Traces of Post-Transcriptional RNA Modifications in Deep Sequencing Data

Sven Findeiß[a], David Langenberger[a,b], Peter F. Stadler[a,b,c,d,e,f,g,*], Steve Hoffmann[a,b]

[a]*Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelsrtraße 16-18, D-04107 Leipzig, Germany*
[b]*LIFE – Leipzig Research Center for Civilization Diseases, University of Leipzig, Germany*
[c]*Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*
[d]*Fraunhofer Institut für Zelltherapie und Immunologie, Perlickstraße 1, D-04103 Leipzig, Germany*
[e]*Department of Theoretical Chemistry University of Vienna, Währinger Straße 17, A-1090 Wien, Austria*
[f]*Center for non-coding RNA in Technology and Health, University of Copenhagen, Gønnegårdsvej 3, 1870 Frederiksberg C, Denmark*
[g]*Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*

Abstract

Many aspects of the RNA maturation leave traces in RNA sequencing data in the form of deviations from the reference genomic DNA. This includes in particular genomically non-encoded nucleotides and chemical modifications. The latter leave their signatures in forms of mismatches and conspicuous patterns of sequencing reads. Modified mapping procedures focusing on particular types of deviations can help to unravel post-transcriptional modification, maturation and degradation processes. Here, we focus on small RNA sequencing data that is produced in large quantities aiming at the analysis of microRNA expression. Starting from the recovery of many well-known modified sites in tRNAs we provide evidence that modified nucleotides are a pervasive phenomenon in these data sets. Regarding non-encoded nucleotides we concentrate on CCA tails, which, surprisingly, can be found in a diverse collection of transcripts, including sub-populations of mature microRNAs. Although small RNA sequencing libraries alone are insufficient to obtain a complete picture, they can inform on many aspects of the complex processes of RNA maturation.

Keywords: High throughput sequencing, RNA editing, tRNAs, microRNAs, CCA enzyme, RNA modification

Introduction

Mature functional RNAs frequently deviate from their DNA templates. The maturation of a primary RNA transcript usually involves various forms of RNA processing (such as endo- and exonucleolytic trimming, splicing, or polyadenylation). More than a dozen mechanistically distinct types of RNA editing, i.e., targeted nucleotide insertions, deletions, and exchanges, have been described in a wide diversity of clades (Knoop, 2010). Chemical modifications, furthermore, introduce a variety of non-standard nucleotides and affect the majority of non-coding RNAs (ncRNAs) (Ishitani et al., 2008). As a consequence, a mature RNA sequence may differ substantially from its genomic DNA template. RNA editing and modification can have massive effects on both the secondary structure and the interpretation of mRNAs. Chemical modifications in tRNAs, for instance, are instrumental for the integrity of their 3D structures. A→I editing, on the other hand, influences protein sequences since I is read as G by the translation machinery.

Most eukaryotic and many prokaryotic RNAs undergo processing at their 3'-ends. Following cleavage or trimming of the primary transcript, additional nucleotides that are not encoded in the genome are added in many cases. The best-known examples are the polyadenylation of most mRNAs (Millevoi and Vagner, 2010) and the addition of CCA to the 3'-end of tRNAs (Phizicky and Hopper, 2010). Several ncRNAs, including signal recognition particle (SRP) RNA, U2 small nuclear RNA (snRNA) and 7SK RNAs are post-transcriptionally adenylated; U6 snRNA and ribosomal 5S RNA can be both adenylated and uridylated on their 3'-ends (Chen et al., 2000; Perumal et al., 2000; Peru-

*Corresponding author at: Bioinformatics Group, Dept. of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany

Email addresses: sven@bioinf.uni-leipzig.de (Sven Findeiß), david@bioinf.uni-leipzig.de (David Langenberger), studla@bioinf.uni-leipzig.de (Peter F. Stadler), steve@bioinf.uni-leipzig.de (Steve Hoffmann)

mal and Reddy, 2002). Several mature microRNAs are also 3'-adenylated and/or 3'-uridylated (Katoh et al., 2009; Lu et al., 2009; Ebhardt et al., 2009; Burroughs et al., 2010; Fernandez-Valverde et al., 2010).

The emergence of high throughput RNA sequencing (RNAseq) technologies offers, for the first time in history, the possibility to systematically analyze whole transcriptomes at a comparably low cost (Wang et al., 2009). Consequentially, a substantial and rapidly growing body of transcriptome sequence data has accumulated covering a large number of species across all kingdoms. These data pose an enormous challenge and opportunity to computer science and biology alike (Editorial Nature Biotech 26). Public data bases are likely to encompass unburied treasures. This gold rush, however, may be significantly slowed down since little is known about potential sources of error and bias. Even less is understood about the biology of many RNA species. Compared to DNA sequencing, cDNA sequences exhibit much higher error rates often resulting in frustrating alignment results. Depending on sequencing technology, cDNA preparation protocol, and organism under investigation about 20% of the sequences may not be alignable to the reference genome (Li et al., 2010). This may be caused by mismatches, insertions, or deletions, as well as strict mapping policies to purge reads with multiple hits in the reference genome. Disregarding technical artifacts in the RNA sequence read and errors or missing data in the reference genome, which make it impossible to map the read at all, there are at least three reasons why RNA reads do not match exactly to the reference genome: (1) sequencing errors, (2) polymorphisms, and (3) RNA maturation. Hence, analysis of RNAseq data in general requires a significantly higher sensitivity in comparison to DNA variation analysis.

In this contribution we are concerned with the question to what extent chemical modifications, editing, and non-encoded nucleotides in matured RNAs are visible in deep sequencing data. Previous work already indicates that such an approach is feasible: Analyzing reads that map with a single mismatch to the genome, more than 1000 sites with possible RNA base modifications were found in *Arabidopsis thaliana* and *Oryza sativa*, predominantly in tRNAs, microRNAs, and rRNAs (Iida et al., 2009; Ebhardt et al., 2009).

