Evolution of Vault RNAs

Peter F. Stadler [a,b,c,d], Julian J.-L. Chen [e,f], Jörg Hackermüller [b], Steve Hoffmann [a], Friedemann Horn [g,b], Phillip Khaitovich [h], Antje K. Kretzschmar [b], Axel Mosig [h,i], Sonja J. Prohaska [c,d,a], Xiaodong Qi [e], Katharina Schutt [b], Kerstin Ullmann [b]

[a] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
{steve,sonja,studla}@bioinf.uni-leipzig.de

[b] *Fraunhofer Institut für Zelltherapie und Immunologie – IZI Perlickstraße 1, D-04103 Leipzig, Germany*
{joerg.hackermueller, antje.kretzschmar, katharina.schutt, kerstin.ullmann}@izi.fraunhofer.de

[c] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[d] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[e] *Department of Chemistry and Biochemistry, Arizona State University, Tempe, AZ 85287, USA*
jlchen@asu.edu

[f] *School of Life Sciences, Arizona State University, Tempe, AZ 85287, USA*

[g] *Institute of Clinical Immunology and Transfusion Medicine, University Hospital Leipzig, D-04103 Leipzig, Germany*
friedemann.horn@medizin.uni-leipzig.de

[h] *CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, 320 Yue Yang Road, 200031 Shanghai, China*
axel.mosig@gmail.com

[i] *Max-Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*

Manuscript *31 October 2008*

Abstract

Vault RNAs (vRNAs) are small, about 100nt long, poly-III transcripts contained in the vault particles of eukaryotic cells. Presumably due to their enigmatic function they have received little attention compared to other ncRNA families. Here we report on a systematic study of this rapidly evolving class of ncRNAs in deuterostomes, providing a comprehensive collection of computationally predicted vRNA genes. Previously known vRNAs are located at a conserved genomic region linked to the protocadherin gene cluster, an association that is conserved throughout gnathostomes. Lineage specific expansions to small vRNA gene clusters are frequently observed at this locus. Expression of several paralogous vRNA genes, most but not all located at the canonical syntenically conserved locus, was verified by RT-PCR in both zebrafish and medaka. Homology search furthermore identifies an additional vRNA gene in eutheria that was misclassified as a microRNA. Lineage specific loss of one of the two loci in several eutherian lineages suggests compensation among vRNA transcripts and supports the annotation of the novel locus as functional vRNA. The comparative analysis of the promoter structure shows substantial differences between the two eutherian vRNA loci, explaining their differential expression patterns in human cancer cell lines.

Key words: vault particle, vault RNA, micro RNA, homology search, RNA secondary structure

1 Introduction

Vaults are large ribonucleoprotein particles in the cyctoplasm of many eukaryotic cells. They have been found as highly conserved ribonucleoproteins in several eukaryotes including many deuterostomes and the slime mold *Dictyostelium discoideum*, while prominent model organisms such as *Drosophila*, *Caenorhabditis*, *Arabidopsis*, and *Saccharomoyces cerevisae* seem to lack vaults [49,65]. Vault particles have received considerable attention because of their large size (about three times that of a ribosome with a weight of about 13 MDa [9]). They have a characteristic hollow barrel-like shape with an unusual symmetry [55,4,26] and a relatively simple molecular composition [27,17,62,54,7]. MVP (Major Vault Protein, also known as LRP for "Lung Resistance-related Protein"), the major constituent of vault particles, appears to be sufficient to form the ultrastructure of the vault particle, which is dynamic enough to allow the incorporation of the two other known protein components (TEP1 and VPARP [53]) after formation of the particle [48].

Despite numerous reports on vaults' expression and composition, the function of these complexes is still poorly understood. Due to their sub-cellular localization in the cytoplasm as well as their association with the nuclear membrane and nuclear pore complex [1], a role in intracellular – in particular nucleocytoplasmic – transport processes has been suggested [61]. It is interesting to note in this context that MVP is related to a bacterial toxic anion resistance protein family [57]. Since MVP is frequently found to be over-expressed in a variety of drug-resistant cancer cells, it was speculated that vaults may be involved in drug sequestration [24,50]. In fact, there is evidence that vault particles contribute to extrusion of anthracyclines from their target site, the nucleus [32,31]. However, other reports using knockout or siRNA mediated down-regulation of MVP failed to confirm such a role [44,62,23].

About 5% of the mass of vault particles consists of vault RNAs (vRNAs). These RNAs are short polymerase III transcripts with a length varying between about 80 and 150 nt. Mammalian vRNA genes share characteristic upstream elements that are reminiscent of snRNA promoters [29]. So far, their function within the vault complex remains elusive, although they have been shown to bind certain chemotherapeutic drugs [15,38]. A recent study, furthermore, reported that vRNAs — together with a handful of other transcripts — are dramatically over-expressed following an Epstein-Barr virus infection in human B cells [45].

To-date, few examples of vRNAs have been studied experimentally. They exhibit little sequence conservation beyond their internal box A and box B internal Pol III promoter elements. In the human genome, three expressed vRNAs, *hvg1-hvg3*, located in a cluster on Chr.5 [63], have long been known. A fourth

putative human vRNA was recently reported [45]. In contrast, mouse and rat have only a single vRNA gene [30,66,29]. Outside mammals, only two vRNAs from the bullfrog *Rana catasbaiana* have been sequenced [30]. In addition, the existence of vRNAs in the sea urchin *Strongylocentrotus purpuratus* was shown [55]. Its sequence, however, has not been determined.

Because of their high sequence variability, vRNAs have been annotated only in mammals and in *Xenopus* by means of `ensembl`'s ncRNA annotation pipeline. Additional mammalian vRNA sequences can easily be found with `blastn`. In contrast, `blast`-based searches beyond mammals have not been successful. In an earlier study, we therefore used vRNAs as an example to benchmark the performance of the pattern-based homology search tool `fragrep2` [42], identifying a few additional vRNA candidates in the genomes of chicken and of some teleost fishes. Here, we report on a systematic computational analysis of vault RNAs covering all deuterostomes, complemented by experimental studies of human vRNAs in a variety of human cancer cells and the experimental validation of multiple vRNAs in medaka and zebrafish.

2 Materials and Methods

2.1 Computational Methods

2.1.1 Homology Search

We used local copies of the genomes included in `ensembl` and `pre.ensembl` (version 50, summer 2008), the genomes of basal deuterostomes (lancelet, sea urchin, *Ciona savignyii*) as provided through the UCSC genome browser, as well as shotgun traces of the ongoing genome projects for *Saccoglossus* and several mammalia. In addition, we used the NCBI web interface to query the various sections of `Genbank` and the `Trace Archive`.

Mammalian genomes were queried iteratively using NCBI `blastn` (version 2.2.17) until no further candidate sequences were found. For teleosts, we started from *Danio rerio* and *Tetraodon nigroviridis* candidate sequences reported in [42]. We used both NCBI `blastn` and the `blastn` version accessible through `ensembl`'s web interface for the five teleost genomes, using the *distant homology* parameter setting for the latter. The results of these searches were then used again as queries. A second repetition did not result in further candidate sequences.

Separate sequence alignments (see below) for tetrapods and teleosts were then

used to create search patterns for `rnabob` [1], a fast implementation of `RNAmot` [13] which searches for RNA secondary structure patterns, and for `fragrep2` [42], a tool that searches fragmented approximate sequence patterns. Candidates were manually inserted into the multiple alignment and used to refine the query patterns for the two programs.

`GotohScan` [19], an implementation of the semi-global alignment variant of Gotoh's dynamic programming algorithm [16], was used as an independent method for retrieving candidates from more distant genomes, with both the experimentally known vRNAs and the candidates found in previous steps as query.

2.1.2 Alignments and Secondary Structure Analysis

Separate multiple sequence alignments for various clades were computed with `clustalw` [60] and `dialign` [40,41]. These were manually refined and then combined using `emacs` in `ralee` mode [18]. Sequence Logos [51] were derived from alignments using `aln2pattern`, a component of the `fragrep` package [42].

Initial structural annotation was produced with `RNAalifold` [20]. Combined sequence-structure alignments were computed using `locarna` (version 1.3.3) [67] based on initial base pairing probability matrices produced by `RNAfold -p`, the `Vienna RNA Package` [21] implementation of McCaskill's algorithm [39]. The `locarna` alignments were manually refined in unpaired regions.

2.1.3 Analysis of upstream regions

In order to investigate the promoter structure, we analyzed the predicted vRNA genes located at the **pcdh** locus and the eutherian vRNA-4 candidates. Genomic sequences were extracted with 500nt 5' flanking sequence since this interval contains all previously described Pol III promoter elements [63]. We also included 50nt 3' flanking sequence to identify possible indications of retrotransposition, in particular poly-A stretches. The following subsets of sequences were searched for conserved elements using `meme` [5,6]:

(1) all mammalian sequences for which linkage with **pcdh**, ZMAT2, SMAD5 or TGFB1 was established unambiguously.
(2) all teleost vRNAs at the canonical loci and those with evidence for expression in zebrafish and medaka.
(3) all vRNA candidates from amphioxus, tunicates, lamprey, shark, latimeria, frog, lizard, and chicken.

[1] `ftp://selab.janelia.org/pub/software/rnabob/`

These sets of sequences are available in the Electronic Supplement [2] . For eutheria, several minimal and maximal motif lengths and different models were explored with largely consistent results in respect to the similarities and differences between candidates from the **pcdh** and **SMAD5** locus. For comparison, conserved sequences associated with Pol III promoters were extracted from the literature on vRNAs and other Pol III transcripts [14,30,66,63,29,11].

Motifs identified by `meme` served as an input for `mast` searches [5] against vault RNA homologs with unknown genomic location (e.g. from shotgun traces or low coverage genomes).

