Project: Exploring the World of non-messenger RNAs: RNomics Meets Proteomics

Juergen Brosius - **Wilhelms University, Münster – RNA.world@uni-muenster.de**, Steffen Hennig - RZPD, Berlin, Zoltán Konthur – Max Planck Institute for Molecular Genetics, Berlin, Michal Janitz - Max Planck Institute for Molecular Genetics, Berlin, Ralf Dressel - Georg-August-University, Goettingen, Lutz Walter – German Primate Center, Goettingen, Jürgen Schmitz – Wilhelms University Münster, Timofey Rozhdestvensky - Wilhelms University Münster

Introduction

Only a few years ago, the pervasive role of non-messenger RNA (nmRNA), also termed non-protein-coding RNA (npcRNA) in cellular function and genome evolution was unimaginable. Today, however, a combination of biocomputational and experimental methods indicates that mammalian genomes harbour a great number of npcRNAs. Among them, probably more than one thousand micro RNAs (miRNAs) alone. These short RNAs (~20-25 nucleotides, nt) play key roles in regulation of gene expression at important stages in organismal development (1). Apart from complementing our understanding of gene regulation, a lot of hope rests on miRNAs and short interfering RNAs (siRNAs) as future tools for therapeutic intervention in a number of diseases (2). Experimental and computational approaches have also significantly increased members of other classes of npcRNAs (3), such as small nucleolar RNAs (snoRNAs). At the other end of the size spectrum, thousands of transcripts (large RNAs, lRNAs) have been predicted that have characteristics of messenger RNAs (mRNAs) but lack a significant open reading frame (4). It is expected that experimental RNomics will expand the tally of other npcRNA families and detect novel individual npcRNA species. Therefore, we will continue the search for novel npcRNAs in mammals including humans. In the cell, most RNAs bind protein partners to form biologically active ribonucleoprotein particles (RNPs). Our explorative project begins to address the quest for the protein components of novel RNAs.

Results/Project Status

Experimental search for novel snmRNAs

We developed a new method for generating full-length cDNA libraries of snmRNAs. This protocol allows not only for cloning of RNA >50 nt, but also RNAs as small as miRNAs. We generated libraries from HeLa cells. After separation of nuclei and cytoplasm, we isolated total RNA from both compartments. Each preparation was subjected to preparative denaturing acrylamide gels and two size fractions (10 - 60 nt; 60 - 500 nt) were collected, resulting in four separate cDNA libraries. To evaluate the quality of the libraries, about one hundred cDNA inserts of each were sequenced. In the fraction of 10 – 60 nt, we identified numerous known miRNAs (most of them full-length) as well as candidates for novel human miRNA species. In the fraction of 60 – 500 nt we also detected candidates for novel snmRNAs, despite the relatively low number of sequences.

Prediction and validation of novel lRNAs

Prediction of genes encoding non-messenger RNAs is a difficult task with only very limited approaches existing today. In our predictive method we focus mainly on the introns of human genes (Figure 1). Among the extremely large number of public human ESTs we identify those which completely fall into introns and, in addition, show no partial sequence similarity to any known proteins nor seem to be products of so-called internal priming events. The accordance of our predictions to recently published experimental data (5) based

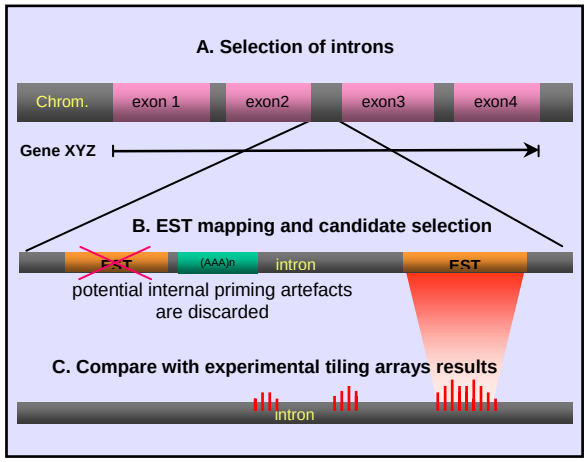

Fig 1: *In-silico selection of candidate regions.* ***(A)*** *Complete intron-exon gene structures are imported from ENSEMBL.* ***(B)*** *Introns are compared versus dbEST (human), significant matches (>90% similarity, >80% coverage) are further analysed with respect to nearby poly(A) regions. In addition, positive ESTs are compared against transcript and protein databases (any hit excluded from the candidate set).* ***(C)*** *Recently published results from Affymetrix tiling-array based transcription screens (5) are imported and used for further confirmation.*

on genomic tiling microarrays is remarkable: On average, more than 60% of the loci predicted to be transcribed show also transcription in the experiment (Table 1). We are currently in the process of experimentally validating the predictions. Even by conservative standards, the total number of unknown genes encoding npcRNAs is expected to be more than fifty thousand. Table 1 gives an overview of the predicted candidates for several human chromosomes.