A primary source of information on ncRNAs are short read libraries that are prepared and analyzed with a focus on microRNAs. Here, total RNA is size-selected so that RNAs larger than 30 nt, and hence all complete "house-keeping" ncRNAs, are removed before sequencing. Surprisingly, these libraries contain a large number of reads deriving from nearly all ncRNA classes (Kawaji and Hayashizaki, 2008). These originate from cleavage of larger transcripts. For instance, tRNAs are under certain conditions specifically cleaved into fragments of different lengths in the anticodon loop or anticodon left arm (Lee and Collins, 2005; Li et al., 2008; Jöchl et al., 2008). MicroRNA-sized products are derived from position specific processing at the 5'- or 3'-end of mature or precursor tRNAs (Cole et al., 2009; Lee et al., 2009). Such small RNA fragments, for which in individual cases a microRNA-like function has been demonstrated, are also derived from small nucleolar RNAs (Taft et al., 2009), vault RNAs (Persson et al., 2009; Stadler et al., 2009), and Y RNAs (Meiri et al., 2010), as well as some long ncRNAs such as MALAT1 (Stadler, 2010). The generation of these small RNA fragments is tied closely to the stable double-stranded regions in the parental RNA (Langenberger et al., 2010). Here, we set out to explore to what extent small RNA sequencing libraries are suitable for a *systematic* investigation of RNA maturation.

Results

We have analyzed a combination of two RNA libraries obtained from *Homo sapiens* (human) and *Macaca mulatta* (Rhesus macaque) brains, respectively (Somel et al., 2010).

Table 1: Statistics of the data sets used in this study. Reads with identical sequences were merged into tags. All tags matching to the mitochondrial genome of human and macaque, respectively, were removed to avoid contamination of nuclear copies of mitochondrial DNA (NUMT). Overlapping tags that survived the filtering steps were joined into blocks.

	Human	Macaque
Entire library		
reads	71,307,445	114,619,534
tags	355,453	14,240,332
3'-CCA tails		
tags	3,925	138,895
NUMT cleaned tags	3,017	118,298
non-genomically encoded and NUMT filtered 3'-CCA tails		
tags	1,431	90,208
blocks	246	1,289

In our analysis we distinguish between individual reads and tags. A *tag* is defined as a DNA sequence that occurs at least once in a set of sequencer reads. Thus a

tag typically corresponds to several identical reads. The advantage of using tags lies in a drastic reduction of data that have to be handled. A statistical overview of the analyzed data sets is given in Table 1.

Inference of chemical modifications from mismatches

Some chemical modifications of nucleotides are detectable as mismatches between RNAseq data and the genomic reference. In contrast to PCR artifacts the mismatches appear in many different tags, and the frequency distribution of nucleotides deviates from that expected for SNPs. Two recent studies showed that tRNA modifications are detectable in plants (Iida et al., 2009; Ebhardt et al., 2009). Figure 1 shows that this is the case also in mammals, using the well-known 1-methyladenosine modification found at position 58 of many tRNAs (Roovers et al., 2004) as an example.

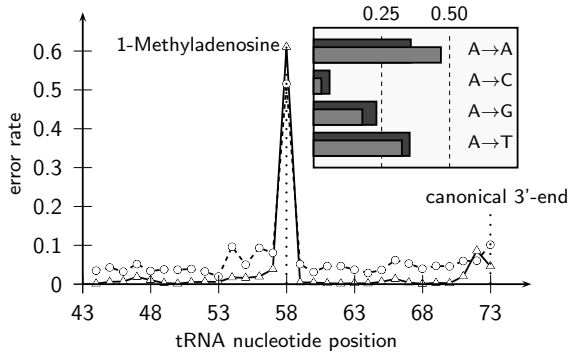

Figure 1: Frequency of mismatches between RNAseq reads and genomic reference sequence for reads mapping close to the 3'-ends of human (solid line; triangles) and macaque (dashed line; circles) tRNAs. Note that the distribution of the error rates is highly correlated between the two species. The inset shows the frequency of nucleotides observed at position 58. Apart from the being read "correctly" as A, this post-transcriptional modification is typically seen as an A-to-T transversion by the sequencer. Again, human (dark gray bars) and macaque (light gray bars) have highly similar substitution patterns.

The 1-methyl-adenosine modification is pivotal for the stability and thus the function of tRNAs (Anderson et al., 1998). It has been reported that the methylated adenosine residue 58 serves as a pause signal for plus-strand strong-stop DNA synthesis and termination site during reverse transcription (Renda et al., 2001). A closer look at the data (inset in Figure 1) shows that this particular modification of adenine is typically interpreted by the sequencer as an A-to-T transversion or an A-to-G transition. Strikingly, the substitution matrix for this modification seems to be largely invariant to the library preparation.

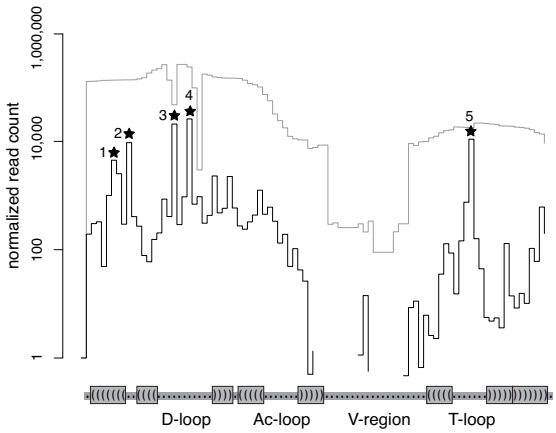

Figure 2: Normalized read counts of coverage (gray) and variation (black) along a tRNA structure, given as dot-bracket notation. Gaps in the coverage within the T-loop and V-region are caused by gaps in the sequence alignment used for normalization. Peaks (numbered stars) along the variation curve correspond to common tRNA modifications at the respective position. The modification 5★ within the T-loop corresponds to the 1-methyladenosine modification present in most tRNA sequences. Other modifications are N2-methylguanosine (1★), 1-methylguanosine (2★), 2-O-methylguanosine (3★) and dihydrouridine (4★).

This modification is the most prominent one that is directly visible from the superposition of the error profiles of all tRNAs. Several other modifications are detectable as conspicuous accumulations of mismatches in individual tRNAs. Notably, most of the detectable variations are located either towards the 5'- or towards the 3'-end of the tRNA. This is caused by the very uneven coverage of tRNAs with small sequencing reads, which is heavily biased towards the ends and the fact that the error-prone 3'- and 5'-termini of sequencing reads naturally coincide with the ends of the tRNA (see Figure 2). Besides the very strong effect of 1-methyladenosine on the accuracy of the cDNA, RNAseq data additionally exhibits moderately increased error rates for dihydrouridines and methylguanosine modifications such as N2-methylguanosines. This is consistent with the findings that the major substitution sites in plant tRNAs correspond to known RNA base modifications: N1-methyladenosine (m1A), N2-methylguanosine (m2G), and N2,N2-methylguanosine (m22G) (Iida et al., 2009; Ebhardt et al., 2009).