2.2 VaultRNA Expression

2.2.1 Expression of Human vRNAs

Cell culture MCF-7, HEK-293, PC3, Du-145 and HeLa cells (ATCC) were grown in DMEM/high glucose with 10% FCS (Biochrom), 100Units/ml of penicillin and 100μg/ml of streptomycin (PAA). LNCaP cells (ATCC) were grown in RPMI1640 supplemented with 10% FCS (Biochrom), 100Units/ml of penicillin and 100μg/ml of streptomycin (PAA) and 10mM Hepes (Biochrom). RWPE-1 cells (ATTC) were grown in Keratinocyte-Serum free medium (Gibco-BRL) supplemented with 5ng/ml human recombinant EGF (Gibco-BRL) and 0.05mg/ml bovine pituary extract (Gibco-BRL). All cells were cultured at 37°C in a humified atmosphere of 5% CO_2 in air.

Real-time quantitative PCR Total RNA was extracted from the different fractions by using TRIzol reagent according to the manufacturer's instructions (Invitrogen, Carlsbad, CA). Quantitative real-time expression analysis (qRT-PCR) of vRNAs was carried out with the primers listed in the Appendix. 5μl total RNA of each fraction was reverse transcribed using random hexamer primers and the High Capacity Reverse Transcription kit (Applied Biosystems). The cDNA was diluted 1:12.5 and served as the template for qPCR analysis using the TaqMan 9700 System (Applied Biosystems) with FAST SYBR green mastermix (Applied Biosystems) and the primers listed above. All amplicons were confirmed by sequencing. For each vaultRNA assay, a standard curve was calculated to check PCR efficiency. Expression of the vaultRNAs in the different cell lines was normalized to genomic DNA.

[2] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-022/`

2.2.2 vRNAs in Short Read Sequencing Libraries

Datasets The datasets analyzed here for vaultRNA fragments were produced in the context of other projects and were (`HeLa` data [12]) or will be published in that context. In brief, total RNA was isolated from the frozen prefrontal cortex tissue using the TRIzol (Invitrogen, USA) protocol with no modifications. Low molecular weight RNA was isolated, ligated to the adapters, amplified, and sequenced following the *Small RNA Preparation Protocol* (Illumina, USA) with no modifications.

`BGI cortex`, `rep1`, and `rep2` libraries represent three technical replicates of the same three pooled samples from prefrontal cortex of humans, chimpanzees, and rhesus macaques. In each case, prior to low molecular weight RNA isolation, total RNA from 20 male human individuals aged between 14 and 58 years, 5 chimpanzees aged between 7 and 44 years, and 5 rhesus macaques aged between 4 and 10 years was combined in equal amounts. Replication was carried out by independent processing of the mixed sample of 20 individuals starting from the low molecular weight RNA isolation step. `Cerebellum` library: total RNA from 5 male human individuals aged between 20 and 56 years, 5 chimpanzees aged between 7 and 44 years, and 5 rhesus macaques aged between 4 and 10 years was combined in equal amounts. `Aging` library: 14 sequencing lanes of sample containing RNA from the prefrontal cortex of 12 humans aged from 0 to 98 years were analyzed.

Short-Read Mapping The mapping of large libraries containing hundreds of thousands of short, inaccurate sequences to large mammalian genomes cannot be performed reliably and efficiently by commonly used heuristics such as `blat` [28] or `blast` [3]. This is due to limitations in both computational resources and accuracy. We therefore used `segemehl`, a new mapping tool based on enhanced suffix arrays (ESA), which was developed by Hoffmann et al. [22]. It uses an alternative heuristics based on the matching statistics [8] to incorporate not only mismatches but also insertions and deletions. Briefly, the matching statistics computes the longest common prefix (lcp) of a query pattern P with a genomic sequence S represented by the ESA. The matching statistics can be computed very quickly, in $O(m)$, where m denotes the length of the query pattern P. However, the traditional version of the matching statistics does not allow for errors in the query pattern. To overcome this problem, we keep track of alternative paths during the traversal of the ESA. Let P_i denote a suffix of P that starts at position i. For some P_j, $j > 0$, the lcp with S is sought. The information on the lcp for P_{j-1} is used to narrow down the search space. To achieve this, the suffix link for P_{j-1} is used to find a start position within ESA [2]. During the search we keep track of all alternative matches with no more than D errors and report the highest scoring match. The human brain library was matched to the complete human genome with

$D = 1$. A match was reported if the percentage of matched pattern nucleotides was $\geq 80\%$ or the E-value was ≤ 0.05.

We also mapped all deep sequencing libraries directly against our four candidate sequences using the `Soap` program [36] allowing up to 1 mismatch position and a seed size of 8.

2.2.3 Expression of Teleost vRNAs

Genomic DNA and total RNA were isolated from fish liver tissue using DNAzol and TRIzol reagents (Invitrogen), respectively, following the manufacturer's protocols. Concentrations of DNA and RNA samples were determined by A260 measurement using the Nanodrop ND-1000 spectrophotometer (Nanodrop Technologies). Putative teleost fish vRNAs were PCR amplified from genomic DNA (0.5μg/50μl reaction) with Taq DNA polymerase (New England Biolabs) and gene specific primers at 1μM final concentration. Each primer (listed in the Appendix) was designed to anneal specifically to the respective teleost vRNA genes.

The PCR was carried out with 1 cycle at 95°C for 2 min, followed by 35 cycles of 94°C for 20s, 58°C for 20s and 72°C for 15s, and finished with a final elongation at 72°C for 2 min. The PCR products were gel purified and cloned into *pZero* vector (Invitrogen) for sequencing confirmation of the specific vRNA genes amplified.

The expression of individual vRNAs were verified by RT-PCR. From 2μg of total RNA, medaka or zebrafish, cDNA libraries were prepared using Thermoscript reverse transcriptase (Invitrogen) and a random-hexamer primer following the manufacturer's instruction. Gene specific primers were used to PCR amplify the putative vRNA sequences from the cDNA libraries under conditions similar to the PCR condition for genomic DNA samples, except 3-5 additional cycles. The RT-PCR products were cloned into *pZero* vector and sequenced. The Mock RT reactions with reverse transcriptase enzyme omitted served as the negative control.

3 Results

3.1 Homology Search

3.1.1 Mammalian vRNAs

To-date three expressed human vRNA genes (*hvg1-hvg3*), forming a small cluster on Chr.5, have been described [63]. Orthologs of the human vRNA cluster can easily be found in the chimp (*Pan troglodytes*) genome. In contrast, both orangutan (*Pongo pygmaeus*) and macaque (*Macaca mulatta*) have two copies instead of three. A sequence alignment shows clearly that human *hvg2* and *hvg3* are very recent duplicates. This "vault RNA cluster" is located between the ZMAT2 and PCDHA (protocadherin-α) genes. Since the protocadherin-α cluster has received quite a bit of attention in recent years [47,56,58,70,69], we will refer to this regions as *the **pcdh** locus* for short. In addition, a single pseudogene has been found on Chr.X [63].

Most mammals have a single vRNA copy at the **pcdh** locus, which in particular includes the experimentally determined mouse and rat vRNAs (Fig. 1). In some cases, it has expanded locally into a multicopy cluster, notably in primates (as mentioned above), pika (*Ochotona princeps*), guinea pig (*Cavia procellus*), and shrew (*Sorex araneus*). The oppossum (*Monodelphis domestica*) also exhibits a (recent) tandem duplication of the vRNA at the **pcdh** locus.

Surprisingly, a search in the dog (*Canis familiaris*) genome returned only a degraded pseudogene of the vRNA at the **pcdh** locus. Instead, we detected a single alternative sequence by `blastn`, which is located (in anti-sense direction) between TGFB1 and SMAD5. Using this sequence as query, one can recover homologs at the syntenic position throughout all eutheria (the lack of `blast` hits in the earliest-branching eutherian, the armadillo (*Dasypus novemcinctus*) may be due to the low coverage of the genome). It is interesting to note that — with the exception of some laurasiatheria — both the TGFB1-SMAD5 and the ZMAT2-PCHDA locus are found on the same chromosome, barely 0.5Mb separated from each other.

To our surprise, the corresponding `blast` hits are annotated as *mir-886* in Human [35] and Macaque [71]. Sequence alignments, however, quite clearly identify this sequence as a vRNA homolog. This was already recognized by [45]: A transcript called *CBL-3* matching our prediction was identified as a putative fourth vRNA in EBV-infected cells. We, as well as the authors of the companion paper [46] choose to investigate this transcript in more detail to confirm its identity as vault RNA (see below).

