Evolution of proteins and non-coding RNA genes studied with comparative genomics

Marcela Davila

Department of Medical Biochemistry and Cell Biology
Institute of Biomedicine
Sahlgrenska Academy
University of Gothenburg

2011

A doctoral thesis at a university in Sweden is produced either as a monograph or as a collection of papers. In the latter case, the introductory part constitutes the formal thesis, which summarizes the accompanying papers. These have either already been published or are in the form of manuscripts at various stages (in press, submitted or accepted).

Department of Medical Biochemistry and Cell Biology
Institute of Biomedicine, Sahlgrenska Academy
at University of Gothenburg, Box 440, SE-405 30 Göteborg.

COVER: Phylogenetic distribution of U1snRNA

ISBN 978-91-628-8183-2
URL: http://hdl.handle.net/2077/23816

Printed by Intellecta Infolog AB

Abstract

The identification of protein and non-coding RNA (ncRNA) genes is one important step in the analysis of a genome. This thesis focuses on the identification and analysis of proteins and ncRNAs homologues by exploiting a variety of computational methods in order to reach conclusions as to their structure, function, evolution and regulation. This work is composed of two different parts. One deals with computational prediction of protein and ncRNA homologues from different ribonucleoprotein (RNP) complexes and the other addresses problems related to non-random gene order in eukaryotes.

In the first part RNPs that were previously not well explored with respect to their phylogenetic distribution were examined. Thus, homology-based methods were employed to analyze the RNP complexes of RNase P, RNase MRP and the spliceosome as well as RNPs and RNA structures involved in the 3' end processing of histone mRNAs. We identified a large number of previously unrecognized homologues that improved our understanding of the evolution of the different RNPs. For example, homology relationships of the RNases P and MRP proteins were identified providing further evidence of homology between the human and the yeast RNPs. We presented evidence that the histone 3’ end processing machinery is more ancient than previously anticipated and can be traced to the root of the eukaryotic phylogenetic tree. We presented a detailed map of the distribution of the spliceosomal U12-type RNA genes, supporting an early origin of the minor spliceosome and pointing to a number of occasions where it was lost during evolution.

In the second part we generated gene order maps to show the localization of both protein and ncRNA genes in a wide range of eukaryotic organisms. Non-random gene order was then examined to identify the most important determinants of gene order conservation. One important conclusion was that gene pairs that are evolutionarily conserved and that are divergently transcribed are much more likely to be related by function as compared to poorly conserved gene pairs. The genes of such pairs are likely to be related also in terms of transcriptional control. Moreover, we presented the eukaryotic Gene Order Browser (eGOB), where data related to this project is available and where researchers can visualize and compare the evolution of gene organization in different organisms. In addition, the browser may be used to identify pairs of adjacent genes that are evolutionarily conserved and likely to be transcriptionally linked. eGOB is available at http://egob.bioimedicine.gu.se.

Keywords: bioinformatics, secondary structure, homologue prediction, evolution, non-coding RNA, RNAse P, RNase MRP, histone, U7 RNA, snRNA, gene order, bidirectional promoter.

List of publications

I Inventory and analysis of the protein subunits of the ribonucleases P and MRP provides further evidence of homology between the yeast and human enzymes
Rosenblad, M.A., **Lopez, M.D.**, Piccinelli, P. and Samuelsson, T.
Nucleic Acids Res. 2006 34, 5145-5156.

II Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components
Davila Lopez, M., Alm Rosenblad, M. and Samuelsson, T.
Nucleic Acids Res. 2008 36, 3001-3010.

III Early evolution of histone mRNA 3' end processing
Davila Lopez, M. and Samuelsson, T.
RNA. 2008 14, 1-10.

IV Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes
Davila Lopez, M. and Samuelsson, T.
PLoS One. 2010 5(5):e10654.