Figure 3 summarizes the genomic distribution of mismatches that can be interpreted as sites of chemical modifications. As a consequence of the short RNA sequencing protocol, tRNAs and miRNAs, as measured by the proportion of their genome sequence, accumulate most of the sequence variations seen in the data sets

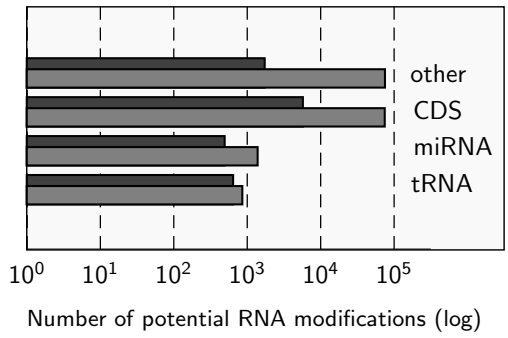

Figure 3: Distribution of potential modifications among human (dark gray) and macaque (light gray) RNA classes. Despite the fact that the data set generated in the macaque RNAseq experiments is about 50-times larger than the human one (Table 1) the total number of possible modifications in tRNA and miRNA does not differ proportionally. In the case of tRNAs, 862 modifications were found in macaque, compared to 657 in human sets; for microRNAs, there are 1400 potential modifications in macaque and about 500 in human.

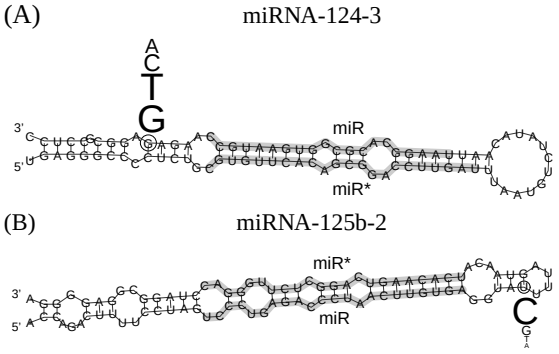

Figure 4: Secondary structure of human precursor miRNA sequences. miR and miR* are shaded in gray. Letter size is proportional to the frequency of observed nucleotides at highlight positions. (A) Sequencing data for the mir124-3 exhibits an unusually high error rate at position 77 (encircled nucleotide). The guanine is frequently replaced by all other three bases indicating RNA modification rather than genomic variation. The modification is located at the 3'-end of the miRNA precursor sequence. (B) mir-125b-2 shows sequence variation at position 43 (encircled nucleotide).

used here. For tRNAs, the numbers are comparable between human and macaque (657 vs. 862), indicating that even the comparably small human library is nearly saturated, while for microRNAs there are about 500 positions with significant sequence variation in human and 1400 in macaque, leading us to expect that the true number of putative modifications in human microRNAs will be larger than observed here. Such an saturation effect cannot be expected for long transcripts, including CDSs, as a consequence of the small RNA sequencing protocol. The much higher difference of observable nucleotide variations within these classes clearly supports this assumption. Surprisingly, more than two third of the detectable loci in human fall into coding sequences, while only 8% are located in tRNAs. In macaque, the ratio is even smaller, suggesting that chemical modifications indicated by the observed sequence variations are a common phenomenon in general.

A detailed analysis of individual modification sites with respect to their substitution patterns requires a sufficiently large coverage. We have identified two frequent potential modifications in human mir-124-3 and mir-125b-2 transcript tags. In the case of the human mir-124-3 (Figure 4) 33 different tags cover nucleotide 77G but only 11 tags are in accordance with the reference sequence. The other nucleotides A, C, and T were counted 6, 5, and 11 times, respectively. No genomic polymorphisms were previously described at this position. Whether the substitutions are caused by a methylation or another modification remains to be clarified. Mir-124-3 is epigenetically silenced by heavy methylation and exhibits tumor-suppressive potential in hepatocellular carcinoma (Furuta et al., 2009). It is also involved in neuronal differentiation (Makeyev et al., 2007) and might be specific for brain libraries. A homolog of mir-124-3 in macaque has not been reported yet.

For position 43T of the human mir-125b-2 the reference base was inferred only 4 times from the cDNA data. Moreover, it was 17 times replaced by C, 3 times by an A and 6 times by a G (see Figure 4). No sequence variation has been found for the homologous position in macaque mml-mir-125b-2. Hence, the modification remains to be verified in independent experiments. Mir-125b has a profound influence on the proliferation of differentiated cancer cells in depletion experiments (Lee et al., 2005). A recent study showed a strong correlation of mir-125b-2 expression and survival rates in childhood leukemia (Gefen et al., 2010).

Inference of chemical modifications from read patterns

Some nucleotide modifications act as road blocks for the reverse transcriptase (Motorin et al., 2007). Thus we expect to observe non-random termination of DNA products from the initial reverse transcription of the RNA. Since the sequencing protocols used to generate the data that we analyzed here are strand-specific, sequencing reads are reported in the reading direction of the original RNA in the sample. An obstacle in the reverse transcription step thus results in the enrichment of

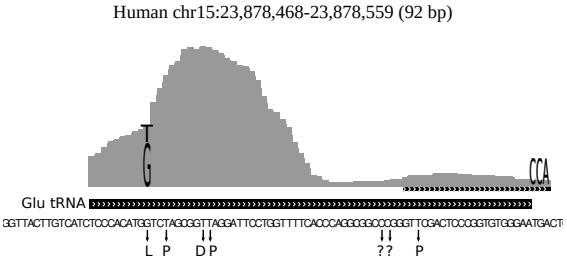

Figure 5: Number of tags covering a given sequence position in a human tRNA-Glu gene. Known chemical modifications are indicated below the genomic reference sequence. (See Table 2 for the key to symbols mapping.) The most prominent modification, the N2-methylguanosine (L) at position 10, is detectable as a G-to-T transversion in 23% of all tags as well as a sharp increase of read starts at the following position. The block of tRNA 3-end reads including the genomically not encoded CCA is indicated. Also note that the read coverage is smallest around the anti-codon.

mapped read starts at the position *following* the chemical modification (Motorin et al., 2007). As shown in Figure 5, this leads to an upward jump of the read and tag coverage at the position following the modification.

