Int. J. Mol. Sci. **2013**, *14*, 15423-15458; doi:10.3390/ijms140815423

OPEN ACCESS
International Journal of
Molecular Sciences
ISSN 1422-0067
www.mdpi.com/journal/ijms

Review

Detecting and Comparing Non-Coding RNAs in the High-Throughput Era

Giovanni Bussotti [1,*], Cedric Notredame [2] and Anton J. Enright [1]

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK; E-Mail: aje@ebi.ac.uk

[2] Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Aiguader, 88, 08003 Barcelona, Spain; E-Mail: Cedric.Notredame@crg.eu

* Author to whom correspondence should be addressed; E-Mail: giovanni.bussotti@ebi.ac.uk; Tel.: +44-1223-49-2680 (ext. 2680); Fax: +44-1223-49-4486.

Received: 31 May 2013; in revised form: 16 July 2013 / Accepted: 17 July 2013 / Published: 24 July 2013

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 has become important. Aligning nucleotide sequences is a key 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 amino-acids. Moreover for many non-coding RNAs, evolution is likely to be mostly constrained at the structural level and not at the sequence level. This results in very poor sequence conservation impeding comparison of these molecules. These difficulties define a context where new methods are urgently needed in order to exploit experimental results to their full potential. This review focuses on the comparative genomics of non-coding RNAs in the context of new sequencing technologies and especially dealing with two extremely important and timely research aspects: the development of new methods to align RNAs and the analysis of high-throughput data.

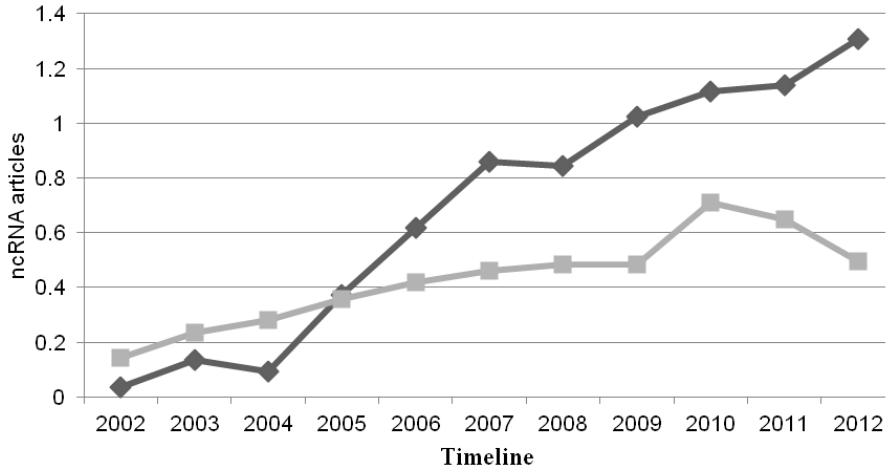
Keywords: lncRNAs; ncRNAs; high-throughput; RNA-seq; comparative biology; sequencing

1. Introduction

1.1. The Non-Coding RNA (New)-World

In recent years, the non-coding RNA (ncRNA) field has rapidly expanded (Figure 1) with a rapid 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 small nucleolar RNAs) and an even more limited collection of regulatory RNAs, such as lin-4 in *Caenorhabditis elegans* [1] and Xist in mammals [2]. Since then, the number of novel ncRNAs has increased dramatically and far more is known about their function, biogenesis, length, structural and sequence features. New and ever more sophisticated high-throughput technologies, such as tiling arrays and next generation sequencing (NGS) have been applied to comprehensively profile the transcriptome of various organisms.

Figure 1. Number of publications in PubMed found using the keyword "ncRNA" (dark grey) and "regulatory RNA" (pale gray). The x-axis represents the timeline, the y-axis the number of times the words "ncRNA" and "regulatory RNA" match 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 [3] and small nucleolar RNAs [4] and has resulted in the compilation or the update of many publicly available databases [5–10]. Furthermore, high-throughput approaches have revealed extensive and pervasive transcription of long ncRNAs (lncRNAs) [11–13], operationally defined as functional RNA longer than 200 base pairs that does not template protein synthesis. In the human genome, for instance, the GENCODE consortium annotated 9,640 lncRNA loci representing 15,512 transcripts [3,14] and in [15] the authors estimated that total number of human lncRNAs genes to be about 50,000, more than two-fold greater than the number of protein-coding genes. These discoveries were very timely in the context of growing concern for the lack of a significant correlation between the number of protein-coding genes and the commonly accepted concept of "organism complexity" [4,16,17]. 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) [18] had to be slightly altered and at least in higher eukaryotes may be inadequate [16,17]. The biological role of most of these novel long untranslated molecules is still a controversial issue. Some authors have even raised doubts about whether these transcripts are functional at all [19]. The lack of shared discernible features hampers our ability to define lncRNA classes, thus impeding function prediction [20]. However mounting experimental evidence illustrates that lncRNAs are implicated in a variety of biological processes [21] and are linked to various diseases including cancer [22]. Additionally, the functional roles of lncRNA transcripts have been uncovered in signaling sensors [23], embryonic stem cell differentiation [11], brain function [24,25], subcellular compartmentalization and chromatin remodeling [26]. Some examples include X chromosome inactivation by Xist, the silencing of autosomal imprinted genes accomplished by Air, nuclear trafficking regulated by NRON and muscle differentiation controlled by linc-MD1 [2,27–29]. In [30] the authors identified a class of lncRNAs named ncRNA-a (ncRNA-activator) able to stimulate the expression of proximal protein-coding genes, and a recent update on ncRNA-a [31] showed that the co-activator complex Mediator plays a central role in the activation process. See [21] and [32] for more examples and lncRNAdb [33] 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 [11,25,34]. It has been shown that lncRNAs can act both in *cis* [30,35] and in *trans* [36], some acting as precursors for short ncRNAs [37–39], while others act independently as long transcripts. As in [40] lncRNAs can be classified as "intergenic" or "genic" depending on their position/orientation with respect to protein-coding genes. lncRNAs not overlapping any protein-coding gene are tagged as intergenic and then further classified according to their transcription orientation with the closest protein-coding loci (same sense, convergent, or divergent). The genic lncRNA set are catalogued as "exonic" if overlapping a protein-coding exon. Otherwise, lncRNAs are labeled as "intronic", when positioned within protein-coding introns or as "overlapping", in presence of a protein-coding transcript located within the intron of the lncRNA [40].

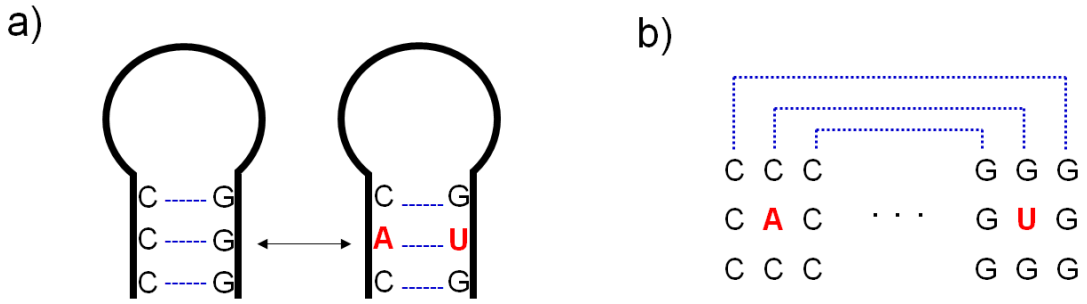
1.2. lncRNA Challenges

Although the conservation level of different lncRNAs may be not always directly comparable (e.g., the evolutionary conservation of genic lncRNAs may be biased by the presence of the protein-coding genes), overall approximately half of reported human lncRNA exhibit significant conservation across mammals [40]. These levels suggest some key cellular function, even though only a small fraction of these transcripts have so far been functionally characterized. Such functional analyses remain however, very superficial and lack precise molecular mechanisms explaining the activity of these novel transcripts. Our low level of understanding can be in part attributed to the difficulty with 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. After ribosomal RNA (rRNA), protein-coding mRNA represents the highest population of RNA species [41]. In previous studies [34,42,43] it has been reported that 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 [43,44]. 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 NGS. Additionally, our ability to assemble and annotate genomes was less advanced than currently and we had simplified notions 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" aspect of genomes [45,46]. Furthermore, there are specific classes of ncRNAs, such as circular RNAs (circRNAs), that have been extremely hard to identify. circRNAs are a class of non-coding RNA family that were discovered more than 20 years ago [47–50]. These RNAs form circles that arise from non-canonical splicing events (also known as exon shuffling) that join a splice donor to an upstream splice acceptor to produce a circular RNA molecule. Recent studies [51,52] show that the human circRNA CDR1as, antisense to the Cerebellar Degeneration-Related protein 1 (CDR1), hosts around 70 binding sites for the miR-7 microRNA and is highly associated with the Argonaute protein Ago2 as demonstrated by PAR-CLIP and HITS-CLIP experiments [51,52]. Mainly because of their non-canonical splicing behavior, circRNAs have eluded detection by next generation sequencing until recently. These latest studies adopted a novel computational approach to identify circRNAs from high-throughput RNA-seq data and demonstrated their widespread abundance within transcriptomes [51,53].

In general, a major obstacle for ncRNA detection is the difficulty to perform informative sequence comparisons. Standard primary sequence alignment is hampered by the low complexity of the nucleic alphabet, making it difficult to produce statistically meaningful RNA alignments. Ribonucleic acid chemistry relies on just four primary 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 [54]. Below this level, common ancestry is hard to infer with certainty. By comparison, this threshold is around ~20%–35% for proteins [55]. Furthermore, ncRNA appears to be evolving rapidly [56] or are under the influence of very specific evolutionary constraints [56]. It was proposed that most ncRNAs evolve at higher mutation rates, with the maintenance of secondary structures being the main source of selection [57,58]. This assumption makes sense from an evolutionary standpoint. As ncRNAs will be left untranslated, the nucleotide sequence itself is not constrained to keep the codon reading frame. Of course many exceptions exist. Specific ncRNAs types can hold functional sequences and act via their primary sequence (*i.e.*, miRNAs). Previous reports have shown that at least some miRNA genes are well conserved across species [59–61], 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 [62,63]. This induces a characteristic pattern of covariance that occurs when a mutation is affecting a nucleotide pairing to another in a structured domain (Figure 2). If the mutation breaks the base pairing so that the functionality of such a domain is compromised, the matching nucleotide is favored 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 (a). The presence of the compensatory mutation is highlighted by the multiple sequence comparison.



For most aligners these features of RNA are hard to account for when using standard alignment procedures that postulate positional independence and seek only to maximize identity. Furthermore, RNA can hold functional pseudo-knots. These are structural configurations where at least two RNA stem-loops are interposed one into the other. Although some comparative approaches including pseudo-knots exist [64,65], these are disregarded by most software due to reasons of computational complexity [66]. As a consequence ncRNA sequences are 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 we review established and more recent methodologies able to make the best of available RNA information (Section 2). Next we discuss different homology based strategies for ncRNA detection (Section 3) and the analysis of high-throughput expression data (Section 4). See Table 1 for a summary of the resources described in the text.