V eGOB: Eukaryotic Gene Order Browser
Davila Lopez, M. and Samuelsson, T.
Submitted for publication

Reprints were made with permission from the publisher

Table of contents

Introduction

A major challenge in molecular biology is to understand the information contents of a genome. The identification of protein and non-coding RNA (ncRNA) genes is one important step in the analysis of a genome. We exploited here a variety of computational methods for identifying and analyzing genes. We took advantage of the vast number of sequenced genomes as well as predicted protein sequences. In this thesis, I will describe the studies that were carried out to examine relationships of homology among selected proteins and ncRNAs in order to reach conclusions as to their structure, function and evolution.

Proteins vs. ncRNAs: a focus on computational identification

There are two general categories of genes in the human genome. First, there are protein-coding genes. In human, protein-coding regions represent less than 1.5% of the total genome sequence. Secondly, there are transcripts that do not code for protein named non-coding RNAs. Recent studies indicate that a large portion of the human genome is transcribed so as to produce ncRNAs[1].

Gene prediction

A crucial task in genomics is the identification of protein and ncRNA genes in order to understand the genome of a species. One approach makes use of statistical properties that indicate the presence of a gene[2]. These methods are referred as *ab initio* or *de novo* methods. A second approach relies on information of previously existing protein or mRNA sequences (Table 1).

A typical eukaryotic protein-coding gene has regulatory regions that determine what portions of the DNA will be transcribed, spliced and translated into a protein. These regulatory regions tend to have consensus sequences that are valuable in methods used for computational gene identification[2]. GenScan[3] is the most widely used *ab initio* prediction program.

Table 1. Examples of protein and ncRNA gene identification methods.

	Non-homology methods	Homology based methods
Protein	• GenScan	• sequence similarity-based (BLAST, FASTA) • profile-based (HMMER, PSI-BLAST)
RNA	• EvoFold • QRNA • RNAz	• sequence similarity-based (BLAST, FASTA) • patterns/motifs (RNABOB) • stochastic context free grammars (CMSEARCH) • custom designed (miRseeker, SRPscan, tRNAscan-SE)

Compared to protein-coding genes, ncRNA genes are in general less conserved in primary sequence. Furthermore, they do not have the statistical signals associated with protein genes, such as splicing signals and nucleotide composition bias in their coding regions. For these reasons, ncRNA prediction is more difficult than protein prediction. However, it is relatively common to find examples where homologous ncRNAs maintain a consensus secondary structure through compensatory base mutations. Unfortunately, ncRNA genes come in more than one flavor[4-5] (Table 2) making it difficult to obtain a single method to reliably identify ncRNAs in a genome-wide fashion.

For the problem of finding all possible ncRNAs with conserved secondary structure, including ncRNAs that have not previously been described, comparative approaches to evaluate evolutionarily conserved secondary structure seem to be a suitable choice. Examples are QRNA[6], RNAz[7] and EvoFold[8] that provide a measure of probability that a given alignment of sequences adopts a conserved RNA fold.

Table 2. Examples of some ncRNAs and their functions.

Process	Non-coding RNA	Function
Genomic stability	Telomerase RNA	telomere synthesis
RNA processing and modification	snRNA	splicing and other functions
	U7 snRNA	histone pre-mRNA 3' end formation
	RNase P	tRNA maturation
	RNase MRP	rRNA maturation
	SmY mRNA	trans-splicing
	snoRNA	nucleotide modification of RNAs
	gRNA	nucleotide modification of RNAs
	Y RNA	RNA processing, DNA replication
Regulation of gene expression	miRNA	gene regulation
	piRNA	transposon defense
	siRNA	gene regulation
	tasiRNA	gene regulation
	rasiRNA	transposon defense
Transcription	7SK RNA	negative regulation of P-TEFb
	6S RNA	bacterial transcriptional regulator
Translation	rRNA	translation
	tmRNA	rescues stalled ribosome
	tRNA	translation
Protein trafficking	SRP RNA	membrane integration

In more detail, QRNA detects conserved structural RNAs by analyzing a pairwise alignment using three models of sequence evolution (protein coding, structured RNA and a null hypothesis) and reporting the highest scoring model[6]. RNAs that do not have well conserved secondary structures will be missed by this approach. RNAz detects ncRNAs and cis-acting RNA elements in a small number of aligned sequences. It calculates the probability that a multiple sequence alignment represents a conserved structured RNA by predicting the thermodynamic stability of a consensus secondary structure[7]. EvoFold identifies functional RNAs in multiple sequence alignments. It uses a combined probabilistic model of RNA secondary structure and sequence evolution (phylo-SCFG) to evaluate how well the substitution pattern in the alignment matches its secondary structure annotation[8]. A problem with all these *de novo* methods is that they exhibit fairly high false discovery rates (50%-70%)[9].

