# The early days of RNomics

**ALEXANDER HÜTTENHOFER**
Division of Genomics and RNomics, Core Facility Deep-Sequencing, Innsbruck Medical University - Biocenter, 6020 Innsbruck, Austria

In the late eighties, I became interested in non-protein-coding RNAs (ncRNAs) during the completion of my PhD in Munich (Germany) and continued my interest, as a postdoc, in one of the main centers for ribosomal RNA and protein synthesis research, the lab of Harry Noller at UCSC in California (USA). Shortly after I arrived at Harry's lab, he published one of his hallmark papers in *Science*, demonstrating that ribosomal RNA, mainly devoid of ribosomal proteins, was able to catalyze protein synthesis. This finding implied a catalytic function for ribosomal RNA rather than a scaffolding function. Many of his former students and postdocs, me included, still regret that Harry didn't receive the Nobel prize for his pioneering work on ribosomes. After all, his lab was the first to sequence the 16S and 23S RNAs (rRNAs) from *E. coli*, proposed the first 2D and 3D structure models of rRNAs and the ribosome, and also investigated, by chemical probing, its interaction with tRNA and antibiotics. In addition, he was the first to publish the crystal structure of the 70S ribosome. I know many scientists who would not have taken it lightly, when they missed the Nobel prize by such a small margin, as Harry probably did. However, when I once asked Harry whether he would like to receive the Nobel Prize at some point in his career, he answered: "Hell no, I would rather know how the ribosome really works…".

So that time at UCSC was a very stimulating environment with the most exciting discussions on RNA structure and function, with people like Ted Powers, Ray Samaha, Rachel Green, and Shura Mankin. During that period, I also investigated the interaction of RNase P with its cognate substrate, tRNA by chemical probing in a collaboration with Norman Pace. By chemical probing, we discovered GG residues within a conserved loop region of RNase P which would base-pair to the CC of the CCA-end of tRNAs, a mechanism similar to the recognition of P-site tRNA by the ribosome. By chemical probing and hydroxyl radical foot-printing, I analyzed the interactions of tRNAs and mRNAs with the *E. coli* ribosome, a technique which had been established some years earlier by Tom Cech.

Scientifically and personally, these were a very enjoyable four years as a postdoc in Harry's lab at UCSC and I remember having a tough time upon returning to Munich, where I had stayed previously as a PhD student. In Munich, I worked on the co-translational incorporation of selenocysteine into proteins, which required an mRNA hairpin loop structure interacting with the SELB protein, which we analyzed by chemical probing and modeling, in collaboration with Eric Westhof from the IBMC in Strasbourg.

In 1997, as an associate professor, I moved to the University of Münster (Germany) to the lab of Jürgen Brosius. Jürgen was also a former postdoc of Harry Noller at UCSC, and had sequenced the entire *rrn*B ribosomal RNA operon of *E. coli* for the first time in Harry's lab, which took him three years, a task done today in milliseconds by deep-sequencing. Remarkably, there was not a single error in the entire 7.5 kb sequence. By the German Human Genome Project (HGP) we received considerable funding for a research proposal, initated by Jürgen, on the identification of novel ncRNAs in model organisms. At that time, surprisingly few people were interested in that topic, however. I remember the most enthusiastic response I ever got on that project, from a colleague from the US, was: "Well, if you get funding, you should do it…". While the contribution of Germany with respect to actually sequencing the human genome was rather modest (in the low percentage range, in contrast to that of the US or the UK), I personally feel that the pioneering work on novel ncRNA identification was one of the more significant contributions of Germany to the Human Genome Project and not merely a national "me-too" project.

The basic idea of the ncRNA project proposal was quite simple: Total RNA, isolated from various model organisms and tissues, from mouse to bacteria, was size-separated by denaturing polyacrylamide gel electrophoresis. Subsequently, the small RNA fraction from about 50–500 nt was excised, eluted, linkers were ligated or C-tails added, followed by cDNA synthesis, cloning, and sequencing. Before sequencing, we spotted cDNAs on filters, in collaboration with the group of Hans Lehrach from Berlin, and hybridized these against the most abundant known ncRNAs, e.g., rRNAs, snRNAs, SRP RNA, etc. The reasoning for the lower size-range of 50 nt was that, at that time, the majority of ncRNAs were known to exhibit a size at or above 70 nt (such as

**Corresponding author: alexander.huettenhofer@i-med.ac.at**
Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.049775.115. Freely available online through the *RNA* Open Access option.