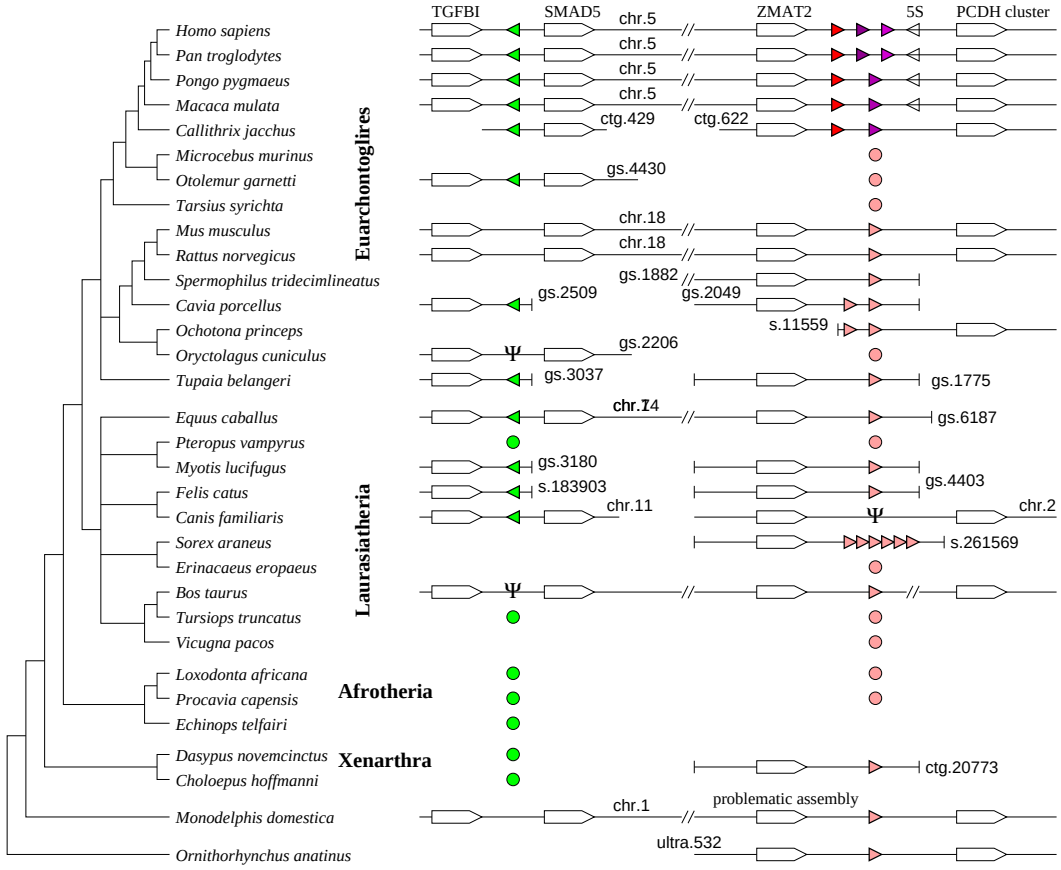


Fig. 1. Overview of the putative functional vRNA genes in Mammalia. The two syntenically conserved mammalian vRNA loci are usually located on the same chromosome. Most investigated genomes are not assembled to chromosomes; in these cases "Contig" (ctg.), "Scaffold" (s.), "GeneScaffold" (gs.) or "UltraContig" (ultra.) numbers are provided. Filled Triangles indicate vRNAs and their reading direction relative to the adjacent genes (the reading direction of the database entries is not indicated). Pseudogenes are marked by Ψ. In primates, a 5S rRNA is located adjacent to the vRNAs (open triangle). For Afrotheria and Xenarthra the presence of recognizable homologs is indicated. The current assemblies do not allow to determine whether they are located at the syntenic positions. Circles indicate vRNA homologs with promoter elements characteristic for one of the two loci (see Fig. 2 for details).
The phylogeny of the eutherian superfamilies follows [34]. The figure represents the genomic data from `ENSEMBL 50` and the NCBI Trace archive in fall 2008.

The vRNA copy at the SMAD5 locus is typically somewhat better conserved than its counterparts at the **pcdh** locus, and no duplicates at this locus have been found in any of the investigated mammalian genomes. The locus harbours a highly derived, probably pseudogenized, sequence in the rabbit (*Oryctolagus cuniculus*) and in the cow (*Bos taurus*) genomes. It has completely lost its vRNA in both mouse and rat.

Our data show that the origin of the SMAD5 locus pre-dates the divergence of Afrotheria, Laurasiatheria, and Euarchontoglires. A search in the trace archive returned two putative vRNAs from the sloth *Choloepus hoffmanni*, while only a single vRNA sequence is found in the survey genome of *Dasypus novemcinctus*. A `clustalw` alignment and Neighbor-Joining analysis tentatively places one of the two sloth sequences together with the human SMAD5, suggesting that Xenarthra also possess both vRNAs in both loci. This result is supported independently by the presence of promoter elements characteristic for one of the two loci, see Fig. 2. In contrast, no vRNA can be identified at the SMAD5 locus in both the oppossum and the platypus genome, placing the mammalian-specific duplication of vRNAs at the root of the Eutheria.

Several mouse, rat, and human vRNA pseudogenes are explicitly discussed in the literature [29,63]. In most species, a `blastn` search returns very few hits, some of which, furthermore are easily identified as pseudogenes because they match only part of a human vRNA. In the guinea pig, however, the vRNA(s) from the **pcdh** locus spawned a larger family of pseudogenes with about 100 members. The vRNA at the SMAD5 locus, on the other hand, is the origin of several dozens of pseudogenes in primates. Many of them are listed as vRNAs by the ENSEMBL genome annotations, since this annotation pipeline at present cannot distinguish reliably between functional loci and pseudogenes. Several mammalian genomes are currently sequenced to 2X coverage only, entailing significant problems for their assembly. In particular it is not possible to determine the exact number of vRNAs in these genomes. There is no evidence that vRNAs outside the **pcdh** and the SMAD5 loci are syntenically conserved, indicating that there is no selection pressure to maintain these copies. We therefore conjecture that they are mostly non-expressed pseudogenes.

Figure 2 summarizes conserved sequence elements associated with vRNAs at the two syntenically conserved eutherian loci. While the tandem copies at the **pcdh** locus have highly similar upstream regions in primates, there are substantial differences in the sequence motifs compared to the SMAD5 locus. While `meme` uncovers different sequence elements depending on parameter settings, the distinction between the **pcdh** and the SMAD5 locus is clearly visible for almost all parameter choices. The sequence motifs in Figure 2 include most of the well-known Pol III associated elements that have been described for both vRNAs and snRNAs, see e.g. [66,63,10,29,11,68]. Interestingly, the Pol III promoters at the **pcdh** have a TATA box, which is derived in rodents [66,29]. In contrast, no TATA box was found for the vRNA genes at the SMAD5 locus.

The rich promoter structure and the clear distinction between the two eutherian loci allows us to use these sequence motifs to identify functional vRNA genes also in cases where the genomic location could not be determined based on flanking genes. This includes low-coverages genomes such as those of the

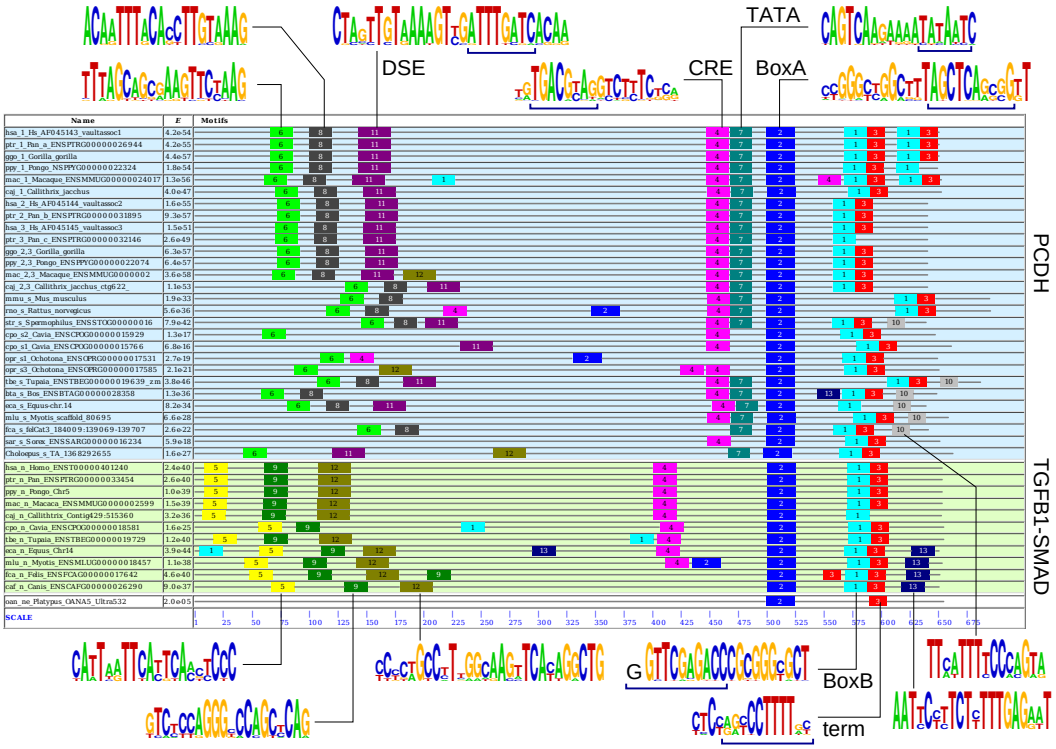


Fig. 2. Conserved elements associated with mammalian vRNAs.
The diagram combines a subset of the patterns discovered by `meme` with three different parameter settings. Previously described Pol III upstream elements (*Distal Sequence Element* DSE, *Cyclic AMP Response Element* CRE [a part of the larger *Proximal Sequence Element* PSE], and the TATA box) are indicated. The vaultRNA itself is delimited by the Box A containing motif 2 (blue) on its 5' side and by the elements 1 and 3 (cyan and red), which contain the Box B and the Pol III terminator signal, on its 3' side. Primate vRNA-1 genes have a second copy of the Box B and terminator downstream of the experimentally determined end of the transcript. The extended structure of the mouse and rat vRNAs [29] is also clearly visible.
Almost all vRNA genes at the **pcdh** locus exhibit elements 6 (green), 8 (gray), and the DSE containing motif 11 (violet), while the distal region of the novel locus is characterized by three unrelated motifs 5 (yellow), 9 (dark green), 12 (olive).

bushbaby or the elephant, and shotgun traces of genome projects in progress. Positive predictions are indicated by colored circles in Figure 1; the corresponding sequences are compiled in the supplemental material. In particular, vRNAs with SMAD5 locus promoters are found in several Xenarthra and Afrotheria, demonstrating that both types of vRNAs are present throughout all major eutherian clades.

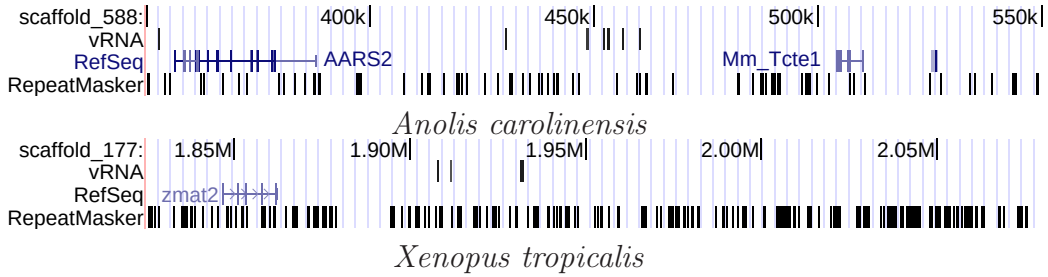


Fig. 3. vRNA locus in the lizard and frog genomes. Pictures taken from the UCSC genome browser.