2. Comparing Non-Coding RNAs

As mentioned, generating meaningful ncRNA alignments is a challenging task and at least in some cases, the best accuracy could be achieved by exploiting 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 [67]. The polyanionic charge of the phosphate backbone makes the nucleotide sequence move 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 uninformative. Even when trying to resolve RNA molecules in solution using NMR, the resonance assignment is more difficult for RNA than for proteins [68]. RNAs have only 4 primary nucleosides instead of the 20 different

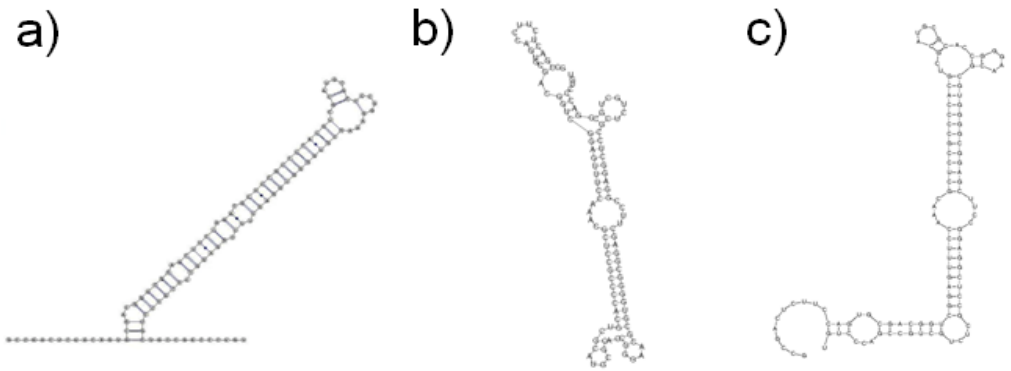
amino-acid side chains found in proteins [69]. Thus, the chemical shift dispersion is narrower in RNA than in proteins, resulting in less informative spectra [69].

2.1. RNA Structure Prediction

Because of these limitations, RNA structure is usually computationally predicted without any experimental support [70,71]. RNA secondary structure inference amounts to the computation of base-pairings that shape the *in vivo* molecule structure. The prediction is performed using primary RNA sequence data. Another possibility is including other sources of statistical information to constrain 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 [72] and knowledge based [73–75]. The first considers a biophysical model to calculate the structure whose folding has the minimum Gibbs free energy (ΔG). In this approach, [76–80] 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 [77–79]. The two most popular implementations of the MFE structure prediction algorithm are Mfold [70] and RNAfold [81] packages. The latter implements McCaskill's algorithm [82], 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 [83]. Knowledge based approaches rely on probabilistic models, where statistical learning procedures are used instead of empirical measurement of thermodynamic parameters. The Stochastic Context Free Grammar (SCFG) model [73] 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 consistency is the main limit of both MFE and knowledge based methods [84]. (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% [73–75,85]. This low figure may be the result of three confounding factors. Firstly, folding *in vivo* can be influenced by RNA chaperones [86], RNA editing [87], and even by the transcriptional process itself [88]. At present, there is no software able to account for these effects. Secondly, looking for a single structure can sometimes be inadequate. There are cases, such as the ribo-switches [89,90], where multiple functional structures can be derived from the same sequence depending on conditions such as temperature or other external factors. Standard predictors are not well suited to deal with such situations and require dedicated tools able to identify potential conformational switches [91,92]. Thirdly, RNAs might contain pseudo-knots, which are ignored by most tools due to reasons of computational complexity [66].

Figure 3. Consistency 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. [93]; (**c**) Mfold web server prediction, running Mfold version 4.6 [71].



The best secondary structure prediction accuracy can be achieved using comparative methods [66]. 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.

2.2. Structure Prediction and Alignment Strategies

Due to the close relationship between sequence and structure, structure prediction and sequence alignment can be described as interdependent problems [63]. As theorized by Sankoff [94], 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 reflected by the wealth of available alternative solutions. The web server WAR [95] is a good example. This tool allows the execution of a total of 14 different strategies to align and predict the common 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" [66,96]. 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 [97], T-Coffee [98]), and the folding of the aligned sequences can be performed by a plethora of tools (e.g., RNAalifold [99], PFOLD [100], ILM [101], ConStruct [102]) optimizing a score based on compensated mutations and thermodynamics. However this strategy is effective just in a determined sequence similarity range. On one hand, sequences that are too similar do not carry any covariance or purifying selection information and are less informative from an evolutionary standpoint. On the other hand, sequences need to be similar enough to be meaningfully aligned as below the "twilight zone" of similarity sequence alignment tends to obscure the covariance signal [96]. Plan B makes it possible to

use evolutionary signals to help improve 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 [94]. 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 [103,104], Foldalign [105,106], Stemloc [107], Consan [108]. Another example is pmmulti [109], 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 [110] or RNAcast [111]. 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 [112] 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 [97]). Plan C is particularly well suited to situations where experimental secondary structures are available.

The situation is radically different when experimental 3D structure information is available. In this 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 [113], DIAL [114], iPARTS [115], ARTS [116] and SARSA [117]. Besides this, recently several 3D RNA structure database search programs were developed, such as LaJolla [118] and FRASS [119].

Giving an exhaustive overview of the methods available for folding and aligning structured RNA sequences is well beyond the scope of this review. Over the last twenty years, more than 30 methods have been described that deal with these issues which is an indication of the complexity of this problem, despite 25 years of research following its formal description by Sankoff.

3. Detecting ncRNA Homologues

In the ncRNA field another critical step is the collection of homologues to genes of interest. Homologues can be used in several situations, such as the detection of functional motifs, inference of possible evolutionary steps or the design of laboratory 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 detection. As previously shown [120], homology search methods can be grouped in three sets: sequence-based, profiles 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 [121]. Nevertheless, this approach is CPU time demanding and implementations of this method have been until recently unpractical for large-scale database and genome wide screenings [122]. For this reason, alternative approaches such as FASTA [123] or BLAST [124] have been frequently preferred. These methods apply heuristics that boost computational speed at the cost of 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. The second class of approaches are based on profiles, including 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 [125,126]. However, such models have a linear architecture best suitable for modeling linear protein sequences (as opposed to secondary structure relationships). A more appropriate modeling 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 too computationally demanding, hence the need for approximate heuristic implementations of this exact algorithm. Such approximations include banded Sankoff tools [104,106,108,127], genetic algorithm implementations such as RAGA [128] and covariance models (CMs). CMs are the most commonly used method 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 [129] and implemented in Infernal [130]. This and other related applications such as RSEARCH [131] 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 modeling 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 matches, mismatches, insertions and deletions. However, generating accurate RNA alignments is difficult. In Rfam [132] CMs parameters are trained on high quality alignments (seed alignment) obtained from literature with manual curation. This input is used to estimate CMs, which are then passed to Infernal for homology searches. This CM/Infernal strategy is analogous to HMM/HMMER used for Pfam [133]. 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 [134]. This can be accomplished by means of *ad hoc* algorithms [135], profile HMMs like ML-heuristic [126] or BLAST with relaxed expectation values (*E*-values) to avoid losing sensitivity as achieved in Rfam [136]. A number of studies have been dedicated to the optimization of BLASTn parameters for seeking RNA homologs. For instance, in one study [120,137] the effectiveness of BLAST and other popular homology search methods tuned for ncRNA screenings were benchmarked. In [138], BlastR is introduced, 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 computationally demanding algorithms to serve as a pre-filtering tool. One merit of this approach is that it neither require profiles nor secondary structure information, but relies solely on information encoded in primary nucleotide sequences.

Together with sequence-based, profiles and structure-based methods, another possibility for detecting inter-species homologs involves the use of multiple genome alignments [43]. Once established reciprocity between blocks of genomes belonging to different organisms (*i.e.*, syntenic regions), coordinate transfer from one gene to its homolog is straightforward and implies the projection of corresponding positions. This has been made possible thanks to the availability of genome sequences [139–142] 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 [143]. Although comparing ncRNAs is currently still a complicated task, there exist several bioinformatics options to workaround the poor sequence conservation and effectively perform homology based prediction of novel ncRNAs.

4. High-Throughput Technologies and Genome-Wide Annotation of ncRNAs

4.1. Approaches for the High-Throughput Expression Detection

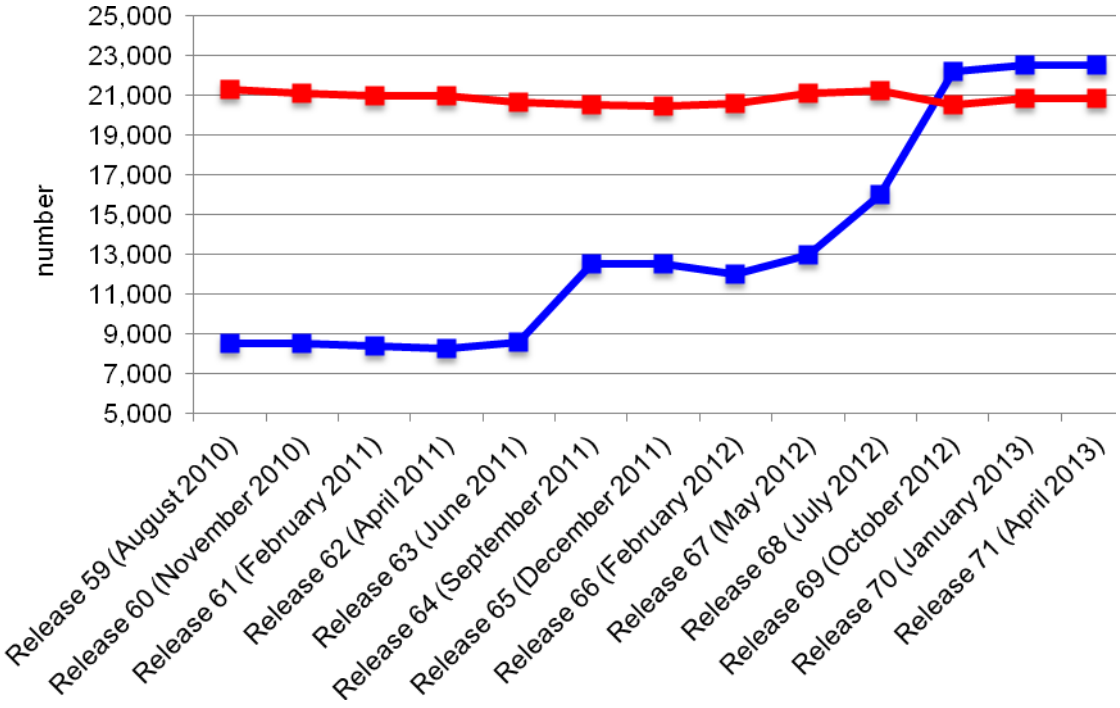
Recent technological advances have allowed the collection of an unprecedented amount of RNA sequence data coming from a wide range of organisms and conditions. For many years the main strategy for transcript discovery had been the sequencing of cloned complementary DNA (cDNA) of expressed sequence tags (ESTs) [144–146]. EST sequencing was then successfully used for the generation of large-scale expression datasets [147], and already by 1991 this approach had been utilized for human gene discovery [148]. 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 and contamination that can lead to redundancy and under-representation or over-representation of host-selected transcripts [149–151]. More recently, oligonucleotide microarray technologies have made high throughput expression analysis much more practical, while the even more recently developed RNA-seq technologies 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 labeled with fluorescent dyes, hybridized to the arrays, washed, and scanned with a laser [152]. 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 [153–155], dye-based identification problems [156–160] and physical manufacturing restrictions, impeding the detection of splicing events and the discovery of unannotated genes [151]. A variant of traditional expression array is represented by tiling arrays. These are chips that use extremely densely spotted and probes representing overlapping contiguous regions of genome. Several works relying on this technology and aiming at transcript discovery have been published [38,161–164]. However tiling arrays require a substantial quantity of RNA and have further limitations affecting their sensitivity, specificity and the detection of splicing [151]. For instance, as shown in [165], 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 [165–167]. 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 [152,166,168]). Additional methods for high-throughput RNA discovery include the serial analysis of gene expression (SAGE) [169,170], several updated variants such as LongSAGE [171], RL-SAGE [172], SuperSAGE [173] and analogous approaches like the massively parallel signature sequencing (MPSS) [174]. 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) [175] and the rapid amplification of cDNA ends (RACE) [176–178]. Both approaches can be used to demarcate transcript