chromosome	no candidates (ESTs)	Affymetrix confirmed cands.	cands./intron (average)
22	1154	705	0,39
21	780	511	0,52
6	2711	1845	0,37
X	1364	785	0,27
Y	39	27	0,11
Genome (expect.)	**50341**		

Tab 1: *Statistics of intronic candidates for novel non-messenger RNA genes. The numbers are based on analysis results for 5 human chromosomes.*

To verify this computational approach, we have selected a set of ESTs annotated from human Embryonic Stem Cells (hES), Placenta, Thyroid and Brain tissues for experimental analysis. In an initial pilot study, we have evaluated twelve candidates derived from ESTs of the human ES cell line H1. RT-PCR suggests that all except one are transcriptionally active. Furthermore, we investigated the presence of three of

these candidate npcRNAs by Quantitative RT-PCR and Real-time PCR on RNA from twenty different tissue samples. Interestingly, while one npcRNA shows similar expression levels in almost all tissues tested, the other two npcRNAs show clearly differential expression patterns. In an alternative approach, we tested twelve candidates annotated from ESTs of Human Placenta for their expression pattern by Northern blot analysis using specially designed oligonucleotides as hybridization probes. Two candidates are expressed at high levels. While the present data is very promising, reliable statistics will be only available after larger sets of data are evaluated. Therefore, as all methodological parameters are established, analysis of 100-150 more candidates annotated from hES cells, placenta and brain are in progress. Towards elucidating the function of novel npcRNAs, efforts are underway to identify protein interacting partners by Phage Display technology, and at final stages the functional analysis will be performed by over-expression and knock-down experiments on transfected-cell arrays.

Subcellular location and protein interaction of novel npcRNAs

The goal of this workpackage is the subcellular localization of nmRNAs as well as determination of the RNA-protein interactions using transfected-cell array technology. Fluorescently labelled RNA molecules will be directly transfected onto arrays and the subcellular location of the RNA will be determined in a high throughput manner. Furthermore, we will optimise two hybrid transfected-cell array technology for analysis of interactions between RNA and proteins. Using this approach not only proteins that bind directly to the RNAs will be identified but also proteins that bind to the RNA via protein-protein interactions. Mammalian vectors, that allow coexpression of novel npcRNAs candidates fused with a RNA tag-sequence and expression of a bacteriophage protein that specifically binds to this tag, have been successfully generated. The protein will be a fusion protein with domains that facilitate detection of expression as well as affinity purification of the formed RNPs and separated by specific protease sites. We also accomplished establishment of the cell array technique for intracellular protein detection (6), such that this technique can now be implemented for large scale RNA localisation studies.

The transcriptome of the MHC locus

We have started to analyse the immunologically important genomic regions of the major histocompatibility complex (MHC), the natural killer cell complex (NKC), and the leukocyte receptor complex (LRC) for the presence of npcRNA transcripts. For this purpose, we employ high-density oligonucleotide arrays (NimbleGen technology) that cover about 12 Mb of the rat genome. For the design of the chip we employed 60-mer oligodeoxynucleotides with an average spacing of 14 nucleotides. The chromosomal regions covered are: chr1, 64200000 – 69100000; chr1, 93800000 – 94200000; chr4, 165700000 – 169350000; chr20, 900000 – 5210000. As hybridising probes, we use RNA derived from unstimulated freshly prepared lymphocytes and from interferon-alpha-stimulated lymphocytes. Experiments are currently performed and data will be available shortly.

Outlook

With demonstrations of the feasibility of our multi-disciplinary and multi-facetted approach we are now in an excellent position to make the transition from discovery of novel npcRNA transcripts towards elucidation of their places and roles in the cell.

*Lit.: **1.** Zamore PD, Haley B. Ribo-gnome: the big world of small RNAs. Science. 2005;309:1519-24. **2.** Gong H, Liu CM, Liu DP, Liang CC. The role of small RNAs in human diseases: potential troublemaker and therapeutic tools. Med Res Rev. 2005;25:361-81. **3.** Huttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, Brosius J. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. EMBO J. 2001;20:2943-53. **4.** Okazaki Y, Furuno M et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature. 2002;420:563-73. **5.** Cheng J, Kapranov P et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005;308:1149-54. **6.** Hu YH, Vanhecke D, Lehrach H and Janitz M. High-throughput subcellular protein localization using cell arrays. Biochem Soc Trans. 2005, in press.*