3.1.2 Other Tetrapod vRNAs

Beyond mammals, genomes for only four more tetrapods, i.e. a frog (*Xenopus tropicalis*), two birds (chicken and zebrafinch), and a lizard (*Anolis carolinensis*), are available. The frog vRNAs form a cluster downstream of the ZMAT2 homolog. While only a single candidate is found in the chicken genome [42], a cluster consisting of 6 potentially expressed vRNAs is found in the lizard genome (Fig. 3). No convincing candidate was found in the genome of the zebrafinch *Taeniopygia guttata*. This may be due to the preliminary status of the genome assembly.

Although somewhat rearranged in chicken and lizard, the genomic locations of the tetrapod vRNA candidates are clearly syntenic to the mammalian **pcdh** locus. The identity of the *Xenopus* sequence, furthermore, is unambiguous because it is almost identical to the two experimentally known bullfrog vRNAs [30].

3.1.3 Latimeria and Shark

In addition to genome projects, we also profit from the interest in the protocadherin gene cluster. This locus was sequences in the coelacanth *Latimeria menadoensis* [47] and most recently also in the elephant shark *Callorhinchus milli* [69]. Searching the region upstream of the PCDH genes with `Gotohscan` resulted in a single shark vRNA and a pair of hits in the coelacanth, one of which is probably the functional vRNA and the other appears to be a truncated pseudogene.

The entire regions from HARS, DND1, and ZMAT2 to the protocadherin cluster is syntenically conserved between shark and human [69]. Our findings thus imply that the **pcdh** locus contained the vRNA already in the ancestral jawed vertebrate.

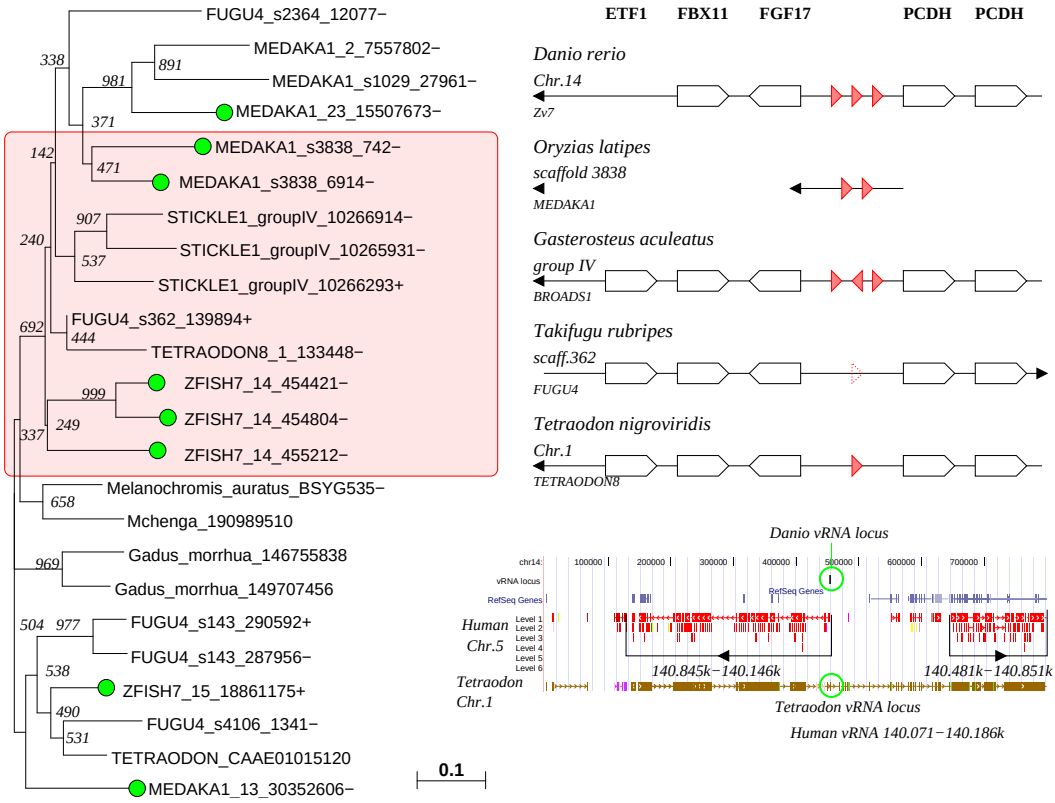


Fig. 4. Evolutions and location of vRNAs in teleosts.
L.h.s.: Neighbor-Joining tree of the teleost vRNA candidates, highlighting the sequences located at the **pcdh2** locus and their most likely paralogs in the medaka genome. Additionally, putative vRNA sequences from two cichlids (*Melanochromis auratus* and *Mchenga conophoros*, NCBI trace archive) and from the herring (*Gadus morrhua*, NCBI EST DB) are included. Sequences with experimental evidence for expression are marked with a green bullet. **R.h.s.(top):** genomic organization of the **pcdh2** locus in the five sequenced teleosts. **R.h.s.(bottom):** UCSC genome browser map of zebrafish **pcdh2** locus showing synteny with the human **pcdh** region on Chr.5 and with the **pcdh2** locus in the pufferfish *Tetraodon nigroviridis*. While largely conserved among teleosts, the **pcdh2** has been broken up and rearranged locally in teleosts relative to the ancestral state still present in the human genome.

3.1.4 Teleost vRNAs

Multiple vRNA candidates were found by a combination of `fragrep`, and `blast` for all five available teleost genomes. By their genomic location, they can be grouped into two classes. The first group, which most likely comprises functional vRNA genes is located at the **pcdh2** locus [58,70], one of the paralogous copies of the ancestral **pcdh** locus that arose through the fish-specific genome duplication (FSGD) [59,64].

Reconstruction of the ancestral karyotypes based on comparisons of teleost fish and tetrapod genomes has suggested that several major chromosomal rearrangements occurred in the fish lineage within a short period after the FSGD [25]. In particular, several local changes have modified the surrounding of the two paralogous **pcdh** loci, see also Fig. 4. The analysis of the teleost loci is complicated by the fact that the quality of the fugu genome assembly is poor in this region, showing a substantial amount of misassembly in the protocadherin loci, presumably in part due to the similarity between the many protocadherin genes [70]. The aberrant fugu candidate vRNA at this position, which lacks almost the entire variable loop region, however, was confirmed by independent sequencing of the **pcdh2** locus by [70]. It remains surprising that the fugu and tetraodon vRNA candidates are much more different than one would expect for these closely related species. This may be due to rapid evolution, as observed also within primates, or due to persisting inaccuracies in the current genome assemblies of one or both species.

The second group of candidates is more heterogeneous and does not show recognizable syntenic conservation. Some of these sequences are almost certainly pseudogenes, judging from mutations that interrupt the closing stems that otherwise are nearly perfectly conserved. In both medaka and zebrafish we verified the expression of the vRNA genes at the syntenically conserved locus as well as expression of some additional homologs, see section 3.3.3. The analysis of the upstream regions with `meme` did not identify significant sequence motifs, suggesting that the promoter elements have evolved fairly rapidly in teleosts.

Interestingly, one of the *Danio rerio* vRNA candidates located in the **pcdh2** locus was annotated as a microRNA, *mir-733*, based on the expression of short RNA resembling a mature microRNA [33]. In complete analogy to the human *mir-866*, this is probably a mis-annotation, see also section 3.3.3.

3.1.5 Basal Deuterostomes

`GotohScan` uncovered a collection of related hits in the genome of the lamprey *Petromyzon marinus*. Using these as queries for `blast` search against the same genome reveals hits at the following contigs: 36465 (3), 33930 (3), 30520 (2), 11460 (1), 21649 (1), 5840 (1) and 9902 (1). Sequence alignments suggest that the first and second group of three contigs, resp., cover the same genomic locations, so that there are only 5 distinct vRNA candidates, at least three of which form a cluster. Due to the fragmented nature of the lamprey genome assembly, we cannot determine whether some or all of the vRNA candidates are located in the vicinity of the lamprey protocadherin.

Using `RNAbob`, two candidate sequences were found in the *Ciona intestinalis*

genome, one on chromosome 3p, the other one on 9q. Only the first has a (syntenically conserved) counterpart in *Ciona savignyi*. Mutations in the putative terminal stem of the copy on chr.3p in *C. intestinalis* makes it likely that this sequence is a pseudogene.

Using `GotohScan` and a subsequent `blastn` search we found six candidates in the genome of the lancelet *Branchiostoma floridae*. The sequences are located on scaffold 41 (assembly 2.0) in two tight clusters containing three vRNAs each. Phylogenetic analysis of the six sequences suggests that the two subclusters arose by independent expansions following an ancestral duplication of the vRNA. The protocadherin homolog(s) are located on different scaffolds in the lancelet.

The existence of a sea urchin vRNA was reported previously [55], although its sequence had not been determined. Starting from the lamprey and lancelet vRNAs, we used `fragrep2` and `Gotohscan` to search the genomic DNA data of the sea urchin *Strongylocentrotus purpuratus* (assembly 2.1) and the acorn worm *Saccoglossus kowalevskii*. While `fragrep2` was not successful with our descriptor (which was based on the conserved A and B boxes and hence also matched some tRNAs), `GotohScan` returned two series of tightly clustered hits for the sea urchin, and more than a dozen shotgun traces for the acorn worm. In both cases we used `blastn` to retrieve all putative homologs from the respective genomes. Our candidates are located on scaffolds *StPu13288* and *StPu11479* of the *Strongylocentrotus purpuratus* genome [52], assembly 2.1. Nearly identical sequences were also found in the 454 data of *Strongylocentrotus franciscanus* and *Allocentrotus fragilis*[3] The two *S. purpuratus* scaffolds share large amounts of almost identical sequence and exhibit large gaps; it is not unlikely, hence, that they represent the same genomic locus. If this assumption is correct, the sea urchin vRNAs genes form a single cluster with a length of $\sim$ 20kb containing 7-10 vRNA genes and pseudogenes. The acorn worm read ***M228602676*** contains two vRNA genes or pseudogenes, suggesting that it may have a similar vRNA cluster.

The motifs identified by `meme` within the upstream regions of the vRNAs of chordates (except teleosts and mammals) are more similar among the clustered copies than between species. This supports the view that the vRNA clusters are subject to concerted evolution.

