
Dombkowski AA, Thibodeau BJ, Starcevic SL, Novak RF (2004) Gene-specific dye bias in microarray reference designs. FEBS Lett 560:120-124.
Doshi KJ, Cannone JJ, Cobaugh CW, Gutell RR (2004) Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. BMC Bioinformatics 5:105.
Dowell RD, Eddy SR (2004) Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. BMC Bioinformatics 5:71.
Dowell RD, Eddy SR (2006) Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. BMC Bioinformatics 7:400.
Dror O, Nussinov R, Wolfson H (2005) ARTS: alignment of RNA tertiary structures. Bioinformatics 21 Suppl 2:ii47-53.
Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis. Probabilistic models of proteins and nucleic acids: Cambridge University Press
Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755-763.
Eddy SR (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. BMC Bioinformatics 3:18.

Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res 22:2079-2088.
Eklund AC, Turner LR, Chen P, Jensen RV, deFeo G, Kopf-Sill AR, Szallasi Z (2006) Replacing cRNA targets with cDNA reduces microarray cross-hybridization. Nat Biotechnol 24:1071-1073.
ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799-816.
ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57-74.
Esteve-Codina A, Kofler R, Palmieri N, Bussotti G, Notredame C, Perez-Enciso M (2011) Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. BMC Genomics 12:552.
Eveland AL, McCarty DR, Koch KE (2008) Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. Plant Physiol 146:32-44.
Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8:186-194.
Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.
Eyras E, Reymond A, Castelo R, Bye JM, Camara F, Flicek P, Huckle EJ, Parra G, Shteynberg DD, Wyss C, Rogers J, Antonarakis SE, Birney E, Guigo R, Brent MR (2005) Gene finding in the chicken genome. BMC Bioinformatics 6:131.
Farazi TA, Juranek SA, Tuschl T (2008) The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. Development 135:1201-1214.
Ferre F, Ponty Y, Lorenz WA, Clote P (2007) DIAL: a web server for the pairwise alignment of two RNA three-dimensional structures using nucleotide, dihedral angle and base-pairing similarities. Nucleic Acids Res 35:W659-668.
Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res 36:D281-288.
Flicek P et al. (2012) Ensembl 2012. Nucleic Acids Res 40:D84-90.
Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA (2011) Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome. BMC Genomics 12:202.
Freyhult EK, Bollback JP, Gardner PP (2007) Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. Genome Res 17:117-125.
Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P (2009) Estimating accuracy of RNA-Seq and microarrays with proteomics. BMC Genomics 10:161.
Furtig B, Richter C, Wohnert J, Schwalbe H (2003) NMR spectroscopy of RNA. Chembiochem 4:936-962.

Gardner PP, Giegerich R (2004) A comprehensive comparison of comparative RNA structure prediction approaches. BMC Bioinformatics 5:140.
Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136-140.
Garg R, Patel RK, Tyagi AK, Jain M (2011) De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. DNA Res 18:53-63.
Geiss GK et al. (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 26:317-325.
Gerhard DS et al. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res 14:2121-2127.
Gorodkin J, Heyer LJ, Stormo GD (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res 25:3724-3732.
Gowda M, Jantasuriyarat C, Dean RA, Wang GL (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. Plant Physiol 134:890-897.
Grabherr MG et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644-652.
Graveley BR et al. (2011) The developmental transcriptome of Drosophila melanogaster. Nature 471:473-479.
Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31:439-441.
Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140-144.
Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33:D121-124.
Guerra-Assuncao JA, Enright AJ (2012) Large-scale analysis of microRNA evolution. BMC Genomics 13:218.
Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503-510.
Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223-227.