In order to determine whether this effect can be seen in the analyzed RNAseq data, we compared the start positions of reads with the positions of known modifications in human tRNA sequences compiled in the `tRNAdb` (Jühling et al., 2009). For example, we observe a nearly 7-fold enrichment of read starts on tRNA position 59 compared to the modified position 58 (Table 2), corresponding to reverse transcription products that terminate before the modified base. Some of these reads extend beyond the tRNase Z processing site and hence derive from the unprocessed precursor. This suggests that these modifications might precede the formation of the 3'-terminus.

A road block function of several modifications is observable in our data in particular for modifications close to the 5'- and 3'-ends of the tRNA. For instance, the N4-acetylcytidines and N2-methylguanosines modifications, which are located close to the 5'-end of mature tRNAs, are detectable by a high incidence of reads starting immediately downstream of the site of modification. Many of the more centrally located modifications are not detectable. This bias is largely caused by the imbalance in the read coverage, which is much higher towards the 5'- and 3'-ends of tRNA (Figure 2).

Nucleotidyltranferases add CCA tails not only to tRNAs

tRNA biogenesis involves multiple maturation steps of the primary RNA polymerase-III transcripts: removal of the 5'-leader, trimming of the 3'-trailer, addition

Table 2: Patterns of read starts and tRNA modifications. The first two columns give the common name of the modification and its RNAMods abbreviation (Dunin-Horkawicz et al., 2006); #: number of experimentally verified modifications in human tRNAs; pos: median position of the modification within tRNAs; exp: number of genomic loci for which a modification is expected and reads are mapped. The last column (ratio) gives the number of read starts one nt downstream of the modification divided by the number of read starts observed at the modified position. Road block modifications that impair reverse transcription are expected to exhibit large ratios.

modification	*	#	exp	pos	ratio
N4-acetylcytidine	M	2	38	2.5	12.80
5-methylcytidine	?	17	298	49	10.70
1-methyladenosine	"	12	188	58	7.17
N2-methylguanosine	L	12	164	9.5	2.66
1-methylguanosine	K	4	52	19	1.92
5-methyluridine	T	6	78	54	1.60
pseudouridine	P	39	529	33	1.28
2-methyladenosine	\	1	18	54	0.59
dihydrouridine	D	29	404	19	0.46
7-methylguanosine	7	6	87	46	0.45
2-O-methylcytidine	B	4	53	32.5	0.08
2-O-methyluridine	J	5	82	33	0.07
2-O-methylguanosine	#	4	46	26	0.02

of CCA, splicing of introns that may be present, and chemical modification of multiple nucleoside residues (Hartmann et al., 2009; Phizicky and Hopper, 2010). Enzymatic CCA addition by nucleotidyltransferases is considered essential for tRNA biosynthesis in all organisms that do not encode CCA termini at the genomic level. These enzymes are able to add specific nucleotides or nucleotide sequences in the absence of genetic templates (Xiong and Steitz, 2004). Although CCA nucleotidyltranferases (also named CCA-adding enzyme) primarily recognize and process tRNAs, see (Vörtler and Mörl, 2010) for a review, they are known to have non-tRNA substrates as well. Many of them are tRNA-like elements such as the mascRNA processed from the 3'-terminus of MALAT1, a long non-coding RNA that is known to be mis-regulated in many human cancers (Wilusz et al., 2008) and viral RNA motifs that act to attract the host's processing machinery, see e.g. (Bogerd et al., 2010). An other prominent example is the U2 small nuclear RNA (snRNA), which undergoes a maturation process similar to that of tRNAs. After removal of a 3'-trailer by a 3' exonuclease, the trimmed U2 snRNA is processed by the human CCA-adding enzyme (Cho et al., 2002). CCA addition has also been observed for several mitochondrial mRNAs (Williams

Table 3: Distribution of blocks with non-genomically encoded CCA tails in different RNA classes. RNA classification was derived from the tRNAscan-SE prediction and the downloaded RNA gene track. Pseudo ncRNAs lack important features (e.g. parts of the characteristic secondary structure) of their functional siblings. A significant proportion of CCA tags was expectedly found in tRNA loci for both human and macaque.

	Human		Macaque	
	3'-CCA blocks			
tRNA	99	40.2%	219	17.0%
miRNA	52	21.1%	111	8.6%
U2	3	1.2%	3	0.2%
rRNA	2	0.8%	0	0.0%
CDS	2	0.8%	2	0.1%
pseudo ncRNA	10	4.1%	13	0.9%
unknown	79	32.1%	941	73.9%
TOTAL	246		1289	

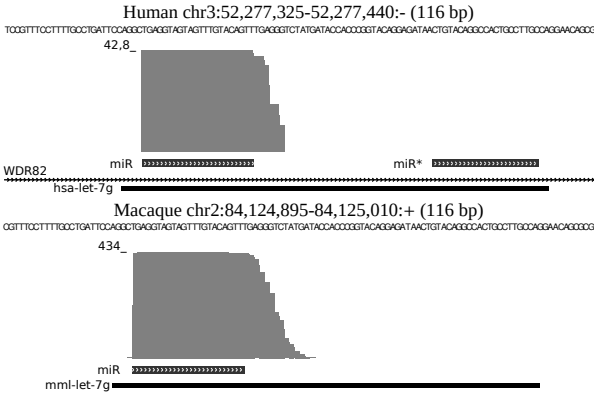

Figure 6: The 5'-end of the let-7g miRNA shows a highly expressed RNA species with non-genomically encoded 3'-CCA modification in human (top) and macaque (bottom). In both species the mature sequence (miR) is covered by this RNA species. Note that the coverage with 3'-CCA tailed reads of the macaque homolog is 10-fold higher than in human but still shows the same pattern.

et al., 2000).