3.2 Vault RNA Secondary Structures

We constructed separate alignments for the two eutherian loci, teleosts, other gnathostomes, and basal deuterostomes. Figure 5 shows an alignment of the

[3] Available at `http://www.hgsc.bcm.tmc.edu/projects/seaurchin/`.

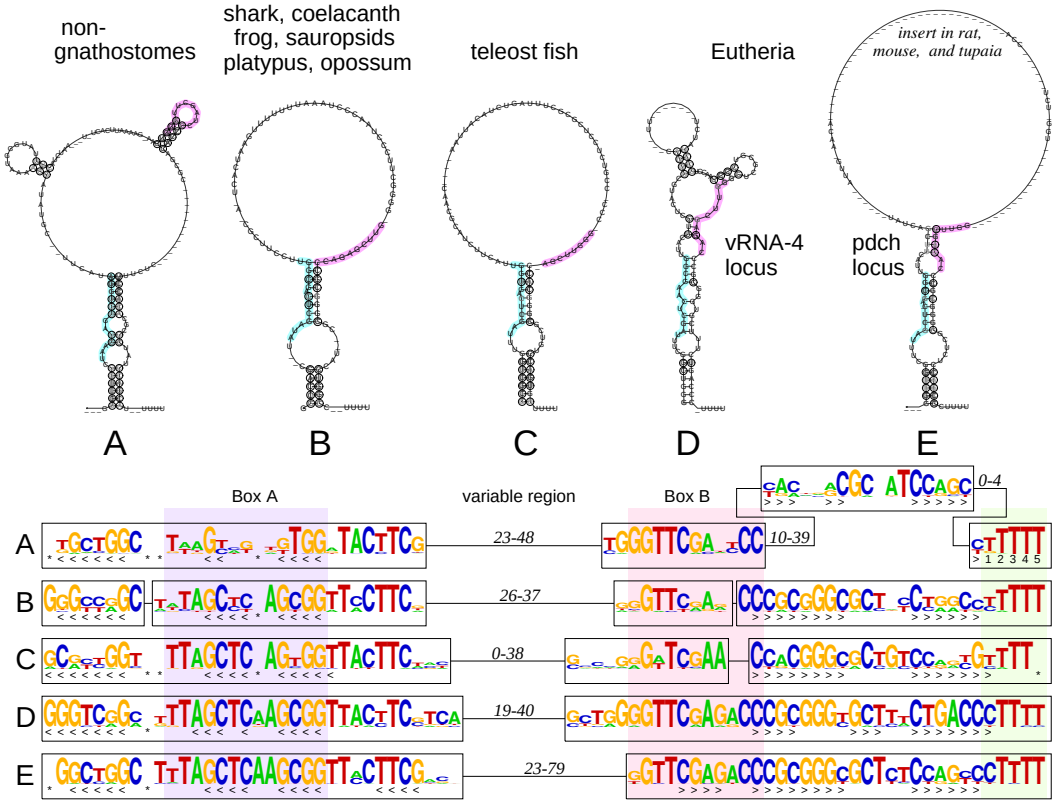


Fig. 5. Comparison of consensus secondary structures and sequence logos derived from separate alignments of of the deuterostome vRNAs. **A**: basal deuterstomes (sea urchins, acorn worm, lancelet, tunicates, and lamprey), **B**: teleosts, **C**: shark and sarcopterygii without eutheria, **D**: eutherian vRNA-4, and **E**: eutherian **pcdh** locus. The upper panel shows the consensus secondary structure computed from `locarna`-based alignments, the sequence logos were derived from the manually curated sequence-based alignments. For the latter, the base pairs of the consensus structures of the latter are indicated by <> pairs. They agree almost exactly with the `locarna`-based structures. Circles in the secondary structure drawings indicate compensatory substitutions, while letters in gray imply that one or two of the aligned sequences cannot form the corresponding base pair. Gaps in the alignment of the logos below are indicated by $*$, framed blocks are essentially gap-less in each of the sequence alignments. For gap-rich regions and the variable loop region of the vRNA, only the length ranges of intervening sequences is shown. The internal Pol III promoter elements are shaded in both panels.

sequence logos derived from these sequence alignments. The internal Pol III promoter elements, Box A and Box B [30,66,29], as well as the terminator signal are clearly visible. Interestingly, we observed the insertion of an 'A' Box A of some of the eutherian vRNAs. Furthermore, vertebrates have lost a junk of variable sequence immediately downstream of the Box B that is present in the more basal lineages.

As reported previously [42], vRNAs form a conserved panhandle-like secondary structure with a well-conserved extended stem-loop structure connecting 5'-end and 3'-end of the molecule. This structure also involves the Box A sequence. The manually generated alignments and the `locarna` alignments agree almost perfectly in this region. The Box B, on the other hand, does not take part in conserved structural features, albeit in vertebrates, the stem-loop structure overlaps the last one or two nucleotides of the Box B. In the basal lineages, Box B and the 3'-side of the stem-loop structure are separated by at least 10nt of intervening sequence, (Fig. 5). The base pairing of Box A likely contributes to the sequence conservation in the 3'-region of the vRNAs. The vRNAs of mouse and rat is peculiar in that it has a duplicated box B [29]. A similar extension is observed in the tupaia.

Both the "loop" region, and the sequence between box B and the closing stem in the basal lineages are highly variable. Even with the rather dense taxon coverage reported here, at best partial sequence alignments within subgroups can be obtained. The sequence/structure alignment algorithm `locarna` finds some partially conserved structural features within the "loop" region of the eutherian vRNAs, in particular in the novel vRNA-4. Interestingly, the regions binding chemotherapeutic drugs are located within the highly variable "loop" region [15].

3.3 Vault RNA Expression

3.3.1 Human vRNAs

The abundance of vault particles differs substantially between different human cell types [62]. Here, we have investigated the relative expression of the four vRNA genes in seven different human cancer cell lines in which vault particles are abundant. Expression levels normalized to genomic DNA for each cell line are shown in Fig. 6. We observe that the four vRNA genes are expressed in different proportions depending on the cell line.

While the expression levels of the three genes at the **pcdh** locus are similar in most cell lines, the novel vRNA-4 gene shows substantial variations. Interestingly, hvg2 expression is missing in Du-145 cells, while the expression ratio of the most recent duplicates, hvg2 and hvg3, is close to 1 otherwise.

3.3.2 Evidence from short RNA sequencing

The human microRNA atlas [35] contains small RNA sequences that uniquely map to the SMAD5 locus, hence this vRNA has been mis-annotated as microRNA *mir-886*. Fig. 7 shows deep sequencing of small RNA libraries uncover-

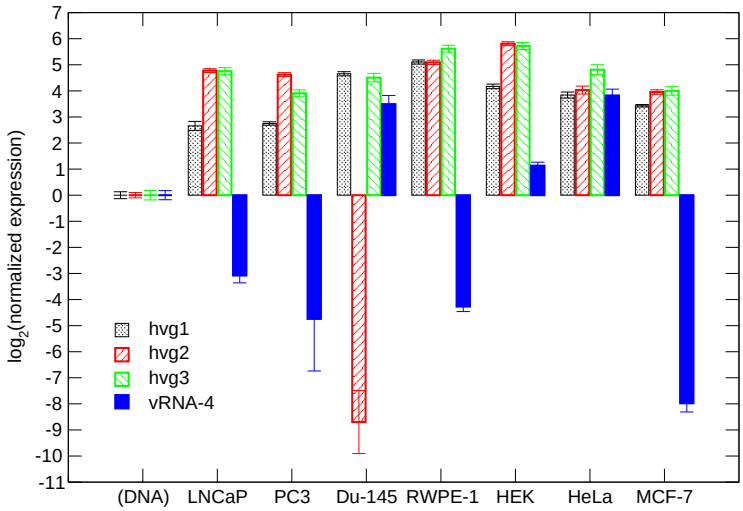


Fig. 6. Expression of the four human vRNA loci in 7 different cancer cell lines. Expression levels are normalized to a genomic DNA standard curve for each cell line.

ing multiple uniquely mapped reads at all four vRNA loci. Although Illumina's *Small RNA Protocol* is tailored towards detecting small RNAs, the data convincingly demonstrate expression from all four vaultRNA loci. In fact, we find a high degree of similarity among the transcription patterns in the human, chimpanzee and rhesus macaque data sets.