Halbeisen RE, Gerber AP (2009) Stress-Dependent Coordination of Transcriptome and Translatome in Yeast. PLoS Biol 7:e105.
Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38:e131.
Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2:495-502.
Harrow J, Denoeud F, Frankish A, Reymond A, Chen CK, Chrast J, Lagarde J, Gilbert JG, Storey R, Swarbreck D, Rossier C, Ucla C, Hubbard T, Antonarakis SE, Guigo R (2006) GENCODE: producing a reference annotation for ENCODE. Genome Biol 7 Suppl 1:S4 1-9.
Harrow J et al. (2012) GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res 22:1760-1774.
Havana team In.
Havgaard JH, Lyngso RB, Stormo GD, Gorodkin J (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics 21:1815-1824.
He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. Science 322:1855-1857.
Helzer KT, Barnes HE, Day L, Harvey J, Billings PR, Forsyth A (2009) Circulating tumor cells are transcriptionally similar to the primary tumor in a murine prostate model. Cancer Res 69:7860-7866.
Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915-10919.
Herschlag D (1995) RNA chaperones and the RNA folding problem. J Biol Chem 270:20871-20874.
Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH (2009) Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res 19:657-666.
Hofacker IL (2003) Vienna RNA secondary structure server. Nucleic Acids Res 31:3429-3431.
Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J Mol Biol 319:1059-1066.
Hofacker IL, Bernhart SH, Stadler PF (2004) Alignment of RNA base pairing probability matrices. Bioinformatics 20:2222-2227.
Hofacker IL, Fontana W., Stadler P.F., Bonhoeffer S., Tacker M., P. S (1994) Fast Folding and Comparison of RNA Secondary Structures. Monatshefte f Chemie 125:167-188.
Holmes I (2004) A probabilistic model for the evolution of RNA structure. BMC Bioinformatics 5:166.
Holmes I (2005) Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics 6:73.
Huang R, Jaritz M, Guenzl P, Vlatkovic I, Sommer A, Tamir IM, Marks H, Klampfl T, Kralovics R, Stunnenberg HG, Barlow DP, Pauler FM (2011) An RNA-Seq

strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. PLoS One 6:e27288.
Jager M, Ott CE, Grunhagen J, Hecht J, Schell H, Mundlos S, Duda GN, Robinson PN, Lienau J (2011) Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing. BMC Genomics 12:158.
Kaikkonen MU, Lam MT, Glass CK (2011) Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res 90:430-440.
Kapranov P, Willingham AT, Gingeras TR (2007a) Genome-wide transcription and the implications for genomic organization. Nat Rev Genet 8:413-423.
Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916-919.
Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR (2005) Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res 15:987-997.
Kapranov P et al. (2007b) RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484-1488.
Katayama S et al. (2005) Antisense transcription in the mammalian transcriptome. Science 309:1564-1566.
Keiler KC (2008) Biology of trans-translation. Annu Rev Microbiol 62:133-151.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12:996-1006.
Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E (2010) Genome-wide measurement of RNA secondary structure in yeast. Nature 467:103-107.
Kirillova S, Tosatto SC, Carugo O (2010) FRASS: the web-server for RNA structural comparison. BMC Bioinformatics 11:327.
Klein RJ, Eddy SR (2003) RSEARCH: finding homologs of single structured RNA sequences. BMC Bioinformatics 4:44.
Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 31:3423-3428.
Kodzius R, Kojima M, Nishiyori H, Nakamura M, Fukuda S, Tagami M, Sasaki D, Imamura K, Kai C, Harbers M, Hayashizaki Y, Carninci P (2006) CAGE: cap analysis of gene expression. Nat Methods 3:211-222.
Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, Odom DT, Marques AC (2012) Rapid Turnover of Long Noncoding RNAs and the Evolution of Gene Expression. PLoS Genet 8:e1002841.
Lander ES et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.
Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.
Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75:843-854.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.
Li H, Ruan J, Durbin R (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851-1858.
Li R, Li Y, Kristiansen K, Wang J (2008b) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713-714.
Li W, Feng J, Jiang T (2011) IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. J Comput Biol 18:1693-1707.
Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H, Skogerbo G, Wu Z, Zhao Y (2011) Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic Acids Res 39:3864-3878.
Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24:2431-2437.
Lin MF, Carlson JW, Crosby MA, Matthews BB, Yu C, Park S, Wan KH, Schroeder AJ, Gramates LS, St Pierre SE, Roark M, Wiley KL, Jr., Kulathinal RJ, Zhang P, Myrick KV, Antone JV, Celniker SE, Gelbart WM, Kellis M (2007) Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res 17:1823-1836.
Lindgreen S, Gardner PP, Krogh A (2007) MASTR: multiple alignment and structure prediction of non-coding RNAs using simulated annealing. Bioinformatics 23:3304-3311.
Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435-1441.
Lipson D, Raz T, Kieu A, Jones DR, Giladi E, Thayer E, Thompson JF, Letovsky S, Milos P, Causey M (2009) Quantification of the yeast transcriptome by single-molecule sequencing. Nat Biotechnol 27:652-658.
Loveland JE, Gilbert JG, Griffiths E, Harrow JL (2012) Community gene annotation in practice. Database (Oxford) 2012:bas009.
Lu ZJ, Turner DH, Mathews DH (2006) A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. Nucleic Acids Res 34:4912-4924.
Lu ZJ, Gloor JW, Mathews DH (2009) Improved RNA secondary structure prediction by maximizing expected pair accuracy. RNA 15:1805-1813.
Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, Kato M, Miller DM, Slack F, Snyder M, Waterston RH, Reinke V, Gerstein MB (2011) Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res 21:276-285.
Luck R, Graf S, Steger G (1999) ConStruct: a tool for thermodynamic controlled prediction of conserved secondary structure. Nucleic Acids Res 27:4208-4217.
Malone JH, Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol 9:34.

Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ (2010) FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat Methods 7:130-132.
Mamidala P, Wijeratne AJ, Wijeratne S, Kornacker K, Sudhamalla B, Rivera-Vega LJ, Hoelmer A, Meulia T, Jones SC, Mittapalli O (2012) RNA-Seq and molecular docking reveal multi-level pesticide resistance in the bed bug. BMC Genomics 13:6.
Mandal M, Breaker RR (2004) Gene regulation by riboswitches. Nat Rev Mol Cell Biol 5:451-463.
Marcy Y, Ouverney C, Bik EM, Losekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, Quake SR (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A 104:11889-11894.
Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30:281-283.
Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509-1517.
Mariotti M, Guigo R (2010) Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. Bioinformatics 26:2656-2663.
Martin-Magniette ML, Aubert J, Cabannes E, Daudin JJ (2005) Evaluation of the gene-specific dye bias in cDNA microarray experiments. Bioinformatics 21:1995-2000.
Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12:671-682.
Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, Govindarajan KR, Tong Y, Wu YL, Lam SH, Yang H, Ruan Y, Korzh V, Gong Z, Liu ET, Lufkin T (2005) Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genet 1:260-276.
Mathews DH (2005) Predicting a set of minimal free energy RNA secondary structures common to two sequences. Bioinformatics 21:2246-2253.
Mathews DH, Turner DH (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317:191-203.
Mathews DH, Turner DH, Zuker M (2007) RNA secondary structure prediction. Curr Protoc Nucleic Acid Chem Chapter 11:Unit 11 12.
Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288:911-940.
Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. Proc Natl Acad Sci U S A 101:7287-7292.

Matsumura H, Ito A, Saitoh H, Winter P, Kahl G, Reuter M, Kruger DH, Terauchi R (2005) SuperSAGE. Cell Microbiol 7:11-18.
Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep 2:986-991.
Mattick JS (2009) The genetic signatures of noncoding RNAs. PLoS Genet 5:e1000459.
Mattick JS, Gagen MJ (2001) The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. Mol Biol Evol 18:1611-1630.
McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105-1119.
Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS (2008) Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci U S A 105:716-721.
Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL (2011) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 30:99-104.
Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31-46.
Mikkelsen TS et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553-560.
Missal K, Rose D, Stadler PF (2005) Non-coding RNAs in Ciona intestinalis. Bioinformatics 21 Suppl 2:ii77-78.
Morozova O, Hirst M, Marra MA (2009) Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet 10:135-151.
Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621-628.
Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344-1349.
Nagaraj SH, Gasser RB, Ranganathan S (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 8:6-21.
Nam JW, Bartel D (2012) Long non-coding RNAs in C. elegans. Genome Res.
Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25:1335-1337.
Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu ET, Ruan Y (2005) Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nat Methods 2:105-111.
Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7:521-527.
Notredame C, O'Brien EA, Higgins DG (1997) RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res 25:4570-4580.