These examples of non-standard 3'-CCA tails are also visible in the RNAseq data sets for both human and macaque (Table 3). We therefore asked whether additional substrates of the CCA nucleotidyltranferase can be identified. A large fraction of the tags matches mitochondrial sequences. After removing those (see Methods), we retained 246 blocks in human (and 1289 blocks in the much larger macaque library). The moderate 5-fold increase of CCA blocks in rhesus monkey compared to human despite a 50-fold higher coverage in the library suggests that the detected set of CCA-tagged RNAs approached saturation in the monkey library. It can only be speculated that for the same reason a majority of 73.9% CCA blocks maps to unknown locations in macaque while only 32.1% of human CCA blocks fall into the same category.

As expected, a substantial fraction of blocks (and the majority of reads and tags) belongs to tRNAs (Table 3). There is, however, a large number of microRNAs with non-encoded CCA ends. Particularly prominent targets are the members of the let-7 family. Almost all of them have read blocks with CCA tails covering the mature miR or miR* sequences (see example in Figure 6). This pattern is shared by 40 human miRNAs and is conserved in macaque. Since the mature miR, as in the example of let-7g, is located on the 5'-arm of the precursor hairpin, it follows that the CCA nucleotidyltranferase does not target the stably base-paired precursor hairpin, as one might expect. Instead, the substrate is either the mature miR or a double-stranded processing intermediate.

One of the few CCA-tagged RNA sequences that arises from coding regions is a small transcript located anti-sense (human chr2:121460406-121460426:-) to the the zinc-finger protein GLI-2 .

Processing of immature and mature tRNAs

The analysis of tRNA loci, surprisingly, shows evidence for the production of small RNA species not only from mature tRNAs but also from unprocessed precursors. This is evidenced, on the one hand, by reads with CCA ends extending the genomically encoded 3'-end and reads showing the hallmarks of chemical modifications, and on the other hand, by reads spanning across the RNase P (5') and RNase Z (3') cleavage sites (Figure 7). Lee et al. (2009) discovered three types of these short RNA fragments: tRF5 and tRF3 sequences are located at the 5'- and 3'-ends of the mature tRNAs, respectively. tRF5 sequences have the RNase P cleavage site at their 5'-end, while tRF3s have a CCA end at the correct position following the tRNase Z processing site. Thus they derive from a matured tRNA. In contrast, tRF1 sequences are entirely located in the 3'-part of the precursor that is cleaved off by tRNase Z. A detailed study (Haussecker et al., 2010) showed that such tRF1-like small RNAs are involved in the global regulation of RNAi, suggesting that many of them could be functional.

In order to obtain at least a rough relative quantification of mature *versus* precursor processing we quantified the fraction of reads that derived from mature tRNAs and their precursors, respectively (Figure 7). While the tRFs arising from the mature tRNA domi-

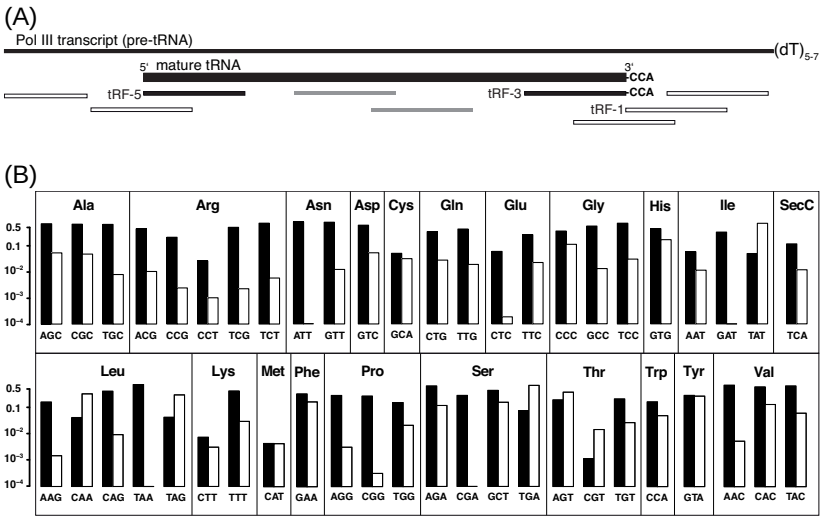

Figure 7: Processing of mature tRNAs and their precursors in human. (A) different types of read blocks derive from different processing stages of tRNAs. Blocks shown as filled boxes are assumed to derive from mature tRNA molecules after RNase P and Z processing. Blocks illustrated as open boxes are located completely or partially outside the mature tRNA region and hence are derived from precursors. The origin of internal reads (grey boxes) cannot be assigned to either mature or precursor tRNA molecules. The classes tRF5, tRF3, and tRF1 of tRNA-derived small RNA fragments were defined in (Lee et al., 2009). (B) fraction of reads mapping to a tRNA locus that are derived from mature tRNAs (black) or precursor sequences (white).

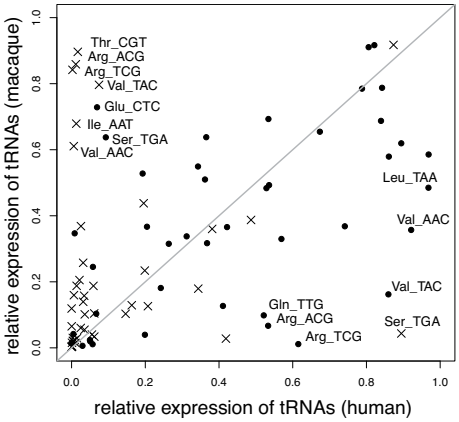

Figure 8: Differences in relative expression of tRFs derived from mature tRNAs (•) and precursors (×) between human an macaque. The tRNAs with the largest deviations between the two species are labeled.

nate in most cases, the situation is different for tRNA-Ile-TAT, tRNA-Leu-CAA, tRNA-Leu-TAG, tRNA-Ser-TGA, tRNA-Thr-AGT, and tRNA-Thr-CGT. Similar data were obtained for human and macaque. When comparing relative expression of tRNAs one would expect that the read counts, normalized relative to the overall coverage and multiple mappings, of different species should show similar expression patterns. There are, however, several significant differences observed between macaque and human (Figure 8).