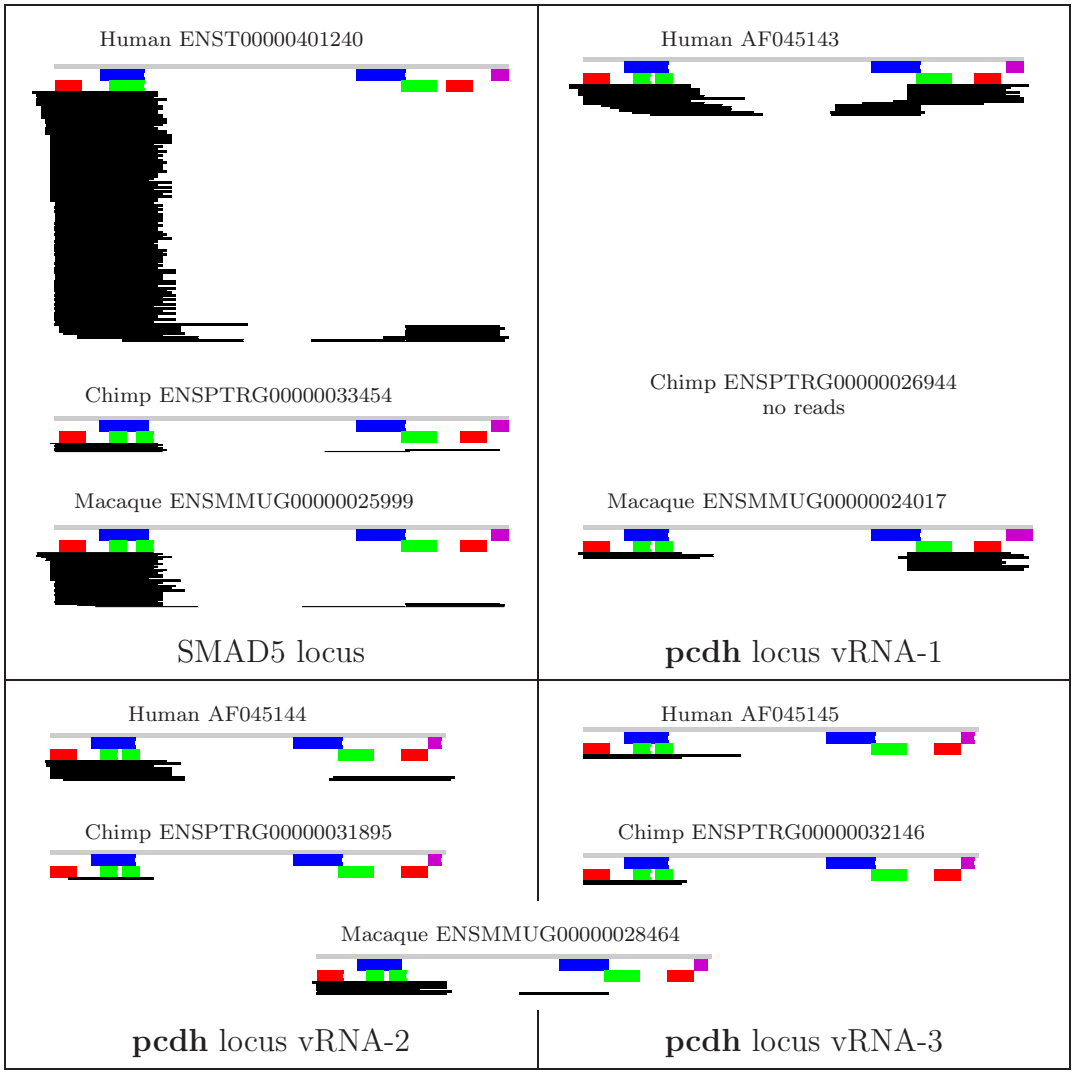
The pattern of short reads that map to the vRNA-4 locus is distinctly different from the pattern typically observed for true microRNAs, as shown in the lower panel of Figure 7: (1) The starting points of the short reads are more heterogeneous in the vRNAs. (2) The location of the reads relative to the peaks of the sequence conservation does not match the pattern expected for microRNAs. (3) The distance between the putative mature miR and miR* sequences is very large for a mammalian microRNA.

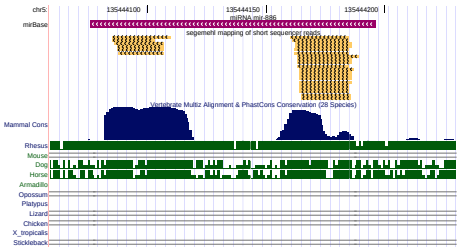
The companion paper [46], furthermore, shows that vRNA-4 co-fractionates with the major vault protein MVP in a succrose gradient, which is a strong indication that vRNA-4 indeed is a functional vRNA.

3.3.3 Expression of Teleost vRNAs

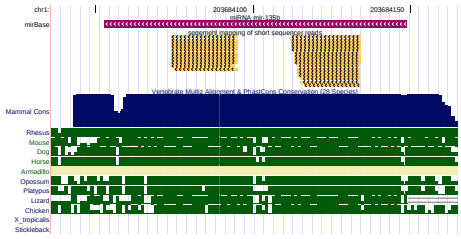
All "canonical" loci are expressed in both zebrafish and medaka, as verified by RT-PCR. Surprisingly, we also find positive RT-PCR results for the fourth zebrafish homolog and for two additional loci in the medaka.

Mapping the four expressed vRNAs and their homologs to the zebrafish genome reveals that two of these loci are annotated as microRNA (ZFISH6_14_2880749





"mir-886"



mir-135b

Fig. 7. Short read sequencing data.
Upper panel: Patterns of short reads mapping to the primate vRNAs. The grey bar indicates the vRNA sequence with annotation for Box A and Box B (in blue) and the terminator signal (in magenta). The the two major conserved stem regions in red and green, resp. Libraries are combined. Expression levels differ significantly between the four (three in Macaca) vRNA genes. The panel was prepared using the custom-made `C++` program `soap2eps`
Lower panel: Comparison of the vRNA-4 (SMAD5 locus), which was annotated as *mir-886*, and *mir-135b*, a *bona fide* microRNA. The panel was produced with the UCSC genome browser.

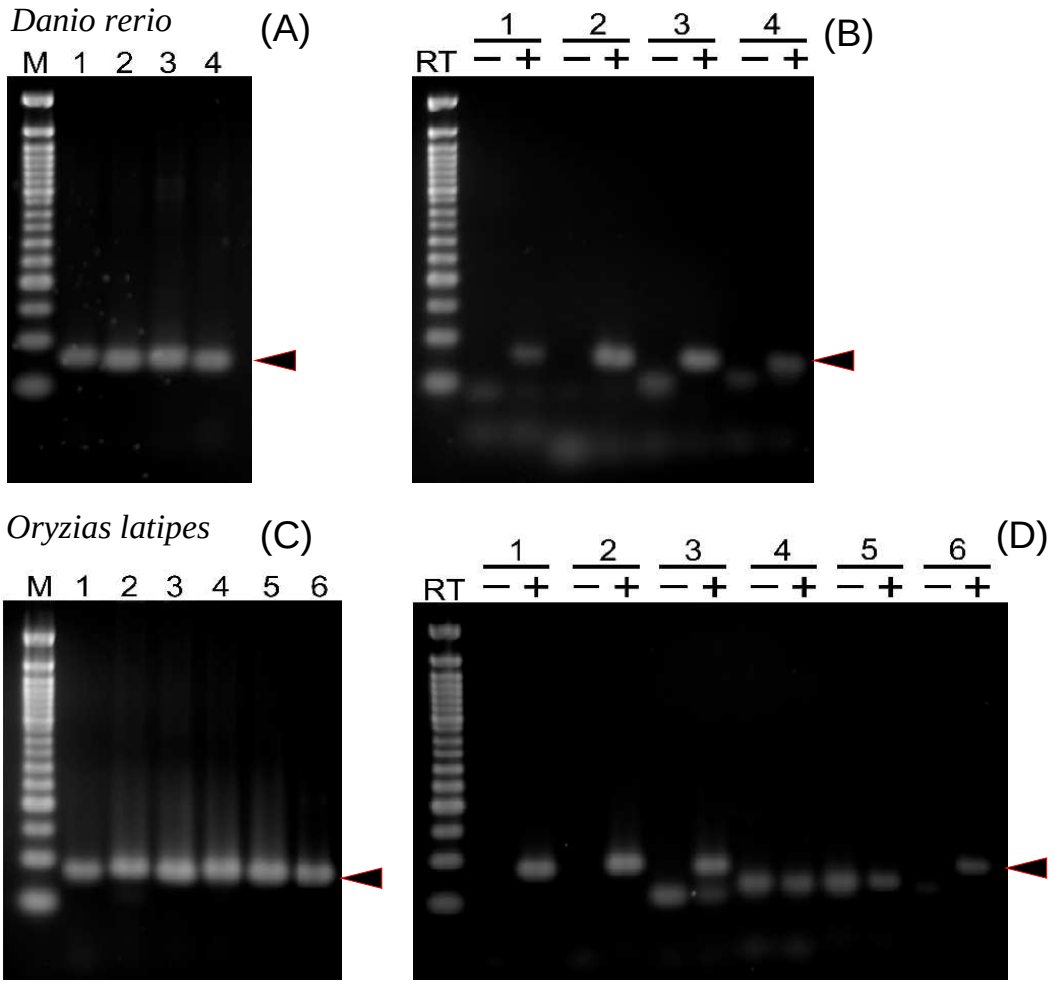


Fig. 8. Detection of teleost vRNA transcripts. Specific primer sets were designed for amplification of each putative teleost vRNA transcript. (A) PCR amplification of zebrafish vRNA genes from genomic DNA. Lane 1: *ZFISH7_14_455212-*, lane 2: *ZFISH7_14_454804-*, lane 3: *ZFISH7_14_454421-*, lane 4: *ZFISH7_15_18861175+*. (B) PCR amplification of zebrafish vRNA genes from cDNA library. (C) PCR amplification of medaka vRNA genes from genomic DNA. Lane 1: *MEDAKA1_s3838_6914-*, lane 2: *MEDAKA1_s3838_742-*, lane 3: *MEDAKA1_23_15507673-*, lane 4: *MEDAKA1_s1029_27961-*, lane 5: *MEDAKA1_2_7557802-*, lane 6: *MEDAKA1_13_30352606*. (D) PCR amplification of medaka vRNA genes from cDNA library. (RT-): PCR amplification using a cDNA library generated from a mock RT reaction in which the thermoscript reverse transcriptase was omitted.

= *dre-mir-733*) and a predicted putative microRNA (ZFISH6_14_2873070 = ENSDARG00000063786), respectively. Both loci are expressed. As in the case of the eutherian *mir-886*, this is most likely a mis-annotation. In addition to the evidence derived from comparative sequence analysis, we observe that a fragment amplified by PCR with a 3'-primer starting well outside the putative

mir-733 amplifies a vRNA-like product.

4 Discussion

We have performed a comprehensive computational analysis of vRNAs in deuterostomes. Starting from a handful of vRNA sequences from the literature, we identified more than 100 homologous genes. The expression of the predicted vRNAs was verified by PCR in human, zebrafish, and medaka.

In gnathostomes, the functional vRNA gene(s) are usually located upstream of the protocadherin cluster. Within eutheria, however, there is a second locus harboring a functional vRNA gene between the TGFB1 and SMAD5 genes. In several eutherian lineages, only one of the two loci contains a vRNA gene, strongly suggesting that the transcript from the SMAD5 is indeed a vRNA. Analysis of the external Pol III elements was used to identify likely functional vRNA genes in cases where information on genomic location was not available, e.g. for shotgun traces. The two eutherian vRNA types exhibit significantly different upstream elements, potentially explaining why the relative expression of vRNA paralogs differs substantially among human cell lines.