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205-217.
Novikova IV, Hennelly SP, Sanbonmatsu KY (2012) Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. Nucleic Acids Res 40:5034-5051.
Ogawa Y, Sun BK, Lee JT (2008) Intersection of the RNA interference and X-inactivation pathways. Science 320:1336-1341.
Okoniewski MJ, Miller CJ (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7:276.
Olivarius S, Plessy C, Carninci P (2009) High-throughput verification of transcriptional starting sites by Deep-RACE. Biotechniques 46:130-132.
Orom UA, Derrien T, Beringer M, Gumireddy K, Gardini A, Bussotti G, Lai F, Zytnicki M, Notredame C, Huang Q, Guigo R, Shiekhattar R (2010) Long noncoding RNAs with enhancer-like function in human cells. Cell 143:46-58.
Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4:14.
Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA (2012) Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics 13:1.
Ozsolak F, Milos PM (2011) RNA sequencing: advances, challenges and opportunities. Nat Rev Genet 12:87-98.
Ozsolak F, Goren A, Gymrek M, Guttman M, Regev A, Bernstein BE, Milos PM (2010a) Digital transcriptome profiling from attomole-level RNA samples. Genome Res 20:519-525.
Ozsolak F, Ting DT, Wittner BS, Brannigan BW, Paul S, Bardeesy N, Ramaswamy S, Milos PM, Haber DA (2010b) Amplification-free digital gene expression profiling from minute cell quantities. Nat Methods 7:619-621.
Palanisamy N et al. (2010) Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med 16:793-798.
Palmieri N, Nolte V, Suvorov A, Kosiol C, Schlötterer C (2012) Evaluation of Different Reference Based Annotation Strategies Using RNA-Seq – A Case Study in Drososphila pseudoobscura. PLoS One.
Pan T, Sosnick T (2006) RNA folding during transcription. Annu Rev Biophys Biomol Struct 35:161-175.
Pang KC, Frith MC, Mattick JS (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. Trends Genet 22:1-5.
Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A (2009) Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res 37:e123.
Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, Lilja P, Mukherjee G, Oezcimen A, Rayner T, Rocca-Serra P, Sharma A, Sansone S, Brazma A (2005)

ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 33:D553-555.
Plessy C et al. (2010) Linking promoters to functional transcripts in small samples with nanoCAGE and CAGEscan. Nat Methods 7:528-534.
Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 37:D32-36.
Pruitt KD, Tatusova T, Brown GR, Maglott DR (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130-135.
Rajagopalan R, Vaucheret H, Trejo J, Bartel DP (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. Genes Dev 20:3407-3425.
Ravasi T, Suzuki H, Pang KC, Katayama S, Furuno M, Okunishi R, Fukuda S, Ru K, Frith MC, Gongora MM, Grimmond SM, Hume DA, Hayashizaki Y, Mattick JS (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. Genome Res 16:11-19.
Ravindran PP, Heroux A, Ye JD (2011) Improvement of the crystallizability and expression of an RNA crystallization chaperone. J Biochem 150:535-543.
Reeder J, Giegerich R (2005) Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. Bioinformatics 21:3516-3523.
Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. Annu Rev Biochem 81:145-166.
Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, Brugmann SA, Goodnough LH, Helms JA, Farnham PJ, Segal E, Chang HY (2007) Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. Cell 129:1311-1323.
Rivas E, Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16:583-605.
Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L (2011) Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol 12:R22.
Roberts JD, Preston BD, Johnston LA, Soni A, Loeb LA, Kunkel TA (1989) Fidelity of two retroviral reverse transcriptases during DNA-dependent DNA synthesis in vitro. Mol Cell Biol 9:469-476.
Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A (2004) Identification of mammalian microRNA host genes and transcription units. Genome Res 14:1902-1910.
Rogers MF, Thomas J, Reddy AS, Ben-Hur A (2012) SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. Genome Biol 13:R4.
Rosenzweig BA, Pine PS, Domon OE, Morris SM, Chen JJ, Sistare FD (2004) Dye bias correction in dual-labeled cDNA microarray gene expression measurements. Environ Health Perspect 112:480-487.

Roshan U, Chikkagoudar S, Livesay DR (2008) Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities. BMC Bioinformatics 9:61.
Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85-94.
Royce TE, Rozowsky JS, Bertone P, Samanta M, Stolc V, Weissman S, Snyder M, Gerstein M (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. Trends Genet 21:466-475.
Ruan J, Stormo GD, Zhang W (2004) An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics 20:58-66.
Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. Cell 127:1193-1207.
Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE (2002) Using the transcriptome to annotate the genome. Nat Biotechnol 20:508-512.
Sankoff D (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM Journal on Applied Mathematics 45:810-825.
Satterlee JS, Barbee S, Jin P, Krichevsky A, Salama S, Schratt G, Wu DY (2007) Noncoding RNAs in the brain. J Neurosci 27:11856-11859.
Schaefer BC (1995) Revolutions in rapid amplification of cDNA ends: new strategies for polymerase chain reaction cloning of full-length cDNA ends. Anal Biochem 227:255-273.
Schatz MC (2009) CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics 25:1363-1369.
Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W (2003) Human-mouse alignments with BLASTZ. Genome Res 13:103-107.
Shendure J (2008) The beginning of the end for microarrays? Nat Methods 5:585-587.
Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawai J, Carninci P, Hayashizaki Y (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100:15776-15781.
Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15:1034-1050.
Simone NL, Bonner RF, Gillespie JW, Emmert-Buck MR, Liotta LA (1998) Laser-capture microdissection: opening the microscopic frontier to molecular analysis. Trends Genet 14:272-276.
Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117-1123.

Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.
Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128.
Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195-197.
Soukup JK, Soukup GA (2004) Riboswitches exert genetic control through metabolite-induced conformational change. Curr Opin Struct Biol 14:344-349.
Stalker J, Gibbins B, Meidl P, Smith J, Spooner W, Hotz HR, Cox AV (2004) The Ensembl Web site: mechanics of a genome browser. Genome Res 14:951-955.
Sun Y, Aljawad O, Lei J, Liu A (2012) Genome-scale NCRNA homology search using a Hamming distance-based filtration strategy. BMC Bioinformatics 13 Suppl 3:S12.
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.
Torarinsson E, Lindgreen S (2008) WAR: Webserver for aligning structural RNAs. Nucleic Acids Res 36:W79-84.
Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105-1111.
Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511-515.
Tuda J, Mongan AE, Tolba ME, Imada M, Yamagishi J, Xuan X, Wakaguri H, Sugano S, Sugimoto C, Suzuki Y (2011) Full-parasites: database of full-length cDNAs of apicomplexa parasites, 2010 update. Nucleic Acids Res 39:D625-631.
Tzakos AG, Grace CR, Lukavsky PJ, Riek R (2006) NMR techniques for very large proteins and rnas in solution. Annu Rev Biophys Biomol Struct 35:319-342.
UniProt (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40:D71-75.
Valen E et al. (2009) Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. Genome Res 19:255-265.
Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484-487.
Venter JC et al. (2001) The sequence of the human genome. Science 291:1304-1351.
Vieira FG, Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol 3:476-490.
Voss B, Meyer C, Giegerich R (2004) Evaluating the predictability of conformational switching in RNA. Bioinformatics 20:1573-1582.

Wang CW, Chen KT, Lu CL (2010) iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. Nucleic Acids Res 38:W340-347.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008a) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470-476.
Wang J, Zhang J, Zheng H, Li J, Liu D, Li H, Samudrala R, Yu J, Wong GK (2004) Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. Nature 431:1 p following 757; discussion following 757.
Wang X, Song X, Glass CK, Rosenfeld MG (2011) The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs. Cold Spring Harb Perspect Biol 3:a003756.
Wang X, Arai S, Song X, Reichart D, Du K, Pascual G, Tempst P, Rosenfeld MG, Glass CK, Kurokawa R (2008b) Induced ncRNAs allosterically modify RNA-binding proteins in cis to inhibit transcription. Nature 454:126-130.
Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57-63.
Wapinski O, Chang HY (2011) Long noncoding RNAs and human disease. Trends Cell Biol 21:354-361.
Waterston RH et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520-562.
Weinberg Z, Ruzzo WL (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. Bioinformatics 22:35-39.
Wilhelm BT, Landry JR (2009) RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249-257.
Wilm A, Higgins DG, Notredame C (2008) R-Coffee: a method for multiple alignment of non-coding RNA. Nucleic Acids Res 36:e52.
Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL (2008) The vertebrate genome annotation (Vega) database. Nucleic Acids Res 36:D753-760.
Willingham AT, Orth AP, Batalov S, Peters EC, Wen BG, Aza-Blanc P, Hogenesch JB, Schultz PG (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 309:1570-1573.
Xia T, SantaLucia J, Jr., Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH (1998) Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. Biochemistry 37:14719-14735.
Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18:821-829.
Zhang S, Borovok I, Aharonowitz Y, Sharan R, Bafna V (2006) A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. Bioinformatics 22:e557-565.
Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31:3406-3415.

Zuker M, Stiegler P (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res 9:133-148.