Interestingly, the most abundant human tRF1-type small RNA fragment, tRF-1001 deriving from tRNA-Ser-TGA, behaves very differently between human and macaque. Lee et al. (2009) observed that tRF-1001 is expressed highly in a wide range of cancer cell lines, where its expression is tightly correlated with proliferation. While the precursor-derived products dominate the human library, most of the macaque sequences arise from the mature tRNA, suggesting that tRF-1001 might be a very recent innovation in human evolution. One might speculate that, like other evolutionarily very recent ncRNAs such as BC1 and BC200 (Kondrashov et al., 2005) or HAR1 (Pollard et al., 2006), at least some of the tRFs detected in brain RNA libraries have functions in brain development.

Discussion

Small RNA sequencing data contain a wealth of information on RNA maturation. We have shown here that potential sites of several types of chemical modifications are detectable as “polymorphic” sites with characteristic substitution patterns as well as through characteristic patterns of read starts. In both cases the signal derives from the obstruction of reverse transcription which constitutes an indispensable first step in RNAseq protocols. In small RNA libraries, the observed signal is necessarily a composite of superimposed effects: (1) the coverage of a particular position by small RNA fragments is strongly biased towards double-stranded regions (Langenberger et al., 2010); (2) modifications

can be detected and reliably distinguished from SNPs or sequencing errors only when the tag coverage is high enough to estimate the substitution frequencies; (3) closely spaced modifications interfere with each other for the following reason: the modification more close to the RNA's 3'-end may act as a road block for reverse transcriptase, as in the case of tRNA position 58, resulting in sharply decreased read coverage 5' of this position. As a consequence, the more 5'-proximal modification will be represented by a much lower read coverage and its signal may vanish below the detection limit.

Our analysis shows that putative modification sites are by no means restricted to tRNAs. In plants, RNA methylation is well known as a crucial step in microRNA biogenesis (Yu et al., 2005). We find that a large number of human microRNAs also appear to be targets of modification processes. While small RNA libraries exhibit signs of chemical modifications mostly for tRNAs, this saturates as the depth of the library increases. Eventually, positions in coding sequences and unannotated genomic regions dominate the data. The fact that several potential modifications as measured by error rates occur so frequently and are, on top, reproducible in different species is a strong indication of functional importance. Given the data presented here, it is not far-fetched to speculate on possible implications of such modifications on transcript structure and function such as alternative splicing, maturation or degradation of RNA.

The most surprising result of our analysis is the widespread 3'-terminal post-transcriptional extension of mature microRNAs by CCA nucleotidyltransferases in addition to the previously reported adenylation and uridylation of mature microRNAs. Interestingly, the CCA tails are attached to the 3'-end of the mature RNAs utilizing a very variable end position. By specifically looking for frequently substituted bases we have identified two microRNAs with potential modification sites.

Although these examples require intensive validation our analysis of deep RNA sequencing data indicate that novel RNAseq technologies may be a time- and cost-effective way to unravel secrets of a poorly understood layer of information: post-transcriptional modification.

Materials and methods

Data Sets and Mapping. We use a combination of small RNA libraries, sequenced on an Illumina platform, from *Homo sapiens* (human) and *Macaca mulatta* (Rhesus macaque) brains, respectively (Somel et al., 2010). The human library comprises 71,307,445 sequencing reads with an average length of 22.04 ± 1.03 nucleotides. With 114,619,534 reads the macaque data set is roughly twice as big as the human one and has a similar read length distribution with an average of 23.19 ± 3.37 nucleotides. To allow error-tolerant mapping of cDNA sequences (>14nt) we used the short read aligner `segemehl` with standard parameters and an E-value cutoff of 500 to also align short sequences (Hoffmann et al., 2009). The `segemehl` software is able to detect mismatches, insertions and deletions alike and reports multiple equally good scoring hits. Multiple best hits to the genome were explicitly allowed. When measuring levels of expression, the number of reads, represented by each tag was divided by the number of hits in the genome with equally good scores. This procedure ensures that the redundancy of multiple (nearly) identical copies (e.g. of tRNAs) is properly taken into account.

Genome sequences and annotation tracks were downloaded from the UCSC genome browser (Kent et al., 2002). Coding sequence (CDS) annotation was taken from the RefSeq gene tracks for both species. An RNA gene track was available for human only. The two primate species are so closely related, however, that all macaque homologs of human ncRNAs considered in this study are reliably identified by a simple `blast` search. MirBase (release 12) was used as source of pre-miRNAs as well as mature miR and miR* sequences and annotation.

CCA ends. To measure the activity of nucleotidyltransferases, tags ending with 3'-CCA were selected. The CCA was removed and the truncated tag was mapped to the reference genome. Tags with a genomically encoded CCA end downstream of the mapped tag were excluded from further analysis. Since short reads deriving from *nu*clear copies of *mit*ochondrial DNA (NUMT) (Hazkani-Covo et al., 2010) and reads truly deriving from the mitochondrial DNA cannot be reliably distinguished in the data at hand, we also excluded all tags matching to the mitochondrial genome of the respective species. Overlapping tags passing the filtering steps were then joined into blocks. Finally, blocks representing less than 10 reads were excluded from further analysis (cf. Table 1).

Analysis of sequence variation. Variation calling was performed with `pileup`, a component of the `SAMTOOLS` package (Li et al., 2009), using standard parameters. In a subsequent filtering step a minimum coverage of 12 sequencing tags that overlap the variation was required.

tRNAs. The tRNAscan-SE program[1] was applied to the reference genomes analyzed in this contribution. The predicted intact tRNAs and pseudogenes, respectively, were treated separately. Positions of tRNA modifications were extracted from the `tRNAdb` (Jühling et al., 2009). To safely map these modifications to the predicted tRNA genes and to avoid biases due to differently sized isoacceptor sequences only tRNA genes coding for the same amino acid and having the same length as those listed in the database were used. This set of mapped tRNA modifications was intersected with the tag variation data obtained from RNAseq read analysis. Raw counts of variant nucleotides were normalized by the number of tags mapping to the position with the variation.

To test whether post-transcriptional modifications are visible in RNAseq data, all blocks overlapping with tRNA 3'-ends were extracted and aligned at the tRNAse Z cleavage site. tRNA sequences, as well as the positions and type of the chemical modifications were retrieved from the `tRNAdb`.

The compiled data set is deposit on the web server of University Leipzig `http://www.bioinf.uni-leipzig.de/publications/supplements/10-036`.