boundaries, *i.e.*, define start and end of a transcript. Such information is extremely valuable *in situ* ations 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) [179,180] is a technique that allows high-throughout profiling of transcriptional starts points. Another promising application for ncRNA discovery, named RNA CaptureSeq, has been recently reported [181]. This approach 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 [181] RNA CaptureSeq can be used to fuel the detection of ncRNAs that are missed by genome-wide standard RNA sequencing.

4.2. Datasets

Undoubtedly, high-throughput technologies enable the tremendous possibility to get both qualitative and quantitative information on whole transcripts mass produced by cells. This has resulted in high-resolution views of RNA expression dynamics throughout different tissues and time points [182–184] and fueled the development of ncRNA specific databases, such as Rfam [132], NRED [20], lncRNAdb [33], RNAdb [185], fRNAdb [186] and NONCODE [187]. Furthermore, various groups and projects, such as RefSeq [188], GENCODE [14,189], HAVANA team [190,191], Ensembl [192] and FANTOM [193] 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 [188,194]. The GENCODE pipeline combines HAVANA and Ensembl automatic annotations to annotate the human gene features generated in the context of the ENCODE project [14,45,189]. 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 [192]. 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 FANTOM consortium aims to provide functional annotations to the full-length cDNAs [193]. The annotations generated by these consortia are freely available through genome browsers, including UCSC [195], Ensembl [196] and VEGA [197]. As new genomic regions get annotated and new transcript sequences become publicly available, these gene sets continue to growth [14,188,194]. A recent publication [14] 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, http://www.gencodegenes.org/archive_stats.html) to version 7 (December 2010, http://www.gencodegenes.org/archive_stats.html), the number of protein-coding transcripts increased from 68,880 to 76,052, and the number of lncRNAs jumped from 10,457 to 15,512. In terms of gene annotations, the number of known protein-coding genes has remained almost unchanged, while the ncRNA gene annotations expanded tremendously (see Figure 4).

Figure 4. Number of non-coding and protein-coding genes annotated over the last Ensembl releases. The *x*-axis indicates the number and the date of the release. The vertical axis reports the number of ncRNA (blue line) and protein-coding genes (red line).



The overall picture, however, remains blurred by inconsistent findings, suggesting that more analyses are still needed. For instance, the 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 [198]. This discrepancy can be in part explained by the fact that GENCODE catalogues transcripts using cDNA/EST alignments [14] rather than RNA-seq short-read data. A classic low-throughput EST sequencing operated by the Sanger technology can identify mostly high abundant transcripts [199], 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 [200].

4.3. NGS Challenges

To make the most out of the extraordinary possibilities that NGS offers, 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 [201]) and as a consequence it becomes necessary to reassemble the full-length transcripts. Small non-coding 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 large memory and by multi-CPU computers able to execute the algorithms in parallel with sufficient high-performance storage to store primary, temporary and output data. Although various short-read assemblers [202–204] were successfully applied to genome assembly, these packages cannot be easily

used to reconstruct transcriptomes. Applying tools normally designed for genome reconstruction to the problem of transcriptome assembly leads to multiple complications. A key issue is that the DNA sequencing depth is supposed to be identical over the entire genome while transcriptome sequencing depth is expected to fluctuate significantly. 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 modeling is involved as transcript variants coming from the same gene can share exons and are difficult to resolve unambiguously [199].

It is possible to work out the transcriptome assembly following a reference-based approach, a *de novo* assembly or combinations of each [199]. The first considers the initial mapping of the reads on a reference genome, and then the usage of transcript assemblers. To the end of labeling 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 [205–211]. Other important issues relate to the mapping of reads crossing exon-junction boundaries [212,213] 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 [214,215]. The quality scores, introduced by the Phred algorithm [216,217], indicate the reliability of each base call in each read in a log-likelihood scale. Since bases with reduced quality scores have an increased probability 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 of large size. This is a result of the fact that 3' ends of longer reads are affected by sequencing errors at higher rates [215]. Besides choosing a threshold for 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 indicates 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 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 computational 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 [218], Isolazo [219] and Scripture [42]. In [220] 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 [200,219]. Transcript assemblers seem to have a better agreement when reconstructing protein-coding transcripts [43] with the agreement dropping dramatically when modeling large intervening ncRNAs (lincRNAs). For instance, Cufflinks and

Scripture share only 46% agreement for lincRNA transcript models [43]. Such discrepancies are caused by the differences in how each assembler reconstruct lowly expressed transcripts [43]. In other words, about half of the isoforms estimated by a method in areas with low read density do not correspond to isoforms called by the other method. This poor agreement between transcript assemblers highlights the need for 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 *de novo* assembly of transcripts. This strategy does not require any reference genome and is therefore independent on the correct alignment of the reads to the splice sites. Advantages of this approach are that it is less reliant on accurate genome annotation and can be applied to organisms without sequenced or fully annotated genomes. Examples of applications adopting this strategy are described in [221–223]. 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 [199].

Once transcriptome dataset is generated, there are additional complications in the downstream analysis if trying to distinguish genuine ncRNAs from mRNAs. Currently, this issue is becoming increasingly important as many researchers are only interested in one or the other. The most straightforward procedure would be to compare a newly generated transcriptome against existing gene annotations. However in most cases annotations are far from 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 [198]. 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 trivial and may require dedicated expert curation [190]. 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 [224] and for genes having both coding and non-coding isoforms [225]. A commonly used approach to predict the coding potential involves the codon substitution frequency (CSF) estimation [226,227]. 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 [226,228,229], 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 [40] many putative human lncRNAs are not found in other species, and cannot be analyzed using CSF. Besides this, primate specific lncRNAs rarely show sufficient changes to highlight sense/nonsense mutations patterns. 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 flavors including BLASTx and RPS-Blast [124,230] that can be used to identify transcripts whose translational product possesses a match in protein databases such as Pfam [133] and UniProt [231]. Other algorithms include CPC (Coding Potential Calculator), a support vector machine (SVM) classifier including both Open Reading Frame (ORF) and homology predictions features [232], PORTRAIT (Prediction of transcriptomic ncRNA by ab initio methods) a SVM classifier not using homology information [233] and CPAT (Coding Potential Assessment Tool), a logistic regression model built with four sequence features including ORF predictions [234]. 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.

4.4. Other Approaches

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 [13]. 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 [235] 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) [235] to generate profiles of chromatin states. This approach revealed 1600 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 [13]. An alternative strategy used for ncRNA detection involves a combination of different high-throughput data sources and their integration using bioinformatics [236]. 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. These results illustrate how the integration of multiple information sources ends in highly accurate predictions of novel ncRNA genes.

Recently, a number of works reporting a massive quantity of novel ncRNA genes in various species have been published [40,43,44,237,238]. Such rapid growth has been possible thanks to the parallel development of new and ever more sophisticated bioinformatics approaches. Nevertheless, such analyses remain superficial with uncertainties of different type and degree affecting most predictions. For example, the homology search pipeline described in [40] is not sensitive enough to map rapidly evolving lncRNAs, hence the limit to play comprehensive evolutionary study. Moreover such 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. For validation purpose, some works provide the number of predicted lncRNA supported by expression evidences. For instance in [13] the authors confirmed by tiling array the expression of ~70% lncRNA predictions. In other cases as in [44], RT-PCR has been used to validate 15 newly identified lncRNAs. On the short run available transcription data is expected to increase very rapidly, and the necessity to accurately and quickly validate ncRNAs is becoming more pressing than ever.

Table 1. A summary of methods, datasets and browsers for non-coding RNA analysis. The first column indicates the resource type. The second column the resource name. The third column reports the PubMed ID when available, if not the web address. The fourth column provides a brief description of the resource.