The evolutionary history of vRNAs is surprisingly complex and exhibits several features that sets them apart from other ncRNA families, and in particular from other Pol III transcripts. In contrast to spliceosomal snRNAs [37] and tRNAs, which behave much like mobile elements, vRNAs, like Y RNAs [43], have maintained a stable genomic location. In contrast to Y RNAs, which are organized in a single cluster comprising 4 or 5 clearly distinguishable ancient paralogs, vRNAs are either single-copy genes or form a cluster of very similar sequences. Presumably, these copies originated by tandem duplication. In most lineages they are either very young or they are subject to some form of concerted evolution. Within primates, however, there is a clear distinction of *hvg1* vs. *hvg2* and *hvg3* at the **pcdh** locus. Since we have no clear idea of the molecular function of vRNAs, we cannot even speculate what the reasons might be for these lineage-specific evolutionary patterns.

In this contribution, we have limited our attention to deuterostomes. While vault particles are evolutionary much older, they appear to be absent from the best-studied invertebrate model organisms (both nematodes and insects). A cursory `tblastn` search indeed uncovers no sign of a MVP homolog in these clades, see also [57], so that we do not expect to find vRNAs in these organisms. While MVP is present in several other invertebrates and basal eukaryotes, this part of the phylogeny appears too distant for direct homology search and at present lacks genomic data of sufficiently closely related species to identify putative vRNAs by comparative methods.

Acknowledgements

We thanks Liang Zhu (PICB) for help with `fragrep` searches and Manja Marz (Leipzig) for providing U6 snRNA promoter sequences for comparison. This work was supported in part by the Interdisciplinary Center for Clinical Research of the Medical Faculty, University of Leipzig (IZKF, projects A22 and D11), by an NSF CAREER Award (MCB0642857) to J.L.C., and by the 6th Framework Programme of the European Union, projects EMBIO (012835) to S.J.P., and SYNLET to J.H. and P.F.S.

References

[1] C. Abbondanza, V. Rossi, A. Roscigno, L. Gallo, A. Belsito, G. Piluso, N. Medici, V. Nigro, A. M. Molinari, B. Moncharmont, and G. A. Puca. Interaction of vault particles with estrogen receptor in the MCF-7 breast cancer cell. *J Cell Biol.*, 141:1301–1310, 1998.

[2] M. I. Abouelhoda, S. Kurtz, and E. Ohlebusch. Replacing suffix trees with enhanced suffix arrays. *J. Discr. Algorithms*, 2:53–86, 2004.

[3] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped `BLAST` and `PSI-BLAST`: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.

[4] D. H. Anderson, V. A. Kickhoefer, S. A. Sievers, L. H. Rome, and D. Eisenberg. Draft crystal structure of the vault shell at 9-Å resolution. *PLoS Biol.*, 5:e318, 2007.

[5] T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, Menlo Park, CA, 1994. AAAI Press.

[6] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, 34:W369–W373, 2006.

[7] W. Berger, E. Steiner, M. Grusch, L. Elbling, and M. Micksche. Vaults and the major vault protein: Novel roles in signal pathway regulation and immunity. *Cell. Mol. Life Sci.*, 2008. DOI 10.1007/s00018-008-8364-z.

[8] W. I. Chang and E. L. Lawler. Approximate string matching in sublinear expected time. In *Foundations of Computer Science, 1990*, volume 1, pages 116–124, 1990.

[9] N. E. Dickenson, D. Moore, K. A. Suprenant, and R. C. Dunn. Vault ribonucleoprotein particles and the central mass of the nuclear pore complex. *Photochem Photobiol.*, 83:686–691, 2007.

[10] A. M. Domitrovich and G. R. Kunkel. Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficencies. *Nucl. Acids Res.*, 31:2344–2352, 2003.

[11] M. Englert, M. Felis, V. Junker, and H. Beier. Novel upstream and intragenic control elements for the RNA polymerase III-dependent transcription of human 7SL RNA genes. *Biochimie*, 86:867–874, 2004.

[12] M. R. Friedländer, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky. Discovering microRNAs from deep sequencing data using `miRDeep`. *Nat Biotechnol.*, 26:407–415, 2008.

[13] D. Gautheret, F. Major, and R. Cedergren. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput. Appl. Biosci.*, 6:325–331, 1990.

[14] E. P. Geiduschek and G. P. Tocchini-Valentini. Transcription by RNA polymerase III. *Annu Rev Biochem.*, 57:873914, 1988.

[15] S. C. Gopinath, A. Matsugami, M. Katahira, and P. K. Kumar. Human vault-associated non-coding RNAs bind to mitoxantrone, a chemotherapeutic compound. *Nucleic Acids Res.*, 33:4874–4881, 2005.

[16] O. Gotoh. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162:705–708, 1982.

[17] M. M. Gottesman, T. Fojo, and S. E. Bates. Multidrug resistance in cancer: role of ATP-dependent transporters. *Nat Rev Cancer*, 2:48–58, 2002.

[18] S. Griffiths-Jones. `RALEE`—RNA alignment editor in `Emacs`. *Bioinformatics*, 21:257–259, 2005.

[19] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, and P. F. Stadler. Non-coding RNA annotation of the genome of *Trichoplax adherens*. 2008. Manuscript in preparation.

[20] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[21] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[22] S. Hoffmann, C. Otto, J. Hackermüller, S. Kurtz, and P. F. Stadler. Short read mapping with enhanced in/del detection. 2008. in preparation.

[23] K. E. Huffman and D. R. Corey. Major vault protein does not play a role in chemoresistance or drug localization in a non-small cell lung cancer cell line. *Biochemistry*, 44:2253–2261, 2005.

[24] M. A. Izquierdo, G. L. Scheffer, M. J. Flens, A. B. Schroeijers, P. van der Valk, and R. J. Scheper. Major vault protein LRP-related multidrug resistance. *Eur J Cancer*, 32A:979–984, 1996.

[25] M. Kasahara, K. Naruse, S. Sasaki, Y. Nakatani, W. Qu, B. Ahsan, T. Yamada, Y. Nagayasu, K. Doi, Y. Kasai, T. Jindo, D. Kobayashi, A. Shimada, A. Toyoda, Y. Kuroki, A. Fujiyama, T. Sasaki, A. Shimizu, S. Asakawa, N. Shimizu, S. Hashimoto, J. Yang, Y. Lee, K. Matsushima, S. Sugano, M. Sakaizumi, T. Narita, K. Ohishi, S. Haga, F. Ohta, H. Nomoto, K. Nogata, T. Morishita, T. Endo, T. Shin-I, H. Takeda, S. Morishita, and Y. Kohara. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447:714–719, 2007.

[26] K. Kato, H. Tanaka, T. Sumizawa, M. Yoshimura, E. Yamashita, K. Iwasaki, and T. Tsukihara. A vault ribonucleoprotein particle exhibiting 39-fold dihedral symmetry. *Acta Crystallogr D Biol Crystallogr*, 64:525–531, 2008.

[27] N. L. Kedersha, M. C. Miquel, D. Bittner, and L. H. Rome. Vaults. II. Ribonucleoprotein structures are highly conserved among higher and lower eukaryotes. *J. Cell Biol.*, 110:895–901, 1990.

[28] W. J. Kent. `blat`—the `blast`-like alignment tool. *Genome Res.*, 12:656–664, 2002.

[29] V. A. Kickhoefer, N. Emre, A. G. Stephen, M. J. Poderycki, and L. H. Rome. Identification of conserved vault RNA expression elements and a non-expressed mouse vault RNA gene. *Gene*, 309:65–70, 2003.

[30] V. A. Kickhoefer, R. P. Searles, N. L. Kedersha, M. E. Garber, D. L. Johnson, and L. H. Rome. Vault ribonucleoprotein particles from rat and bullfrog contain a related small RNA that is transcribed by RNA polymerase III. *J Biol Chem*, 268:7868–7873, 1993.

[31] M. Kitazono, H. Okumura, R. Ikeda, T. Sumizawa, T. Furukawa, S. Nagayama, K. Seto, T. Aikou, and A. S. Reversal of LRP-associated drug resistance in colon carcinoma SW-620 cells. *Int J Cancer*, 91:126–131, 2001.

[32] M. Kitazono, T. Sumizawa, Y. Takebayashi, Z. S. Chen, T. Furukawa, S. Nagayama, A. Tani, S. Takao, T. Aikou, and S. Akiyama. Multidrug resistance and the lung resistance-related protein in human colon carcinoma SW-620 cells. *J Natl Cancer Inst*, 91:1647–1653, 1999.

[33] W. P. Kloosterman, F. A. Steiner, E. Berezikov, E. de Bruijn, J. van de Belt, M. Verheul, E. Cuppen, and R. H. Plasterk. Cloning and expression of new microRNAs from zebrafish. *Nucleic Acids Res.*, 34:2558–2569, 2006.

[34] J. O. Kriegs, G. Churakov, M. Kiefmann, U. Jordan, J. Brosius, and J. Schmitz. Retroposed elements as archives for the evolutionary history of placental mammals. *PLoS Biol*, 4:e91, 2006.

[35] P. Landgraf, M. Rusu, R. Sheridan, A. Sewer, N. Iovino, A. Aravin, S. Pfeffer, A. Rice, A. O. Kamphorst, M. Landthaler, C. Lin, N. D. Socci, L. Hermida, V. Fulci, S. Chiaretti, R. Foa, J. Schliwka, U. Fuchs, A. Novosel, R. U. Muller, B. Schermer, U. Bissels, J. Inman, Q. Phan, M. Chien, D. Weir, R. Choksi, G. De Vita, D. Frezzetti, H. I. Trompeter, V. Hornung, G. Teng, G. Hartmann,

M. Palkovits, R. Di Lauro, P. Wernet, G. Macino, C. E. Rogler, J. W. Nagle, J. Ju, F. N. Papavasiliou, T. Benzing, P. Lichter, W. Tam, M. J. Brownstein, A. Bosio, A. Borkhardt, J. J. Russo, C. Sander, M. Zavolan, and T. T. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129:1401–1414, 2007.

[36] R. Li, Y. Li, K. Kristiansen, and J. Wang. `SOAP`: short oligonucleotide alignment program. *Bioinformatics*, 24:713–714, 2008.