© 2015 Hüttenhofer This article, published in *RNA*, is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at http://creativecommons.org/licenses/by-nc/4.0/.

tRNAs). Unfortunately, by this misconception we missed cloning of the now well-known class of microRNAs (miRNAs), located about 2 cm below our "excision window." Recently, Jürgen and I debated—jokingly—about who actually was responsible for this decision (Jürgen claims it was he who suggested the 50 nt margin, while I frankly do not recall whose decision this really was…). Nevertheless, I was truly excited some years later to see Tom Tuschl, Victor Ambros, Gary Ruvkun, and Dave Bartel publishing the identification of numerous novel miRNAs by our ncRNA cloning approach, thereby demonstrating that miRNAs represented a large and abundant class of small RNA species.

At that time in Münster, we analyzed a relatively moderate number of cDNA clones from our ncRNA libraries, i.e., in the range of 2000–5000 clones in total, by Sanger sequencing. Surprisingly, despite this small number of clones, we discovered numerous, and now known to be functional, ncRNAs, in particular novel members of C/D and H/ACA snoRNAs, as well as entirely novel ncRNA candidates. The first cDNA library, encoding ncRNAs sized between 50 and 500 nt, we generated from whole mouse brain. Thereby, we identified, for the first time, brain-specifically expressed snoRNAs from the C/D and H/ACA box class, respectively, and their human homologs. Five of the snoRNA genes were located on human chromosome 15 q11-q13, with two present in multiple gene copies. A 4 Mb deletion of this paternally imprinted locus on chromosome 15, harboring also several protein-coding genes, was previously reported to lead to a neuro-developmental disease, designated as Prader Willi Syndrome (PWS). In 2008, the group of A. Beaudet unambiguously showed that a micro-deletion of about 170,000 bp, encompassing two of the clustered snoRNA genes only, designated as HBII-85 and HBII-52, respectively, was responsible for the etiology of PWS. Thus, this was the first demonstration that brain-specific snoRNAs could cause a severe neurological disease. In the commentary to the paper we published in *PNAS* in 2000, Witek Filipowicz wrote: "it's time for RNomics," a term which we subsequently applied in the following publications to define our research on ncRNAs.

By performing RNAseq in Archaea, i.e., *Archeaglobus fulgidus* and *Sulfolobus solfataricus*, respectively, we also identified small ncRNAs transcribed from repetitive sequences of these archaeal genomes, interrupted by unique sequences. By Northern blotting, we could demonstrate a ladder-like expression and processing pattern of these repeats of one, two, three or longer repeat units, respectively. In the paper, we therefore concluded: "these patterns suggest that in all Archaea examined the clustered repeats are transcribed in the form of long precursor(s), subsequently processed into monomers or multimers of the repeat motif." Up to that point, these repeats had been designated as SRSRs (short regular spaced repeats) and were proposed to be involved in genome replication at the DNA level. Some years later, however, it was demonstrated that these small ncRNAs, which we had identified earlier in Archaea as well as in Bacteria, represented, in fact, the gRNAs (guide RNAs) from the CRISPR/Cas system, an antiviral defence mechanism of bacteria. Nowadays, the gRNA/CRISPR system has widely been employed in manipulation of higher eukaryal genomes and bears great potential of site directed mutagenesis within these genomes. Next to gRNAs, we also identified the first H/ACA snoRNAs in Archaea by RNAseq, exhibiting a tripartite stem–loop structure and sharing a common protein, L7Ae, with C/D box snoRNAs.