	Resource	Pubmed ID	Description
Comparing ncRNAs (Section 2)	Mfold	6163133	Single sequence RNA secondary structure prediction.
	RNAfold	12824340	
	WAR	18492721	WEB server allowing the execution of different alignment methods
	RNAalifold	12079347	Folding previously aligned RNAs (Plan A)
	PFOLD	12824339	
	ILM	14693809	
	Construct	10518612	
	Dynalign	11902836	Sankoff derived algorithm for the simultaneous alignment and secondary structure prediction (Plan B)
	Foldalign	9278497	
	Stemloc	15790387	
	Consan	16952317	
	pmmulti	15073017	
	R-Coffee	18420654	Aligners taking into account previously estimated secondary structure (Plan C)
	RNAcast	16020472	
	SARA	18689811	3D structure alignment method
	DIAL	17567620	
	iPARTS	20507908	
	ARTS	16204124	
	SARSA	18502774	
	LaJolla	http://www.mdpi.com/1999-4893/2/2/692	
	FRASS	20553602	

Table 1. *Cont.*

	Resource	Pubmed ID	Description
Detecting ncRNAs (section 3)	ML-heuristic	16267089	Profile HMM
	RAGA	9358168	Genetic algorithm
	RSEARCH	14499004	Covariance model
	Infernal	12095421	
	BlastR	21624887	BLAST-based dinucleotide homology search
Datasets and browsers (section 4)	ENCODE	22955616	Consortium
	Ensembl	22086963	
	FANTOM	11217851	
	HAVANA	http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/	Annotation team
	GENCODE	22955987	Project for the annotation of all human gene features
	UCSC	12045153	Genome browser
	VEGA	18003653	
	RefSeq	18927115	Collection of DNA, transcripts, and proteins
	Rfam	12520045	ncRNA database
	NRED	18829717	
	lncRNAdb	21112873	
	RNAdb	17145715	
	fRNAdb	17099231	
	NONCODE	15608158	

5. Discussion and Conclusions

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 remodeling. 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 [239]. This "pervasive transcription" is strikingly high, especially when considering that a mere 3% of the human genome codes for protein-coding exons. [198]. Numerous novel, previously uncharacterized RNA species have been recently detected. A sizeable fraction of them are defined as lncRNA, *i.e.*, functional 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 has been experimentally validated, it is believed that many of these new ncRNAs belong to an important, relatively 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 sequencing technologies makes it possible to perform 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 [183,239]. Similar studies have been carried out in a number of genetic model organisms including rodents [44,151], plants [240], insects [184], worms [241] and yeasts [242]. In [243] the author argues that RNA-seq represents the most promising technology for transcriptome research. The main strength of RNA-seq approaches are the high degree of dynamic range they offers, returning better sensitivities than microarrays without the need of *a priori* speculation regarding the genomic loci being transcribed [244]. If the pace of scientific progress is maintained and if costs keep decreasing, 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 [152,245].

From a functional perspective, much remains to be done for accurate characterization and functional analysis of ncRNAs. To infer the function of novel ncRNAs one possibility is looking for functional motifs. This can be done by running motif finders algorithms to predict structurally conserved and potentially functional motifs [246–252]. Furthermore, the functional characterization of a novel ncRNA can be aided by the detection of protein-RNA binding motifs and the identification of protein interaction partners. Experimental approaches suited for this include RIP (Rna Immunoprecipitation) and CLIP (Cross-Linking and ImmunoPrecipitation) [253]. Comparative studies also offer a very efficient way to have functional insights and 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 major 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 (~22 nt) and up to ~17 kb long in the case of Xist [2]. Another 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 ORFs that help powering the statistical component of machine learning approaches when doing protein prediction [254]. The strongest signal contained by RNA sequences is 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 modeling. Emerging technologies allowing the high-throughput generation of experimentally derived secondary structures [255] will hopefully help addressing this problem. Unfortunately, taking into account secondary structures while comparing sequences is a challenging procedure, too intensive from a computational point of view to be practical in most circumstances [108]. 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). It has been shown [40,237] that BLAST can be effectively used for lncRNA homolog prediction, in combination with splicing informed heuristics such as exonerate [256] or GeneWise [257]. This strategy is not new, and similar approaches have already been used for the discovery of protein-coding homologs [258–260]. As one would expect,

homology based RNA searches are severely limited by the capacity to align distant homologues. For instance, when searching the human lncRNA complement against mammalian genomes [40] or when using an estimated pig complement [237], the authors 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 lncRNA queries that failed to map are really absent in the target species or undetected. In this context, high quality templates, such as the GENCODE queries used in [40], offer better likelihood to return precise annotation. In the same publication it is also predicted that sizeable fraction of the human lncRNAs is primate specific [40]. This result is in agreement with a recently published study [44] where the authors identified lncRNAs expressed in rodents' adult liver, and then compared the expression of the orthologous genomic regions. This work illustrates 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 homology based approaches is the possibility to identify novel human lncRNA genes candidates by using non-human templates as query [237]. As shown in the paper [237], 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 [40] the authors also extend the lncRNA conservation study to a multiple genome alignments strategy based on PhastCons conservation scores. The analysis is in agreement with previous reports [13,30] and confirms that lncRNAs sequences are less constrained than those of protein-coding genes. Remarkably, it was shown 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 that seen for proteins and suggests that a sizeable fraction of lncRNA genes are probably functional. The fraction of lncRNAs having a mutation rate almost indistinguishable from repeats suggests that at least some lncRNAs (close to a third) may be transcriptional noise. However, despite this abundance of lncRNA sequences that do not appear to be under selection, the transcript product itself might still have a biological role and as shown in [261,262] the transcription process itself of some ncRNA can bear regulative functions.

Despite the difficulties encountered when comparing ncRNAs, homology search of ncRNAs can be successfully used to detect new genes. New and ever more sophisticated algorithms will help addressing the challenges brought by NGS technologies. The ultimate goal is the creation of thorough transcriptome annotations and unbiased expression profiling of each individual transcript. It is still too early to tell. However, if they live up to their promises and expectation, the discovery of this new large class of RNAs may well define one of the turning points of modern biology.

Acknowledgments

The authors wish to acknowledge Roderic Guigó for the helpful comments and suggestions.

Conflict of Interest

The authors declare no conflict of interest.

References

1. Lee, R.C.; Feinbaum, R.L.; Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **1993**, *75*, 843–854.
2. Brown, C.J.; Hendrich, B.D.; Rupert, J.L.; Lafreniere, R.G.; Xing, Y.; Lawrence, J.; Willard, H.F. 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* **1992**, *71*, 527–542.
3. Farazi, T.A.; Juranek, S.A.; Tuschl, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **2008**, *135*, 1201–1214.
4. Bachellerie, J.P.; Cavaille, J.; Huttenhofer, A. The expanding snoRNA world. *Biochimie* **2002**, *84*, 775–790.
5. Barrett, T.; Suzek, T.O.; Troup, D.B.; Wilhite, S.E.; Ngau, W.C.; Ledoux, P.; Rudnev, D.; Lash, A.E.; Fujibuchi, W.; Edgar, R. NCBI GEO: Mining millions of expression profiles—database and tools. *Nucleic Acids Res.* **2005**, *33*, D562–D566.
6. Parkinson, H.; Sarkans, U.; Shojatalab, M.; Abeygunawardena, N.; Contrino, S.; Coulson, R.; Farne, A.; Lara, G.G.; Holloway, E.; *et al.* ArrayExpress—A public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* **2005**, *33*, D553–D555.
7. Griffiths-Jones, S.; Grocock, R.J.; van Dongen, S.; Bateman, A.; Enright, A.J. miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **2006**, *34*, D140–D144.
8. Fraser, B.A.; Weadick, C.J.; Janowitz, I.; Rodd, F.H.; Hughes, K.A. Sequencing and characterization of the guppy (Poecilia reticulata) transcriptome. *BMC Genomics* **2011**, *12*, 202.
9. Tuda, J.; Mongan, A.E.; Tolba, M.E.; Imada, M.; Yamagishi, J.; Xuan, X.; Wakaguri, H.; Sugano, S.; Sugimoto, C.; Suzuki, Y. Full-parasites: database of full-length cDNAs of apicomplexa parasites, 2010 update. *Nucleic Acids Res.* **2011**, *39*, D625–D631.
10. Mamidala, P.; Wijeratne, A.J.; Wijeratne, S.; Kornacker, K.; Sudhamalla, B.; Rivera-Vega, L.J.; Hoelmer, A.; Meulia, T.; Jones, S.C.; Mittapalli, O. RNA-Seq and molecular docking reveal multi-level pesticide resistance in the bed bug. *BMC Genomics* **2012**, *13*, 6.
11. Dinger, M.E.; Amaral, P.P.; Mercer, T.R.; Pang, K.C.; Bruce, S.J.; Gardiner, B.B.; Askarian-Amiri, M.E.; Ru, K.; Solda, G.; Simons, C.; *et al.* Long noncoding RNAs in mouse embryonic stem cell pluripotency and differentiation. *Genome Res.* **2008**, *18*, 1433–1445.
12. Okazaki, Y.; Furuno, M.; Kasukawa, T.; Adachi, J.; Bono, H.; Kondo, S.; Nikaido, I.; Osato, N.; Saito, R.; Suzuki, H.; *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **2002**, *420*, 563–573.
13. Guttman, M.; Amit, I.; Garber, M.; French, C.; Lin, M.F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B.W.; Cassady, J.P.; *et al.* Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **2009**, *458*, 223–227.

14. Harrow, J.; Frankish, A.; Gonzalez, J.M.; Tapanari, E.; Diekhans, M.; Kokocinski, F.; Aken, B.L.; Barrell, D.; Zadissa, A.; Searle, S.; *et al.* GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res.* **2012**, *22*, 1760–1774.
15. Managadze, D.; Lobkovsky, A.E.; Wolf, Y.I.; Shabalina, S.A.; Rogozin, I.B.; Koonin, E.V. The vast, conserved mammalian lincRNome. *PLoS Comput. Biol.* **2013**, *9*, e1002917.
16. Mattick, J.S.; Gagen, M.J. The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Mol. Biol. Evol.* **2001**, *18*, 1611–1630.
17. Mattick, J.S. Non-coding RNAs: The architects of eukaryotic complexity. *EMBO Rep.* **2001**, *2*, 986–991.
18. Crick, F.H. On protein synthesis. *Symp. Soc. Exp. Biol.* **1958**, *12*, 138–163.
19. Wang, J.; Zhang, J.; Zheng, H.; Li, J.; Liu, D.; Li, H.; Samudrala, R.; Yu, J.; Wong, G.K. Mouse transcriptome: Neutral evolution of "non-coding" complementary DNAs. *Nature* **2004**, *431*, doi:10.1038/nature03016.
20. Dinger, M.E.; Pang, K.C.; Mercer, T.R.; Crowe, M.L.; Grimmond, S.M.; Mattick, J.S. NRED: A database of long noncoding RNA expression. *Nucleic Acids Res.* **2009**, *37*, D122–D126.
21. Mattick, J.S. The genetic signatures of noncoding RNAs. *PLoS Genet.* **2009**, *5*, e1000459.
22. Wapinski, O.; Chang, H.Y. Long noncoding RNAs and human disease. *Trends Cell. Biol.* **2011**, *21*, 354–361.
23. Wang, X.; Song, X.; Glass, C.K.; Rosenfeld, M.G. The long arm of long noncoding RNAs: roles as sensors regulating gene transcriptional programs. *Cold Spring Harb. Perspect. Biol.* **2011**, *3*, a003756.
24. Satterlee, J.S.; Barbee, S.; Jin, P.; Krichevsky, A.; Salama, S.; Schratt, G.; Wu, D.Y. Noncoding RNAs in the brain. *J. Neurosci*. **2007**, *27*, 11856–11859.
25. Mercer, T.R.; Dinger, M.E.; Sunkin, S.M.; Mehler, M.F.; Mattick, J.S. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 716–721.
26. Kaikkonen, M.U.; Lam, M.T.; Glass, C.K. Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovasc. Res.* **2011**, *90*, 430–440.
27. Braidotti, G.; Baubec, T.; Pauler, F.; Seidl, C.; Smrzka, O.; Stricker, S.; Yotova, I.; Barlow, D.P. The Air noncoding RNA: An imprinted *cis*-silencing transcript. *Cold Spring Harb. Symp. Quant. Biol.* **2004**, *69*, 55–66.
28. Willingham, A.T.; Orth, A.P.; Batalov, S.; Peters, E.C.; Wen, B.G.; Aza-Blanc, P.; Hogenesch, J.B.; Schultz, P.G. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **2005**, *309*, 1570–1573.
29. Cesana, M.; Cacchiarelli, D.; Legnini, I.; Santini, T.; Sthandier, O.; Chinappi, M.; Tramontano, A.; Bozzoni, I. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* **2011**, *147*, 358–369.
30. Ørom, U.A.; Derrien, T.; Beringer, M.; Gumireddy, K.; Gardini, A.; Bussotti, G.; Lai, F.; Zytnicki, M.; Notredame, C.; Huang, Q.; *et al.* Long noncoding RNAs with enhancer-like function in human cells. *Cell* **2010**, *143*, 46–58.