Acknowledgments

We thank Philipp Khaitovich who kindly provided a subset of the sequencing data (recently published (Somel et al., 2010)) before publication. We thank Mark Helm, University Mainz, for his suggestion to investigate positions of read starts as a signature of chemical modifications. This work is supported in part by the *Deutsche Forschungsgemeinschaft* (SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes": STA 850/7-2 to PFS). This publication is supported by LIFE Leipzig Research Center for Civilization Diseases, Universität Leipzig. This project was funded by means of the European Social Fund and the Free State of Saxony.

Anderson, J., Phan, L., Cuesta, R., Carlson, B. A., Pak, M., Asano, K., Björk, G. R., Tamame, M., and Hinnebusch, A. G. (1998). The essential Gcd10p-Gcd14p nuclear complex is required for 1-methyladenosine modification and maturation of initiator methionyl-tRNA. Genes Dev. *12*, 3650–3662.

Bogerd, H. P., Karnowski, H. W., Cai, X., Shin, J. S., Pohlers, M., and Cullen, B. R. (2010). A mammalian herpesvirus uses noncanonical expression and processing mechanisms to generate viral microRNAs. Mol. Cell. *37*, 135–142.

Burroughs, A. M., Ando, Y., de Hoon, M. J., Tomaru, Y., Nishibu, T., Ukekawa, R., Funakoshi, T., Kurokawa, T., Suzuki, H., Hayashizaki, Y., and Daub, C. O. (2010). A comprehensive survey of 3' animal miRNA modification events and a possible role for 3' adenylation in modulating miRNA targeting effectiveness. Genome Res. *20*, 1398–1410.

Chen, Y., Sinha, K., Perumal, K., and Reddy, R. (2000). Effect of 3' terminal adenylic acid residue on the uridylation of human small RNAs *in vitro* and in frog oocytes. RNA *6*, 1277–1288.

Cho, H. D., Tomita, K., Suzuki, T., and Weiner, A. M. (2002). U2 small nuclear RNA is a substrate for the CCA-adding enzyme (tRNA nucleotidyltransferase). J. Biol. Chem. *277*, 3447–3455.

Cole, C., Sobala, A., Lu, C., Thatcher, S. R., Bowman, A., Brown, J. W., Green, P. J., Barton, G. J., and Hutvagner, G. (2009). Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. RNA *15*, 2147–2160.

Dunin-Horkawicz, S., Czerwoniec, A., Gajda, M. J., Feder, M., Grosjean, H., and Bujnicki, J. M. (2006). MODOMICS: a database of RNA modification pathways. Nucleic Acids Res. *34*, D145–D149.

Ebhardt, H. A., Tsang, H. H., Dai, D. C., Liu, Y., Bostan, B., and Fahlman, R. P. (2009). Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. Nucleic Acids Res. *37*, 2461–2470.

Fernandez-Valverde, S. L., Taft, R. J., and Mattick, J. S. (Oct 2010). Dynamic isomiR regulation in Drosophila development. RNA *16*, 1881–1888.

Furuta, M., Kozaki, K.-i., Tanaka, S., Arii, S., Imoto, I., and Inazawa, J. (2009). miR-124 and miR-203 are epigenetically silenced tumor-suppressive microRNAs in hepatocellular carcinoma. Carcinogenesis *31*, 766–776.

Gefen, N., Binder, V., Zaliova, M., Linka, Y., Morrow, M., Novosel, A., Edry, L., Hertzberg, L., Shomron, N., Williams, O., Trka, J., Borkhardt, A., and Izraeli, S. (2010). Hsa-mir-125b-2 is highly expressed in childhood ETV6/RUNX1 (TEL/AML1) leukemias and confers survival advantage to growth inhibitory signals independent of p53. Leukemia *24*, 89–96.

Hartmann, R. K., Gossringer, M., Spath, B., Fischer, S., and Marchfelder, A. (2009). The making of tRNAs and more—RNase P and tRNase Z. Prog. Mol. Biol. Transl. Sci. *85*, 319–368.

Haussecker, D., Huang, Y., Lau, A., Parameswaran, P., Fire, A. Z., and Kay, M. A. (2010). Human tRNA-derived small RNAs in the global regulation of RNA silencing. RNA *16*, 673–695.

Hazkani-Covo, E., Zeller, R., and Martin, W. (2010). Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet. *6*, e1000834.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F., and Hackermuller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput. Biol. *5*, e1000502.

Iida, K., Jin, H., and Zhu, J. K. (2009). Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in *Arabidopsis thaliana*. BMC Genomics *10*, 155.

Ishitani, R., Yokoyama, S., and Nureki, O. (2008). Structure, dynamics, and function of RNA modification enzymes. Curr. Opin. Struct. Biol. *18*, 330–339.

Jöchl, C., Rederstorff, M., Hertel, J., Stadler, P. F., Hofacker, I. L., Schrettl, M., Haas, H., and Hüttenhofer, A. (2008). Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein-synthesis. Nucleic Acids Res. *36*, 2677–2689.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F., and Pütz, J. (2009). `tRNAdb` 2009: compilation of tRNA sequences and tRNA genes. Nucleic Acids Res. *37*, D159–D162.

Katoh, T., Sakaguchi, Y., Miyauchi, K., Suzuki, T., Kashiwabara, S.,

[1]downloaded from `ftp://selab.janelia.org/pub/software/tRNAscan-SE/tRNAscan-SE-1.23.tar.Z`

Baba, T., and Suzuki, T. (2009). Selective stabilization of mammalian microRNAs by 3' adenylation mediated by the cytoplasmic poly(A) polymerase GLD-2. Genes Dev. *23*, 433–438.

Kawaji, H. and Hayashizaki, Y. (2008). Exploration of small RNAs. PLoS Genet. *4*, e22.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Knoop, V. (2010). When you can't trust the DNA: RNA editing changes transcript sequences. Cell Mol. Life Sci.Doi: 10.1007/s00018-010-0538-9.

Kondrashov, A. V., Kiefmann, M., Ebnet, K., Khanam, T., Muddashetty, R. S., and Brosius, J. (2005). Inhibitory effect of naked neural BC1 RNA or BC200 RNA on eukaryotic in vitro translation systems is reversed by poly(A)-binding protein (PABP). J. Mol. Biol. *353*, 88–103.