We not only had a first glance at the small RNA transcriptomes of Eukarya (*M. musculus*, *D. melanogaster*, and *A. thaliana*), but also Bacteria (*E. coli*) as well as Archaea (*A. fulgidus* and *S. solfataricus*). When I was appointed as a full professor at the University of Innsbruck (Austria), we investigated, in addition, the small ncRNA transcriptomes in chloroplasts and mitochondria as well as from the EBV virus, where we discovered the first viral snoRNA. In all these studies, as mentioned above, a very moderate number of cDNA clones, enconding ncRNAs, was sequenced (i.e., in the range of 2000–5000 clones). Nevertheless, we still discovered a multitude of functional ncRNAs in these model organisms.

Nowadays, by ultra-deep RNAseq, millions of reads are generated. However, we should ask whether these high-throughput approaches are really more suitable to define the ncRNA transcriptome than smaller-scale Sanger sequencing? That is, by sequencing only small numbers of cDNA clones, in the early days of RNomics, we surely had only identified the "tip of the ncRNA iceberg," but maybe only this part indeed represents the functional portion of the ncRNA transcriptome? By ENCODE and other projects, we now know that a large portion of the genome (up to 90%) is transcribed into RNA, but not translated into proteins, resulting in up to 450,000 ncRNA transcripts in Eukarya. However, many of these ncRNA species might represent spurious transcription or turnover products of longer RNAs such as mRNAs (e.g., might be derived from 3′ or 5′ UTRs). Since no gene promoter (i.e., for mRNAs or ncRNAs, respectively) will be absolutely "silent" at all times, this might imply that the deeper one "digs" by ultra-deep sequencing the more likely it is even to pick up single copy RNA transcripts or simply "RNA crap." Personally, my feeling is that one has to be extremely careful to define these rare transcripts as functional ncRNAs, which is not to say that low abundant transcripts might not be functional at least in some cases.

Later on, as a full professor at the Innsbruck Biocenter (Austria), we optimized ncRNA identification based on functionality of ncRNAs. To that end, we isolated ribonucleoprotein particles (RNPs), rather than naked RNA, from cells or tissues; the rationale behind this approach was that, in particular in Eukarya, all functional ncRNAs are known to form RNPs. In addition, differential expression of ncRNAs might also hint to functionality. Thus, we invented a subtractive hybridization method for differential expression of ncRNAs, in analogy to mRNAs, and also generated customized ncRNA micro-arrays for that purpose. Lastly, by the

custom ncRNA micro-array approach we also investigated differential expression of ncRNAs in neurological diseases, another means to enrich functional ncRNAs from the large background of RNA transcripts.

At this time of ncRNA research, one of the most pending and important questions in the field, however, still remains unsolved: While we now know pretty well the number of protein-coding genes in the human genome (i.e., around 20,000), the number of functional ncRNAs in humans and other Eukaryal genomes is still rather ill-defined. Thus, one of the major challenges in the upcoming years will be to define the ncRNA transcriptome (small and large) in model organisms, including humans.

Lastly, as one of those "early scientists" who was, and still is, interested in the identification and function of ncRNAs, what really strikes me these days is the exclusive focus of ncRNAs in gene regulation. It almost seems to me that proteins have completely been forgotten and disappeared from the screen of regulatory factors.

As my colleague Jürgen Brosius pointed out recently: "After a long lag phase on the sidelines, functional RNA currently is in the spotlight of biology, even medicine, as RNomics shows promise to detect additional disease genes, greatly develop the diagnostic toolbox, and revolutionize therapeutic possibilities. However, the development from only two decades ago, when the mere mention of RNA generally exposed grant proposals to monkey hammering, resulting in poorer scores, and the current situation in which a feeding frenzy of RNA discovery, fueled by ultra-deep RNA-sequencing technologies, endorses almost any detected transcript or degradation product as functional RNA borders on the grotesque. Clearly, the pendulum has swung to the other extreme…".

I couldn't agree more. I think we should go back and look at the whole picture of gene regulation by proteins, such as transcription factors, polymerases, UTR binding proteins *and* by ncRNAs and we have to analyze the regulatory network not in a "one on one" fashion but in a Systems Biology approach. Only if we understand the interplay between ncRNAs *and* proteins, we will be able to understand how cells translate their genetic information into biological molecules that determine their structure, function, and specify tissues as well as entire organisms, including humans.