31. Lai, F.; Orom, U.A.; Cesaroni, M.; Beringer, M.; Taatjes, D.J.; Blobel, G.A.; Shiekhattar, R. Activating RNAs associate with Mediator to enhance chromatin architecture and transcription. *Nature* **2013**, *494*, 497–501.
32. Rinn, J.L.; Chang, H.Y. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* **2012**, *81*, 145–166.
33. Amaral, P.P.; Clark, M.B.; Gascoigne, D.K.; Dinger, M.E.; Mattick, J.S. lncRNAdb: A reference database for long noncoding RNAs. *Nucleic Acids Res.* **2011**, *39*, D146–D151.
34. Ravasi, T.; Suzuki, H.; Pang, K.C.; Katayama, S.; Furuno, M.; Okunishi, R.; Fukuda, S.; Ru, K.; Frith, M.C.; Gongora, M.M.; *et al.* Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.* **2006**, *16*, 11–19.
35. Wang, X.; Arai, S.; Song, X.; Reichart, D.; Du, K.; Pascual, G.; Tempst, P.; Rosenfeld, M.G.; Glass, C.K.; Kurokawa, R. Induced ncRNAs allosterically modify RNA-binding proteins in *cis* to inhibit transcription. *Nature* **2008**, *454*, 126–130.
36. Rinn, J.L.; Kertesz, M.; Wang, J.K.; Squazzo, S.L.; Xu, X.; Brugmann, S.A.; Goodnough, L.H.; Helms, J.A.; Farnham, P.J.; Segal, E.; *et al.* Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* **2007**, *129*, 1311–1323.
37. Rodriguez, A.; Griffiths-Jones, S.; Ashurst, J.L.; Bradley, A. Identification of mammalian microRNA host genes and transcription units. *Genome Res.* **2004**, *14*, 1902–1910.
38. Kapranov, P.; Cheng, J.; Dike, S.; Nix, D.A.; Duttagupta, R.; Willingham, A.T.; Stadler, P.F.; Hertel, J.; Hackermuller, J.; Hofacker, I.L.; *et al.* RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science* **2007**, *316*, 1484–1488.
39. Ogawa, Y.; Sun, B.K.; Lee, J.T. Intersection of the RNA interference and X-inactivation pathways. *Science* **2008**, *320*, 1336–1341.
40. Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer, A.; Djebali, S.; Tilgner, H.; Guernec, G.; Martin, D.; Merkel, A.; Knowles, D.G.; *et al.* The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **2012**, *22*, 1775–1789.
41. Chen, Z.; Duan, X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol. Biol.* **2011**, *733*, 93–103.
42. Guttman, M.; Garber, M.; Levin, J.Z.; Donaghey, J.; Robinson, J.; Adiconis, X.; Fan, L.; Koziol, M.J.; Gnirke, A.; Nusbaum, C.; *et al. Ab. initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **2010**, *28*, 503–510.
43. Cabili, M.N.; Trapnell, C.; Goff, L.; Koziol, M.; Tazon-Vega, B.; Regev, A.; Rinn, J.L. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* **2011**, *25*, 1915–1927.
44. Kutter, C.; Watt, S.; Stefflova, K.; Wilson, M.D.; Goncalves, A.; Ponting, C.P.; Odom, D.T.; Marques, A.C. Rapid turnover of long noncoding rnas and the evolution of gene expression. *PLoS Genet.* **2012**, *8*, e1002841.
45. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **2007**, *447*, 799–816.

46. Clark, M.B.; Amaral, P.P.; Schlesinger, F.J.; Dinger, M.E.; Taft, R.J.; Rinn, J.L.; Ponting, C.P.; Stadler, P.F.; Morris, K.V.; Morillon, A.; *et al.* The reality of pervasive transcription. *PLoS Biol.* **2011**, *9*, doi:10.1371/journal.pbio.1000625.
47. Capel, B.; Swain, A.; Nicolis, S.; Hacker, A.; Walter, M.; Koopman, P.; Goodfellow, P.; Lovell-Badge, R. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* **1993**, *73*, 1019–1030.
48. Cocquerelle, C.; Mascrez, B.; Hétuin, D.; Bailleul, B. Mis-splicing yields circular RNA molecules. *FASEB J.* **1993**, *7*, 155–160.
49. Nigro, J.M.; Cho, K.R.; Fearon, E.R.; Kern, S.E.; Ruppert, J.M.; Oliner, J.D.; Kinzler, K.W.; Vogelstein, B. Scrambled exons. *Cell* **1991**, *64*, 607–613.
50. Zaphiropoulos, P.G. Exon skipping and circular RNA formation in transcripts of the human cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol. Cell. Biol.* **1997**, *17*, 2985–2993.
51. Memczak, S.; Jens, M.; Elefsinioti, A.; Torti, F.; Krueger, J.; Rybak, A.; Maier, L.; Mackowiak, S.D.; Gregersen, L.H.; Munschauer, M.; *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **2013**, *495*, 333–338.
52. Hansen, T.B.; Jensen, T.I.; Clausen, B.H.; Bramsen, J.B.; Finsen, B.; Damgaard, C.K.; Kjems, J. Natural RNA circles function as efficient microRNA sponges. *Nature* **2013**, *495*, 384–388.
53. Jeck, W.R.; Sorrentino, J.A.; Wang, K.; Slevin, M.K.; Burd, C.E.; Liu, J.; Marzluff, W.F.; Sharpless, N.E. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **2013**, *19*, 141–157.
54. Capriotti, E.; Marti-Renom, M.A. Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinforma.* **2010**, *11*, 322.
55. Rost, B. Twilight zone of protein sequence alignments. *Protein Eng.* **1999**, *12*, 85–94.
56. Pang, K.C.; Frith, M.C.; Mattick, J.S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **2006**, *22*, 1–5.
57. Bernhart, S.H.; Hofacker, I.L. From consensus structure prediction to RNA gene finding. *Brief Funct. Genomic Proteomic* **2009**, *8*, 461–471.
58. Sun, Y.; Aljawad, O.; Lei, J.; Liu, A. Genome-scale NCRNA homology search using a Hamming distance-based filtration strategy. *BMC Bioinforma.* **2012**, *13*, S12.
59. Bentwich, I.; Avniel, A.; Karov, Y.; Aharonov, R.; Gilad, S.; Barad, O.; Barzilai, A.; Einat, P.; Einav, U.; Meiri, E.; *et al.* Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **2005**, *37*, 766–770.
60. 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.; *et al.* Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.* **2006**, *16*, 1289–1298.
61. Guerra-Assuncao, J.A.; Enright, A.J. Large-scale analysis of microRNA evolution. *BMC Genomics* **2012**, *13*, 218.
62. Missal, K.; Rose, D.; Stadler, P.F. Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics* **2005**, *21*, ii77–ii78.
63. Lindgreen, S.; Gardner, P.P.; Krogh, A. MASTR: Multiple alignment and structure prediction of non-coding RNAs using simulated annealing. *Bioinformatics* **2007**, *23*, 3304–3311.

64. Sperschneider, J.; Datta, A.; Wise, M.J. Predicting pseudoknotted structures across two RNA sequences. *Bioinformatics* **2012**, *28*, 3058–3065.
65. Wong, T.K.F.; Wan, K.-L.; Hsu, B.-Y.; Cheung, B.W.Y.; Hon, W.-K.; Lam, T.-W.; Yiu, S.-M. RNASAlign: RNA structural alignment system. *Bioinformatics* **2011**, *27*, 2151–2152.
66. Gardner, P.P.; Giegerich, R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinforma.* **2004**, *5*, 140.
67. Ravindran, P.P.; Heroux, A.; Ye, J.D. Improvement of the crystallizability and expression of an RNA crystallization chaperone. *J. Biochem.* **2011**, *150*, 535–543.
68. Furtig, B.; Richter, C.; Wohnert, J.; Schwalbe, H. NMR spectroscopy of RNA. *Chembiochem* **2003**, *4*, 936–962.
69. Tzakos, A.G.; Grace, C.R.; Lukavsky, P.J.; Riek, R. NMR techniques for very large proteins and rnas in solution. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 319–342.
70. Zuker, M.; Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **1981**, *9*, 133–148.
71. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **2003**, *31*, 3406–3415.
72. Mathews, D.H.; Turner, D.H.; Zuker, M. RNA secondary structure prediction. *Curr. Protoc. Nucleic Acid Chem.* **2007**, doi:10.1002/0471142700.nc1102s28.
73. Dowell, R.D.; Eddy, S.R. Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction. *BMC Bioinforma.* **2004**, *5*, 71.
74. Dima, R.I.; Hyeon, C.; Thirumalai, D. Extracting stacking interaction parameters for RNA from the data set of native structures. *J. Mol. Biol.* **2005**, *347*, 53–69.
75. Do, C.B.; Woods, D.A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, e90–e98.
76. Xia, T.; SantaLucia, J.J.; Burkard, M.E.; Kierzek, R.; Schroeder, S.J.; Jiao, X.; Cox, C.; Turner, D.H. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **1998**, *37*, 14719–14735.
77. Mathews, D.H.; Sabina, J.; Zuker, M.; Turner, D.H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **1999**, *288*, 911–940.
78. Mathews, D.H.; Disney, M.D.; Childs, J.L.; Schroeder, S.J.; Zuker, M.; Turner, D.H. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7287–7292.
79. Lu, Z.J.; Turner, D.H.; Mathews, D.H. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.* **2006**, *34*, 4912–4924.
80. Lu, Z.J.; Gloor, J.W.; Mathews, D.H. Improved RNA secondary structure prediction by maximizing expected pair accuracy. *RNA* **2009**, *15*, 1805–1813.
81. Hofacker, I.L.; Fontana W.; Stadler, P.F.; Bonhoeffer S.; Tacker M.; Schuster, P. Fast folding and comparison of rna secondary structures. *Monatshefte f. Chem.* **1994**, *125*, 167–188.
82. McCaskill, J.S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **1990**, *29*, 1105–1119.