[37] M. Marz, T. Kirsten, and P. F. Stadler. Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, 2008. in press.

[38] T. Mashima, M. Kudo, Y. Takada, A. Matsugami, S. C. Gopinath, P. K. Kumar, and M. Katahira. Interactions between antitumor drugs and vault RNA. *Nucleic Acids Symp Ser (Oxf)*, 52:217–218, 2008.

[39] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[40] B. Morgenstern. `DIALIGN 2`: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.

[41] B. Morgenstern, S. J. Prohaska, D. Pohler, and P. F. Stadler. Multiple sequence alignment with user-defined anchor points. *Algo. Mol. Biol.*, 1:6, 2006.

[42] A. Mosig, J. L. Chen, and P. F. Stadler. Homology search with fragmented nucleic acid sequence patterns. In R. Giancarlo and S. Hannenhalli, editors, *WABI 2007*, volume 4645 of *Lecture Notes in Computer Science*, pages 335–345, Berlin, Heidelberg, 2007. Springer Verlag.

[43] A. Mosig, M. Guofeng, B. M. R. Stadler, and P. F. Stadler. Evolution of the vertebrate Y RNA cluster. *Th. Biosci.*, 126:9–14, 2007.

[44] M. H. Mossink, A. van Zon, E. Fränzel-Luiten, M. Schoester, V. A. Kickhoefer, G. L. Scheffer, R. J. Scheper, P. Sonneveld, and E. A. Wiemer. Disruption of the murine major vault protein (MVP/LRP) gene does not induce hypersensitivity to cytostatics. *Cancer Res.*, 62:7298–7304, 2002.

[45] J. Mrázek, S. B. Kreutmayer, F. A. Grässer, N. Polacek, and A. Hüttenhofer. Subtractive hybridization identifies novel differentially expressed ncRNA species in ebv-infected human B cells. *Nucleic Acids Res.*, 35:e73, 2007.

[46] C. Nandy, J. Mrázek, H. Stoiber, F. A. Grässer, A. Hüttenhofer, and N. Polacek. Epstein-Barr virus-induced expression of a novel human vault RNA. *J. Mol. Biol.*, 2008. submitted back-to-back with this manuscript.

[47] J. P. Noonan, J. Grimwood, J. Danke, J. Schmutz, M. Dickson, C. T. Amemiya, and R. M. Myers. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res.*, 14:2397–2405, 2004.

[48] M. J. Poderycki, V. A. Kickhoefer, C. S. Kaddis, S. Raval-Fernandes, E. Johansson, J. I. Zink, J. A. Loo, and L. H. Rome. The vault exterior shell is a dynamic structure that allows incorporation of vault-associated proteins into its interior. *Biochemistry*, 45:12184–12193, 2006.

[49] L. Rome, N. Kedersha, and D. Chugani. Unlocking vaults: organelles in search of a function. *Trends Cell Biol.*, 1:47–50, 1991.

[50] G. L. Scheffer, A. B. Schroeijers, M. A. Izquierdo, E. A. Wiemer, and R. J. Scheper. Lung resistance-related protein/major vault protein and vaults in multidrug-resistant cancer. *Curr Opin Oncol.*, 12:550–556, 2000.

[51] T. D. Schneider and R. M. Stephens. Sequence Logos: A new way to display consensus sequences. *Nucleic Acids Res.*, 18:6097–6100, 1990.

[52] Sea Urchin Genome Sequencing Consortium. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science*, 314:941–952, 2006.

[53] S. Smith. The world according to PARP. *Trends Biochem. Sci.*, 26:174–179, 2001.

[54] E. Steiner, K. Holzmann, L. Elbling, M. Micksche, and W. Berger. Cellular functions of vaults and their involvement in multidrug resistance. *Curr. Drug Targets*, 7:923–934, 2006.

[55] P. L. Stewart, M. Makabi, J. Lang, C. Dickey-Sims, A. J. Robertson, J. A. Coffman, and K. A. Suprenant. Sea urchin vault structure, composition, and differential localization during development. *BMC Dev. Biol.*, 5:3, 2005.

[56] H. Sugino, H. Yanase, S. Hamada, K. Kurokawa, S. Asakawa, N. Shimizu, and T. Yagi. Distinct genomic sequence of the CNR/Pcdhalpha genes in chicken. *Biochem Biophys Res Commun.*, 316:437–445, 2004.

[57] K. A. Suprenant, N. Bloom, J. Fang, and G. Lushington. The major vault protein is related to the toxic anion resistance protein (TelA) family. *J. Exp. Biol.*, 210:946–955, 2007.

[58] M. N. Tada, K. Senzaki, Y. Tai, H. Morishita, Y. Z. Tanaka, Y. Murata, Y. Ishii, S. Asakawa, N. Shimizu, H. Sugino, and T. Yagi. Genomic organization and transcripts of the zebrafish protocadherin genes. *Gene*, 340:197–211, 2004.

[59] J. Taylor, I. Braasch, T. Frickey, A. Meyer, and Y. Van De Peer. Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.*, 13:382–390, 2003.

[60] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[61] A. van Zon, M. H. Mossink, A. B. Houtsmuller, M. Schoester, G. L. Scheffer, R. J. Scheper, P. Sonneveld, and E. A. Wiemer. Vault mobility depends in part on microtubules and vaults can be recruited to the nuclear envelope. *Exp Cell Res.*, 312:245–255, 2006.

[62] A. van Zon, M. H. Mossink, R. J. Scheper, P. Sonneveld, and E. A. C. Wiemer. The vault complex. *Cell. Mol. Life Sci.*, 60:1828–1837, 2003.

[63] A. van Zon, M. H. Mossink, M. Schoester, G. L. Scheffer, R. J. Scheper, P. Sonneveld, and E. A. C. Wiemer. Multiple human vault RNAs. *J. Biol. Chem.*, 276:37715–37721, 2001.

[64] K. Vandepoele, W. De Vos, J. S. Taylor, A. Meyer, and Y. Van de Peer. Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. USA*, 101:1638–1643, 2004.

[65] S. K. Vasu and L. H. Rome. *Dictyostelium* vaults: Disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. *J. Biol. Chem.*, 270:16588–16594, 1995.

[66] A. Vilalta, V. A. Kickhoefer, L. H. Rome, and D. L. Johnson. The rat vault RNA gene contains a unique RNA polymerase III promoter composed of both external and internal elements that function synergistically. *J Biol Chem.*, 269:29752–29759, 1994.

[67] S. Will, K. Missal, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comp. Biol.*, 3:e65, 2007.

[68] T. G. Wise, D. J. Schafer, L. S. Lambeth, S. G. Tyack, M. P. Bruce, R. J. Moore, and T. J. Doran. Characterization and comparison of chicken U6 promoters for the expression of short hairpin RNAs. *Animal Biotechnology*, 18:153–162, 2007.

[69] W.-P. Yu, V. Rajasegaran, K. Yew, W.-l. Loh, B.-H. Tay, C. T. Amemiya, S. Brenner, and B. Venkatesh. Elephant shark sequence reveals unique insights into the evolutionary history of vertebrate genes: A comparative analysis of the protocadherin cluster. *Proc. Natl. Acad. Sci.*, 105:38193824, 2008.

[70] W.-P. Yu, K. Yew, V. Rajasegaran, and V. Byrappa. Sequencing and comparative analysis of fugu protocadherin clusters reveal diversity of protocadherin genes among teleosts. *BMC Evol. Biol.*, 7:49, 2007.

[71] J. Yue, Y. Sheng, and K. E. Orwig. Identification of novel homologous microRNA genes in the rhesus macaque genome. *BMC Genomics*, 9:8, 2008.

Appendix: Primer Sequences

	forward	reverse
	Real time qPCR in human cells	
vRNA1	5'-TTCGACAGTTCTTTAATTGAAACAAGC-3'	5'-GGGCGCTCTCCAGTCCTT-3'
vRNA2	5'-AGCTCAGCGGTTACTTCGAGTAC-3'	*vRNA1*
vRNA3	5'-TTAGCTCAGCGGTTACTTCGC-3'	*vRNA1*
vRNA4	5'-TTAGCTCAAGCGGTTACCTCCT-3'	5'-GTTCGAGACCCGCGGG-3'
	Zebrafish RT PCR	
ZFISH7_14_455212-	5'-GGTTAGCTCAGTGGTTACTTCTCA-3'	5'-CCGTGGTTCGATCTCTGGCTT-3'
ZFISH7_14_454804-	5'-TAGCTCAGTGGTTACTTCTGTGAC-3'	5'-CCGTGGTTCGATCTCTGGCCG-3'
ZFISH7_14_454421-	5'-TAGCTCAGTGGTTACTTCTGTGAT-3'	5'-CCGTGGTTCGATCTCTGGCAC-3'
ZFISH7_15_18861175+	5'-GGTTAGCTCAGTGGTTACTTCCGAA-3'	5'-GGTGCCCGTGGTTCGATCCCATGA-3'
	Medaka RT PCR	
MEDAKA1_s3838_6914-	5'-CGGTTTAGCTCAGTGGTTACTTCCAA-3'	5'-CATCGGACAGCGCCCGTGGTCGAACTTGA-3'
MEDAKA1_s3838_742-	5'-CGGTTTAGCTCAGCGGTTACTTCATCGT-3'	5'-CATCGGACGGCACCCGCGGTCGAACCCGTC-3'
MEDAKA1_23_15507673-	5'-CGGTTTAGCTCAGAGGTTATTTCTACG-3'	5'-CGAACGACAGCACCGTGGTTGAACTCGG-3'
MEDAKA1_s1029_27961-	5'-CGGACACAGGGTCAGTGGTTGTTGCTAG-3'	5'-CATCGGACAGCGCCCACGGTCAAACTTGC-3'
MEDAKA1_2_7557802-	5'-TGGTTTGGCTCAGTGGTTTTTTCTACCGG-3'	5'-CATCGGGCAGTCCCTGTAGTCAAACTGTT-3'
MEDAKA1_13_30352606-	5'-TGGTTTAGCTCAGCGGTTACTTCAACTC-3'	5'-CACTGAACAGCGGCCGTGGTTTCAAACCCA-3'