Langenberger, D., Bermudez-Santana, C., Stadler, P. F., and Hoffmann, S. (2010). Identification and classification of small RNAs in transcriptome sequence data. Pac. Symp. Biocomput. *15*, 80–87.

Lee, S. R. and Collins, K. (2005). Starvation-induced cleavage of the tRNA anticodon loop in *Tetrahymena thermophila*. J. Biol. Chem. *280*, 42744–42749.

Lee, Y. S., Kim, H. K., Chung, S., Kim, K.-S., and Dutta, A. (2005). Depletion of human micro-RNA miR-125b reveals that it is critical for the proliferation of differentiated cells but not for the down-regulation of putative targets during differentiation. J. Biol. Chem. *280*, 16635–16641.

Lee, Y. S., Shibata, Y., Malhotra, A., and Dutta, A. (2009). A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). Genes Dev. *23*, 2639–2649.

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010). RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics *26*, 493–500.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, Y., Luo, J., Zhou, H., Liao, J.-Y., Ma, L.-M., Chen, Y.-Q., and Qu, L.-H. (2008). Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote *Giardia lamblia*. Nucleic Acids Res. *36*, 6048–6055.

Lu, S., Sun, Y. H., and Chiang, V. L. (2009). Adenylation of plant miRNAs. Nucleic Acids Res. *37*, 1878–1885.

Makeyev, E. V., Zhang, J., Carrasco, M. A., and Maniatis, T. (2007). The microRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. Mol. Cell *27*, 435–448.

Meiri, E., Levy, A., Benjamin, H., Ben-David, M., Cohen, L., Dov, A., Dromi, N., Elyakim, E., Yerushalmi, N., Zion, O., Lithwick-Yanai, G., and Sitbon, E. (2010). Discovery of microRNAs and other small RNAs in solid tumors. Nucleic Acids Res. *38*, 6234–6246.

Millevoi, S. and Vagner, S. (2010). Molecular mechanisms of eukaryotic pre-mRNA 3'-end processing regulation. Nucleic Acids Res. *38*, 2757–2774.

Motorin, Y., Muller, S., Behm-Ansmant, I., and Branlant, C. (2007). Identification of modified residues in RNAs by reverse transcription-based methods. Methods Enzymol. *425*, 21–53.

Persson, H., Kvist, A., Vallon-Christersson, J., Medstrand, P., Borg, A., and Rovira, C. (2009). The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. Nat. Cell Biol. *11*, 1268–1271.

Perumal, K., Gu, J., and Reddy, R. (2000). Evolutionary conservation of post-transcriptional 3' end adenylation of small RNAs: *S. cerevisiae* signal recognition particle RNA and U2 small nuclear RNA are post-transcriptionally adenylated. Mol. Cell Biochem. *208*, 99–109.

Perumal, K. and Reddy, R. (2002). The 3' end formation in small RNAs. Gene Expr. *10*, 59–78.

Phizicky, E. M. and Hopper, A. K. H. (2010). tRNA biology charges to the front. Genes Dev. *24*, 1832–1860.

Pollard, K. S., Salama, S. R., Lambert, N., Lambot, M. A., Coppens, S., Pedersen, J. S., Katzman, S., King, B., Onodera, C., Siepel, A., Kern, A. D., Dehay, C., Igel, H., Ares Jr, M., Vanderhaeghen, P., and Haussler, D. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.

Renda, M. J., Rosenblatt, J. D., Klimatcheva, E., Demeter, L. M., Bambara, R. A., and Planelles, V. (2001). Mutation of the methylated tRNA(Lys)(3) residue A58 disrupts reverse transcription and inhibits replication of human immunodeficiency virus type 1. J. Virol. *75*, 9671–9678.

Roovers, M., Wouters, J., Bujnicki, J. M., Tricot, C., Stalon, V., Grosjean, H., and Droogmans, L. (2004). A primordial RNA modification enzyme: the case of tRNA (m1A) methyltransferase. Nucleic Acids Res. *32*, 465–476.

Somel, M., Guo, S., Fu, N., Yan, Z., Hu, H. Y., Xu, Y., Yuan, Y., Ning, Z., Hu, Y., Menzel, C., Hu, H., Lachmann, M., Zeng, R., Chen, W., and Khaitovitch, P. (2010). MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. Genome Res. *20*, 1207–1218.

Stadler, P. F. (2010). Evolution of the long non-coding RNAs MALAT1 and MENβ/ϵ. In: Ferreira, C. E., Miyano, S., and Stadler, P. F. (Eds.), Advances in Bioinformatics and Computational Biology, 5th Brazilian Symposium on Bioinformatics. Vol. 6268 of Lecture Notes in Computer Science. Springer Verlag, Heidelberg, pp. 1–12.

Stadler, P. F., Chen, J. J.-L., Hackermüller, J., Hoffmann, S., Horn, F., Khaitovich, P., Kretzschmar, A. K., Mosig, A., Prohaska, S. J., Qi, X., Schutt, K., and Ullmann, K. (2009). Evolution of vault RNAs. Mol. Biol. Evol. *26*, 1975–1991.

Taft, R. J., Glazov, E. A., Lassmann, T., Hayashizaki, Y., Carninci, P., and Mattick, J. S. (2009). Small RNAs derived from snoRNAs. RNA *15*, 1233–1240.

Vörtler, S. and Mörl, M. (2010). tRNA-nucleotidyltransferases: Highly unusual RNA polymerases with vital functions. FEBS Letters *584*, 297–302.

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

Williams, M. A., Johzuka, Y., and Mulligan, R. M. (2000). Addition of nongenomically encoded nucleotides to the 3'-terminus of maize mitochondrial mRNAs: Truncated rps12 mRNAs frequently terminate with CCA. Nucleic Acids Res. *28*, 4444–4451.

Wilusz, J. E., Freier, S. M., and Spector, D. L. (2008). 3énd processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. Cell *135*, 919–932.

Xiong, Y. and Steitz, T. A. (2004). Mechanism of transfer RNA maturation by CCA-adding enzyme without using an oligonucleotide template. Nature *430*, 640–645.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R. W., Steward, R., and Chen, X. (2005). Methylation as a crucial step in plant microRNA biogenesis. Science *307*, 932–935.