83. Durbin, R.; Eddy, S.R.; Krogh, A.; Mitchison, G. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*; Cambridge University Press: Cambridge, UK, 1998.
84. Deigan, K.E.; Li, T.W.; Mathews, D.H.; Weeks, K.M. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 97–102.
85. Doshi, K.J.; Cannone, J.J.; Cobaugh, C.W.; Gutell, R.R. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinforma.* **2004**, *5*, 105.
86. Herschlag, D. RNA chaperones and the RNA folding problem. *J. Biol. Chem.* **1995**, *270*, 20871–20874.
87. Brennicke, A.; Marchfelder, A.; Binder, S. RNA editing. *FEMS Microbiol. Rev.* **1999**, *23*, 297–316.
88. Pan, T.; Sosnick, T. RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.* **2006**, *35*, 161–175.
89. Mandal, M.; Breaker, R.R. Gene regulation by riboswitches. *Nat. Rev. Mol. Cell. Biol.* **2004**, *5*, 451–463.
90. Soukup, J.K.; Soukup, G.A. Riboswitches exert genetic control through metabolite-induced conformational change. *Curr. Opin. Struct. Biol.* **2004**, *14*, 344–349.
91. Bengert, P.; Dandekar, T. Riboswitch finder—A tool for identification of riboswitch RNAs. *Nucleic Acids Res.* **2004**, *32*, W154–W159.
92. Voss, B.; Meyer, C.; Giegerich, R. Evaluating the predictability of conformational switching in RNA. *Bioinformatics* **2004**, *20*, 1573–1582.
93. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **2003**, *31*, 3429–3431.
94. Sankoff, D. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.* **1985**, *45*, 810–825.
95. Torarinsson, E.; Lindgreen, S. WAR: Webserver for aligning structural RNAs. *Nucleic Acids Res.* **2008**, *36*, W79–W84.
96. Bremges, A.; Schirmer, S.; Giegerich, R. Fine-tuning structural RNA alignments in the twilight zone. *BMC Bioinforma.* **2010**, *11*, 222.
97. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.
98. Notredame, C.; Higgins, D.G.; Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **2000**, *302*, 205–217.
99. Hofacker, I.L.; Fekete, M.; Stadler, P.F. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **2002**, *319*, 1059–1066.
100. Knudsen, B.; Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* **2003**, *31*, 3423–3428.
101. Ruan, J.; Stormo, G.D.; Zhang, W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* **2004**, *20*, 58–66.
102. Luck, R.; Graf, S.; Steger, G. ConStruct: A tool for thermodynamic controlled prediction of conserved secondary structure. *Nucleic Acids Res.* **1999**, *27*, 4208–4217.

103. Mathews, D.H.; Turner, D.H. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* **2002**, *317*, 191–203.
104. Mathews, D.H. Predicting a set of minimal free energy RNA secondary structures common to two sequences. *Bioinformatics* **2005**, *21*, 2246–2253.
105. Gorodkin, J.; Heyer, L.J.; Stormo, G.D. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res.* **1997**, *25*, 3724–3732.
106. Havgaard, J.H.; Lyngso, R.B.; Stormo, G.D.; Gorodkin, J. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* **2005**, *21*, 1815–1824.
107. Holmes, I. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinforma.* **2005**, *6*, 73.
108. Dowell, R.D.; Eddy, S.R. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinforma.* **2006**, *7*, 400.
109. Hofacker, I.L.; Bernhart, S.H.; Stadler, P.F. Alignment of RNA base pairing probability matrices. *Bioinformatics* **2004**, *20*, 2222–2227.
110. Wilm, A.; Higgins, D.G.; Notredame, C. R-Coffee: A method for multiple alignment of non-coding RNA. *Nucleic Acids Res.* **2008**, *36*, e52.
111. Reeder, J.; Giegerich, R. Consensus shapes: An alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* **2005**, *21*, 3516–3523.
112. Bernhart, S.H.; Hofacker, I.L.; Stadler, P.F. Local RNA base pairing probabilities in large sequences. *Bioinformatics* **2006**, *22*, 614–615.
113. Capriotti, E.; Marti-Renom, M.A. RNA structure alignment by a unit-vector approach. *Bioinformatics* **2008**, *24*, i112–i118.
114. Ferre, F.; Ponty, Y.; Lorenz, W.A.; Clote, P. 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.* **2007**, *35*, W659–W668.
115. Wang, C.W.; Chen, K.T.; Lu, C.L. iPARTS: An improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res.* **2010**, *38*, W340–W347.
116. Dror, O.; Nussinov, R.; Wolfson, H. ARTS: Alignment of RNA tertiary structures. *Bioinformatics* **2005**, *21*, ii47–ii53.
117. Chang, Y.F.; Huang, Y.L.; Lu, C.L. SARSA: A web tool for structural alignment of RNA using a structural alphabet. *Nucleic Acids Res.* **2008**, *36*, W19–W24.
118. Bauer, R.P.; Rother, K.P.; Moor, P.P.; Reinert, K.P.; Steinke, T.P.; Bujnicki, J.P.; Preissner, R.P. Fast structural alignment of biomolecules using a hash table, N-grams and string descriptors. *Algorithms* **2009**, *2*, 692–709.
119. Kirillova, S.; Tosatto, S.C.; Carugo, O. FRASS: The web-server for RNA structural comparison. *BMC Bioinforma.* **2010**, *11*, 327.
120. Freyhult, E.K.; Bollback, J.P.; Gardner, P.P. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res.* **2007**, *17*, 117–125.
121. Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197.

122. Rognes, T. Faster Smith-Waterman database searches with inter-sequence SIMD parallelisation. *BMC Bioinforma.* **2011**, *12*, 221.
123. Lipman, D.J.; Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **1985**, *227*, 1435–1441.
124. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
125. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755–763.
126. Weinberg, Z.; Ruzzo, W.L. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* **2006**, *22*, 35–39.
127. Holmes, I. A probabilistic model for the evolution of RNA structure. *BMC Bioinforma.* **2004**, *5*, 166.
128. Notredame, C.; O'Brien, E.A.; Higgins, D.G. RAGA: RNA sequence alignment by genetic algorithm. *Nucleic Acids Res.* **1997**, *25*, 4570–4580.
129. Eddy, S.R.; Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **1994**, *22*, 2079–2088.
130. Eddy, S.R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinforma.* **2002**, *3*, 18.
131. Klein, R.J.; Eddy, S.R. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinforma.* **2003**, *4*, 44.
132. Griffiths-Jones, S.; Bateman, A.; Marshall, M.; Khanna, A.; Eddy, S.R. Rfam: An RNA family database. *Nucleic Acids Res.* **2003**, *31*, 439–441.
133. Finn, R.D.; Tate, J.; Mistry, J.; Coggill, P.C.; Sammut, S.J.; Hotz, H.R.; Ceric, G.; Forslund, K.; Eddy, S.R.; Sonnhammer, E.L.; *et al.* The Pfam protein families database. *Nucleic Acids Res.* **2008**, *36*, D281–D288.
134. Griffiths-Jones, S.; Moxon, S.; Marshall, M.; Khanna, A.; Eddy, S.R.; Bateman, A. Rfam: Annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **2005**, *33*, D121–D124.
135. Zhang, S.; Borovok, I.; Aharonowitz, Y.; Sharan, R.; Bafna, V. A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. *Bioinformatics* **2006**, *22*, e557–e565.
136. Gardner, P.P.; Daub, J.; Tate, J.G.; Nawrocki, E.P.; Kolbe, D.L.; Lindgreen, S.; Wilkinson, A.C.; Finn, R.D.; Griffiths-Jones, S.; Eddy, S.R.; *et al.* Rfam: Updates to the RNA families database. *Nucleic Acids Res.* **2009**, *37*, D136–D140.
137. Roshan, U.; Chikkagoudar, S.; Livesay, D.R. Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities. *BMC Bioinforma.* **2008**, *9*, 61.
138. Bussotti, G.; Raineri, E.; Erb, I.; Zytnicki, M.; Wilm, A.; Beaudoing, E.; Bucher, P.; Notredame, C. BlastR—Fast and accurate database searches for non-coding RNAs. *Nucleic Acids Res.* **2011**, *39*, 6886–6895.
139. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; *et al.* Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.

140. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; *et al.* The sequence of the human genome. *Science* **2001**, *291*, 1304–1351.
141. Aparicio, S.; Chapman, J.; Stupka, E.; Putnam, N.; Chia, J.M.; Dehal, P.; Christoffels, A.; Rash, S.; Hoon, S.; Smit, A.; *et al.* Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **2002**, *297*, 1301–1310.
142. Waterston, R.H.; Lindblad-Toh, K.; Birney, E.; Rogers, J.; Abril, J.F.; Agarwal, P.; Agarwala, R.; Ainscough, R.; Alexandersson, M.; An, P.; *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **2002**, *420*, 520–562.
143. Schwartz, S.; Kent, W.J.; Smit, A.; Zhang, Z.; Baertsch, R.; Hardison, R.C.; Haussler, D.; Miller, W. Human-mouse alignments with BLASTZ. *Genome Res.* **2003**, *13*, 103–107.
144. Boguski, M.S.; Tolstoshev, C.M.; Bassett, D.E.J. Gene discovery in dbEST. *Science* **1994**, *265*, 1993–1994.
145. Dias Neto, E.; Correa, R.G.; Verjovski-Almeida, S.; Briones, M.R.; Nagai, M.A.; da Silva, W.J.; Zago, M.A.; Bordin, S.; Costa, F.F.; Goldman, G.H.; *et al.* Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 3491–3496.
146. Gerhard, D.S.; Wagner, L.; Feingold, E.A.; Shenmen, C.M.; Grouse, L.H.; Schuler, G.; Klein, S.L.; Old, S.; Rasooly, R.; Good, P.; *et al.* The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). *Genome Res.* **2004**, *14*, 2121–2127.
147. Boguski, M.S.; Lowe, T.M.; Tolstoshev, C.M. dbEST—Database for "expressed sequence tags." *Nat. Genet.* **1993**, *4*, 332–333.
148. Adams, M.D.; Kelley, J.M.; Gocayne, J.D.; Dubnick, M.; Polymeropoulos, M.H.; Xiao, H.; Merril, C.R.; Wu, A.; Olde, B.; Moreno, R.F.; *et al.* Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **1991**, *252*, 1651–1656.
149. Bonaldo, M.F.; Lennon, G.; Soares, M.B. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **1996**, *6*, 791–806.
150. Nagaraj, S.H.; Gasser, R.B.; Ranganathan, S. A hitchhiker's guide to expressed sequence tag (EST) analysis. *Brief. Bioinforma.* **2007**, *8*, 6–21.
151. Mortazavi, A.; Williams, B.A.; McCue, K.; Schaeffer, L.; Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **2008**, *5*, 621–628.
152. Malone, J.H.; Oliver, B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* **2011**, *9*, 34.
153. Eklund, A.C.; Turner, L.R.; Chen, P.; Jensen, R.V.; deFeo, G.; Kopf-Sill, A.R.; Szallasi, Z. Replacing cRNA targets with cDNA reduces microarray cross-hybridization. *Nat. Biotechnol.* **2006**, *24*, 1071–1073.
154. Okoniewski, M.J.; Miller, C.J. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinforma.* **2006**, *7*, 276.
155. Casneuf, T.; van de Peer, Y.; Huber, W. *In situ* analysis of cross-hybridisation on microarrays and the inference of expression correlation. *BMC Bioinforma.* **2007**, *8*, 461.
156. Cox, W.G.; Beaudet, M.P.; Agnew, J.Y.; Ruth, J.L. Possible sources of dye-related signal correlation bias in two-color DNA microarray assays. *Anal. Biochem.* **2004**, *331*, 243–254.

157. Dombkowski, A.A.; Thibodeau, B.J.; Starcevic, S.L.; Novak, R.F. Gene-specific dye bias in microarray reference designs. *FEBS Lett.* **2004**, *560*, 120–124.
158. Rosenzweig, B.A.; Pine, P.S.; Domon, O.E.; Morris, S.M.; Chen, J.J.; Sistare, F.D. Dye bias correction in dual-labeled cDNA microarray gene expression measurements. *Environ. Health Perspect.* **2004**, *112*, 480–487.
159. Dobbin, K.K.; Kawasaki, E.S.; Petersen, D.W.; Simon, R.M. Characterizing dye bias in microarray experiments. *Bioinformatics* **2005**, *21*, 2430–2437.
160. Martin-Magniette, M.L.; Aubert, J.; Cabannes, E.; Daudin, J.J. Evaluation of the gene-specific dye bias in cDNA microarray experiments. *Bioinformatics* **2005**, *21*, 1995–2000.
161. Bertone, P.; Stolc, V.; Royce, T.E.; Rozowsky, J.S.; Urban, A.E.; Zhu, X.; Rinn, J.L.; Tongprasit, W.; Samanta, M.; Weissman, S.; *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **2004**, *306*, 2242–2246.
162. Cheng, J.; Kapranov, P.; Drenkow, J.; Dike, S.; Brubaker, S.; Patel, S.; Long, J.; Stern, D.; Tammana, H.; Helt, G.; *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **2005**, *308*, 1149–1154.
163. Royce, T.E.; Rozowsky, J.S.; Bertone, P.; Samanta, M.; Stolc, V.; Weissman, S.; Snyder, M.; Gerstein, M. Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet.* **2005**, *21*, 466–475.
164. Kapranov, P.; Willingham, A.T.; Gingeras, T.R. Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.* **2007**, *8*, 413–423.
165. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63.
166. Marioni, J.C.; Mason, C.E.; Mane, S.M.; Stephens, M.; Gilad, Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **2008**, *18*, 1509–1517.
167. Fu, X.; Fu, N.; Guo, S.; Yan, Z.; Xu, Y.; Hu, H.; Menzel, C.; Chen, W.; Li, Y.; Zeng, R.; *et al.* Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **2009**, *10*, 161.
168. Ozsolak, F.; Milos, P.M. RNA Sequencing: Advances, challenges and opportunities. *Nat. Rev. Genet.* **2011**, *12*, 87–98.
169. Velculescu, V.E.; Zhang, L.; Vogelstein, B.; Kinzler, K.W. Serial analysis of gene expression. *Science* **1995**, *270*, 484–487.
170. Harbers, M.; Carninci, P. Tag-based approaches for transcriptome research and genome annotation. *Nat. Methods* **2005**, *2*, 495–502.
171. Saha, S.; Sparks, A. B.; Rago, C.; Akmaev, V.; Wang, C.J.; Vogelstein, B.; Kinzler, K.W.; Velculescu, V.E. Using the transcriptome to annotate the genome. *Nat. Biotechnol.* **2002**, *20*, 508–512.
172. Gowda, M.; Jantasuriyarat, C.; Dean, R.A.; Wang, G.L. Robust-LongSAGE (RL-SAGE): A substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant. Physiol.* **2004**, *134*, 890–897.
173. Matsumura, H.; Ito, A.; Saitoh, H.; Winter, P.; Kahl, G.; Reuter, M.; Kruger, D.H.; Terauchi, R. SuperSAGE. *Cell. Microbiol.* **2005**, *7*, 11–18.

174. Brenner, S.; Johnson, M.; Bridgham, J.; Golda, G.; Lloyd, D.H.; Johnson, D.; Luo, S.; McCurdy, S.; Foy, M.; Ewan, M.; *et al.* Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat. Biotechnol.* **2000**, *18*, 630–634.
175. Ng, P.; Wei, C.L.; Sung, W.K.; Chiu, K.P.; Lipovich, L.; Ang, C.C.; Gupta, S.; Shahab, A.; Ridwan, A.; Wong, C.H.; *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nat. Methods* **2005**, *2*, 105–111.
176. Schaefer, B.C. Revolutions in rapid amplification of cDNA ends: New strategies for polymerase chain reaction cloning of full-length cDNA ends. *Anal. Biochem.* **1995**, *227*, 255–273.
177. Kapranov, P.; Drenkow, J.; Cheng, J.; Long, J.; Helt, G.; Dike, S.; Gingeras, T.R. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res.* **2005**, *15*, 987–997.
178. Olivarius, S.; Plessy, C.; Carninci, P. High-throughput verification of transcriptional starting sites by Deep-RACE. *BioTechniques* **2009**, *46*, 130–132.
179. Kodzius, R.; Kojima, M.; Nishiyori, H.; Nakamura, M.; Fukuda, S.; Tagami, M.; Sasaki, D.; Imamura, K.; Kai, C.; Harbers, M.; *et al.* CAGE: Cap analysis of gene expression. *Nat. Methods* **2006**, *3*, 211–222.
180. Shiraki, T.; Kondo, S.; Katayama, S.; Waki, K.; Kasukawa, T.; Kawaji, H.; Kodzius, R.; Watahiki, A.; Nakamura, M.; Arakawa, T.; *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 15776–15781.
181. Mercer, T.R.; Gerhardt, D.J.; Dinger, M.E.; Crawford, J.; Trapnell, C.; Jeddeloh, J.A.; Mattick, J.S.; Rinn, J.L. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.* **2011**, *30*, 99–104.
182. Mathavan, S.; Lee, S.G.; Mak, A.; Miller, L.D.; Murthy, K.R.; Govindarajan, K.R.; Tong, Y.; Wu, Y.L.; Lam, S.H.; Yang, H.; *et al.* Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet.* **2005**, *1*, 260–276.
183. Wang, E.T.; Sandberg, R.; Luo, S.; Khrebtukova, I.; Zhang, L.; Mayr, C.; Kingsmore, S.F.; Schroth, G.P.; Burge, C.B. Alternative isoform regulation in human tissue transcriptomes. *Nature* **2008**, *456*, 470–476.
184. Graveley, B.R.; Brooks, A.N.; Carlson, J.W.; Duff, M.O.; Landolin, J.M.; Yang, L.; Artieri, C.G.; van Baren, M.J.; Boley, N.; Booth, B.W.; *et al.* The developmental transcriptome of Drosophila melanogaster. *Nature* **2011**, *471*, 473–479.
185. Pang, K.C.; Stephen, S.; Dinger, M.E.; Engström, P.G.; Lenhard, B.; Mattick, J.S. RNAdb 2.0—An expanded database of mammalian non-coding RNAs. *Nucleic Acids Res.* **2007**, *35*, D178–D182.
186. Kin, T.; Yamada, K.; Terai, G.; Okida, H.; Yoshinari, Y.; Ono, Y.; Kojima, A.; Kimura, Y.; Komori, T.; Asai, K. fRNAdb: A platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Res.* **2007**, *35*, D145–D148.
187. Liu, C.; Bai, B.; Skogerbø, G.; Cai, L.; Deng, W.; Zhang, Y.; Bu, D.; Zhao, Y.; Chen, R. NONCODE: An integrated knowledge database of non-coding RNAs. *Nucleic Acids Res.* **2005**, *33*, D112–D125.

188. Pruitt, K.D.; Tatusova, T.; Klimke, W.; Maglott, D.R. NCBI reference Sequences: Current status, policy and new initiatives. *Nucleic Acids Res.* **2009**, *37*, D32–D36.
189. Harrow, J.; Denoeud, F.; Frankish, A.; Reymond, A.; Chen, C.K.; Chrast, J.; Lagarde, J.; Gilbert, J.G.; Storey, R.; Swarbreck, D.; *et al.* GENCODE: Producing a reference annotation for ENCODE. *Genome Biol.* **2006**, *7*, S4 1–S4 9.
190. Havana team. Available online: http://www.sanger.ac.uk/research/projects/vertebrategenome/havana/ (accessed 6 March 2013).
191. Loveland, J.E.; Gilbert, J.G.; Griffiths, E.; Harrow, J.L. Community gene annotation in practice. *Database (Oxford)* **2012**, *2012*, doi:10.1093/database/bas009.
192. Flicek, P.; Amode, M.R.; Barrell, D.; Beal, K.; Brent, S.; Carvalho-Silva, D.; Clapham, P.; Coates, G.; Fairley, S.; Fitzgerald, S.; *et al.* Ensembl 2012. *Nucleic Acids Res.* **2012**, *40*, D84–D90.
193. Kawai, J.; Shinagawa, A.; Shibata, K.; Yoshino, M.; Itoh, M.; Ishii, Y.; Arakawa, T.; Hara, A.; Fukunishi, Y.; Konno, H.; *et al.* Functional annotation of a full-length mouse cDNA collection. *Nature* **2001**, *409*, 685–690.
194. Pruitt, K.D.; Tatusova, T.; Brown, G.R.; Maglott, D.R. NCBI reference sequences (RefSeq): Current status, new features and genome annotation policy. *Nucleic Acids Res.* **2012**, *40*, D130–D135.
195. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006.
196. Stalker, J.; Gibbins, B.; Meidl, P.; Smith, J.; Spooner, W.; Hotz, H.R.; Cox, A.V. The Ensembl web site: Mechanics of a genome browser. *Genome Res.* **2004**, *14*, 951–955.
197. Wilming, L.G.; Gilbert, J.G.; Howe, K.; Trevanion, S.; Hubbard, T.; Harrow, J.L. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.* **2008**, *36*, D753–D760.
198. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **2012**, *489*, 57–74.
199. Martin, J.A.; Wang, Z. Next-generation transcriptome assembly. *Nat. Rev. Genet.* **2011**, *12*, 671–682.
200. Rogers, M.F.; Thomas, J.; Reddy, A.S.; Ben-Hur, A. SpliceGrapher: Detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.* **2012**, *13*, R4.
201. Metzker, M.L. Sequencing technologies—the next generation. *Nat. Rev. Genet.* **2010**, *11*, 31–46.
202. Butler, J.; MacCallum, I.; Kleber, M.; Shlyakhter, I.A.; Belmonte, M.K.; Lander, E.S.; Nusbaum, C.; Jaffe, D.B. ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **2008**, *18*, 810–820.
203. Zerbino, D.R.; Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **2008**, *18*, 821–829.
204. Simpson, J.T.; Wong, K.; Jackman, S.D.; Schein, J.E.; Jones, S.J.; Birol, I. ABySS: A parallel assembler for short read sequence data. *Genome Res.* **2009**, *19*, 1117–1123.
205. Li, R.; Li, Y.; Kristiansen, K.; Wang, J. SOAP: Short oligonucleotide alignment program. *Bioinformatics* **2008**, *24*, 713–714.
206. Lin, H.; Zhang, Z.; Zhang, M.Q.; Ma, B.; Li, M. ZOOM! Zillions of oligos mapped. *Bioinformatics* **2008**, *24*, 2431–2437.

207. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25.
208. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760.
209. Schatz, M.C. CloudBurst: Highly sensitive read mapping with MapReduce. *Bioinformatics* **2009**, *25*, 1363–1369.
210. Ahmadi, A.; Behm, A.; Honnalli, N.; Li, C.; Weng, L.; Xie, X. Hobbes: Optimized gram-based methods for efficient read alignment. *Nucleic Acids Res.* **2012**, *40*, e41.
211. Derrien, T.; Estelle, J.; Marco Sola, S.; Knowles, D.G.; Raineri, E.; Guigo, R.; Ribeca, P. Fast computation and applications of genome mappability. *PLoS One* **2012**, *7*, e30377.
212. Cloonan, N.; Xu, Q.; Faulkner, G.J.; Taylor, D.F.; Tang, D.T.; Kolle, G.; Grimmond, S.M. RNA-MATE: A recursive mapping strategy for high-throughput RNA-sequencing data. *Bioinformatics* **2009**, *25*, 2615–2616.
213. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25*, 1105–1111.
214. Li, H.; Ruan, J.; Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **2008**, *18*, 1851–1858.
215. Smith, A.D.; Xuan, Z.; Zhang, M.Q. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC Bioinforma.* **2008**, *9*, 128.
216. Ewing, B.; Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **1998**, *8*, 186–194.
217. Ewing, B.; Hillier, L.; Wendl, M.C.; Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **1998**, *8*, 175–185.
218. Trapnell, C.; Williams, B.A.; Pertea, G.; Mortazavi, A.; Kwan, G.; van Baren, M.J.; Salzberg, S.L.; Wold, B.J.; Pachter, L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **2010**, *28*, 511–515.
219. Li, W.; Feng, J.; Jiang, T. IsoLasso: A LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* **2011**, *18*, 1693–1707.
220. Palmieri, N.; Nolte, V.; Suvorov, A.; Kosiol, C.; Schlötterer, C. Evaluation of different reference based annotation strategies using RNA-Seq—A case study in *Drososphila pseudoobscura*. *PLoS One* **2012**, doi:10.1371/journal.pone.0046415.
221. Garg, R.; Patel, R.K.; Tyagi, A.K.; Jain, M. *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* **2011**, *18*, 53–63.
222. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652.
223. Jager, M.; Ott, C.E.; Grunhagen, J.; Hecht, J.; Schell, H.; Mundlos, S.; Duda, G.N.; Robinson, P.N.; Lienau, J. Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing. *BMC Genomics* **2011**, *12*, 158.
224. Keiler, K.C. Biology of trans-translation. *Annu. Rev. Microbiol.* **2008**, *62*, 133–151.

225. Novikova, I.V.; Hennelly, S.P.; Sanbonmatsu, K.Y. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucleic Acids Res.* **2012**, *40*, 5034–5051.
226. Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M.F.; Kellis, M.; Lindblad-Toh, K.; Lander, E.S. Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 19428–19433.
227. Lin, M.F.; Jungreis, I.; Kellis, M. PhyloCSF: A comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* **2011**, *27*, i275–i282.
228. Lin, M.F.; Carlson, J.W.; Crosby, M.A.; Matthews, B.B.; Yu, C.; Park, S.; Wan, K.H.; Schroeder, A.J.; Gramates, L.S.; St Pierre, S.E.; *et al.* Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. *Genome Res.* **2007**, *17*, 1823–1836.
229. Liao, Q.; Liu, C.; Yuan, X.; Kang, S.; Miao, R.; Xiao, H.; Zhao, G.; Luo, H.; Bu, D.; Zhao, H.; *et al.* Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.* **2011**, *39*, 3864–3878.
230. Marchler-Bauer, A.; Panchenko, A.R.; Shoemaker, B.A.; Thiessen, P.A.; Geer, L.Y.; Bryant, S.H. CDD: A database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **2002**, *30*, 281–283.
231. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **2012**, *40*, D71–D75.
232. Kong, L.; Zhang, Y.; Ye, Z.-Q.; Liu, X.-Q.; Zhao, S.-Q.; Wei, L.; Gao, G. CPC: Assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.* **2007**, *35*, W345–W349.
233. Arrial, R.T.; Togawa, R.C.; Brigido, M.M. Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT: Case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinforma.* **2009**, *10*, 239.
234. Wang, L.; Park, H.J.; Dasari, S.; Wang, S.; Kocher, J.-P.; Li, W. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **2013**, *41*, e74.
235. Mikkelsen, T.S.; Ku, M.; Jaffe, D.B.; Issac, B.; Lieberman, E.; Giannoukos, G.; Alvarez, P.; Brockman, W.; Kim, T.K.; Koche, R.P.; *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **2007**, *448*, 553–560.
236. Lu, Z.J.; Yip, K.Y.; Wang, G.; Shou, C.; Hillier, L.W.; Khurana, E.; Agarwal, A.; Auerbach, R.; Rozowsky, J.; Cheng, C.; *et al.* Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. *Genome Res.* **2011**, *21*, 276–285.
237. Esteve-Codina, A.; Kofler, R.; Palmieri, N.; Bussotti, G.; Notredame, C.; Perez-Enciso, M. Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics* **2011**, *12*, 552.
238. Nam, J.W.; Bartel, D. Long non-coding RNAs in *C. elegans*. *Genome Res.* **2012**, *22*, 2529–2540.
239. Djebali, S.; Davis, C.A.; Merkel, A.; Dobin, A.; Lassmann, T.; Mortazavi, A.; Tanzer, A.; Lagarde, J.; Lin, W.; Schlesinger, F.; *et al.* Landscape of transcription in human cells. *Nature* **2012**, *489*, 101–108.

240. Eveland, A.L.; McCarty, D.R.; Koch, K.E. Transcript profiling by 3'-untranslated region sequencing resolves expression of gene families. *Plant Physiol.* **2008**, *146*, 32–44.
241. Hillier, L.W.; Reinke, V.; Green, P.; Hirst, M.; Marra, M.A.; Waterston, R.H. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Genome Res.* **2009**, *19*, 657–666.
242. Nagalakshmi, U.; Wang, Z.; Waern, K.; Shou, C.; Raha, D.; Gerstein, M.; Snyder, M. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **2008**, *320*, 1344–1349.
243. Shendure, J. The beginning of the end for microarrays? *Nat. Methods* **2008**, *5*, 585–587.
244. Morozova, O.; Hirst, M.; Marra, M.A. Applications of new sequencing technologies for transcriptome analysis. *Annu. Rev. Genomics Hum. Genet.* **2009**, *10*, 135–151.
245. Auer, P.L.; Doerge, R.W. Statistical design and analysis of RNA sequencing data. *Genetics* **2010**, *185*, 405–416.
246. Hiller, M.; Pudimat, R.; Busch, A.; Backofen, R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.* **2006**, *34*, e117.
247. Chang, T.-H.; Huang, H.-D.; Chuang, T.-N.; Shien, D.-M.; Horng, J.-T. RNAMST: Efficient and flexible approach for identifying RNA structural homologs. *Nucleic Acids Res.* **2006**, *34*, W423–W428.
248. Yao, Z.; Weinberg, Z.; Ruzzo, W.L. CMfinder—A covariance model based RNA motif finding algorithm. *Bioinformatics* **2006**, *22*, 445–452.
249. Ji, Y.; Xu, X.; Stormo, G.D. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics* **2004**, *20*, 1591–1602.
250. Riccitelli, N.J.; Lupták, A. Computational discovery of folded RNA domains in genomes and *in vitro* selected libraries. *Methods* **2010**, *52*, 133–140.
251. Gautheret, D.; Major, F.; Cedergren, R. Pattern searching/alignment with RNA primary and secondary structures: An effective descriptor for tRNA. *Comput. Appl. Biosci.* **1990**, *6*, 325–331.
252. Gautheret, D.; Lambert, A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **2001**, *313*, 1003–1011.
253. König, J.; Zarnack, K.; Luscombe, N.M.; Ule, J. Protein-RNA interactions: New genomic technologies and perspectives. *Nat. Rev. Genet.* **2011**, *13*, 77–83.
254. Rivas, E.; Eddy, S.R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **2000**, *16*, 583–605.
255. Kertesz, M.; Wan, Y.; Mazor, E.; Rinn, J.L.; Nutter, R.C.; Chang, H.Y.; Segal, E. Genome-wide measurement of RNA secondary structure in yeast. *Nature* **2010**, *467*, 103–107.
256. Slater, G.S.; Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinforma.* **2005**, *6*, 31.
257. Birney, E.; Clamp, M.; Durbin, R. GeneWise and genomewise. *Genome Res.* **2004**, *14*, 988–995.
258. Eyras, E.; Reymond, A.; Castelo, R.; Bye, J.M.; Camara, F.; Flicek, P.; Huckle, E.J.; Parra, G.; Shteynberg, D.D.; Wyss, C.; *et al.* Gene finding in the chicken genome. *BMC Bioinforma.* **2005**, *6*, 131.
259. Mariotti, M.; Guigo, R. Selenoprofiles: Profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics* **2010**, *26*, 2656–2663.

260. Vieira, F.G.; Rozas, J. 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.* **2011**, *3*, 476–490.
261. Latos, P.A.; Pauler, F.M.; Koerner, M.V.; Şenergin, H.B.; Hudson, Q.J.; Stocsits, R.R.; Allhoff, W.; Stricker, S.H.; Klement, R.M.; Warczok, K.E.; *et al.* Airn transcriptional overlap, but not its lncRNA products, induces imprinted Igf2r silencing. *Science* **2012**, *338*, 1469–1472.
262. Santoro, F.; Pauler, F.M. Silencing by the imprinted Airn macro lncRNA: Transcription is the answer. *Cell. Cycle* **2013**, *12*, 711–712.