Homologue identification

Another common task in genomics is to identify homologous sequences. In order to identify homologous proteins commonly used tools are BLAST[10] and FASTA[11], where a query sequence is compared to database sequences, resulting in a list of sequences that best match the query sequence. However, in cases where the sequences of homologous proteins have diverged significantly, more sensitive tools are applied. Thus, profile-based searches as implemented in PSI-BLAST[12] and HMMER[13] are more accurate and reliable, since they make use of information in a multiple sequence alignment instead of a single sequence. Therefore, evolutionarily related sequences, orthologues and paralogues, are more easily identified using profile-based searches.

Whereas *de novo* ncRNA gene prediction is a difficult problem, the problem of identifying RNA genes based on homology is less troublesome. In the simplest case, sequence similarity searches are enough to identify homologues. Examples are fairly conserved RNA classes such as ribosomal RNAs or whenever we are considering RNA sequences from closely related species.

However, when this is not the case, secondary structure and more complex models of RNA sequence need to be incorporated. One approach makes use of patterns or motifs, such as in RNABOB, PatSearch[14] and RNAMotif[15], where primary sequence as well as secondary and tertiary motifs are combined. Another approach involves statistical models called profile-stochastic context free grammars (profile SCFG), also known as covariance models (CMs)[16]. Here an RNA multiple alignment including a consensus secondary structure is statistically represented in a model, which then can be used to analyze a sequence or a whole genome. The INFERNAL package[17] contains methods to generate and employ covariance models.

This approach is extremely computationally intensive and therefore filtering steps are commonly applied to reduce the amount of sequence information that needs to be analyzed.

Moreover, there are algorithms that target a particular RNA class to search for homologues such as miRseeker[18] for microRNAs, tRNAscan-SE[19] for tRNAs, snoScan[20] for box C/D snorNAS and SRPscan[21] for SRP RNAs, among others.

A word on gene expression

Expression from a genome is clearly much more complicated than 'DNA makes RNA makes protein'. Eukaryotic gene expression is a complex and extensively coupled process network carried out by distinct machineries that interface physically and functionally with each other[22-27]. In general it comprises: synthesis of RNA (transcription), processing of RNA, RNA degradation, protein synthesis (translation) as well as protein folding, processing and degradation. "Processing of RNA" refers to all reactions the primary transcript is subject to in order to yield the corresponding mature functional RNAs[28], such as end modifications, splicing, cleavage processes, and chemical modifications.

ncRNAs are involved in every step of gene expression (Table 2). In this first section we focused on a number of RNAs involved in the processing of RNA as well as the proteins that are known to associate with them in ribonucleoprotein complexes (RNPs) to examine their structure, function and evolution. We employed homology-based methods (Table 1) to analyze RNPs of RNase P, RNase MRP, U7 snRNP and the spliceosome as well as an RNA regulatory element, the stem-loop structure of the replication-dependent histone mRNAs.

Ribonucleases P and MRP

The RNase P complex is a ribonucleoprotein that processes tRNA precursors[29] and acts as a transcription factor for Pol III[30]. It is found in all living cells in all three domains of life as well as in mitochondria and chloroplasts[29,31]. RNase P is structurally and evolutionarily related to the RNase MRP[32-33] that processes ribosomal RNA precursors at the A3 site[34-35]. However MRP is found only in eukaryotes. Mutations in MRP RNA lead to a variety of inherited diseases[36] such as cartilage-hair hypoplasia[37].

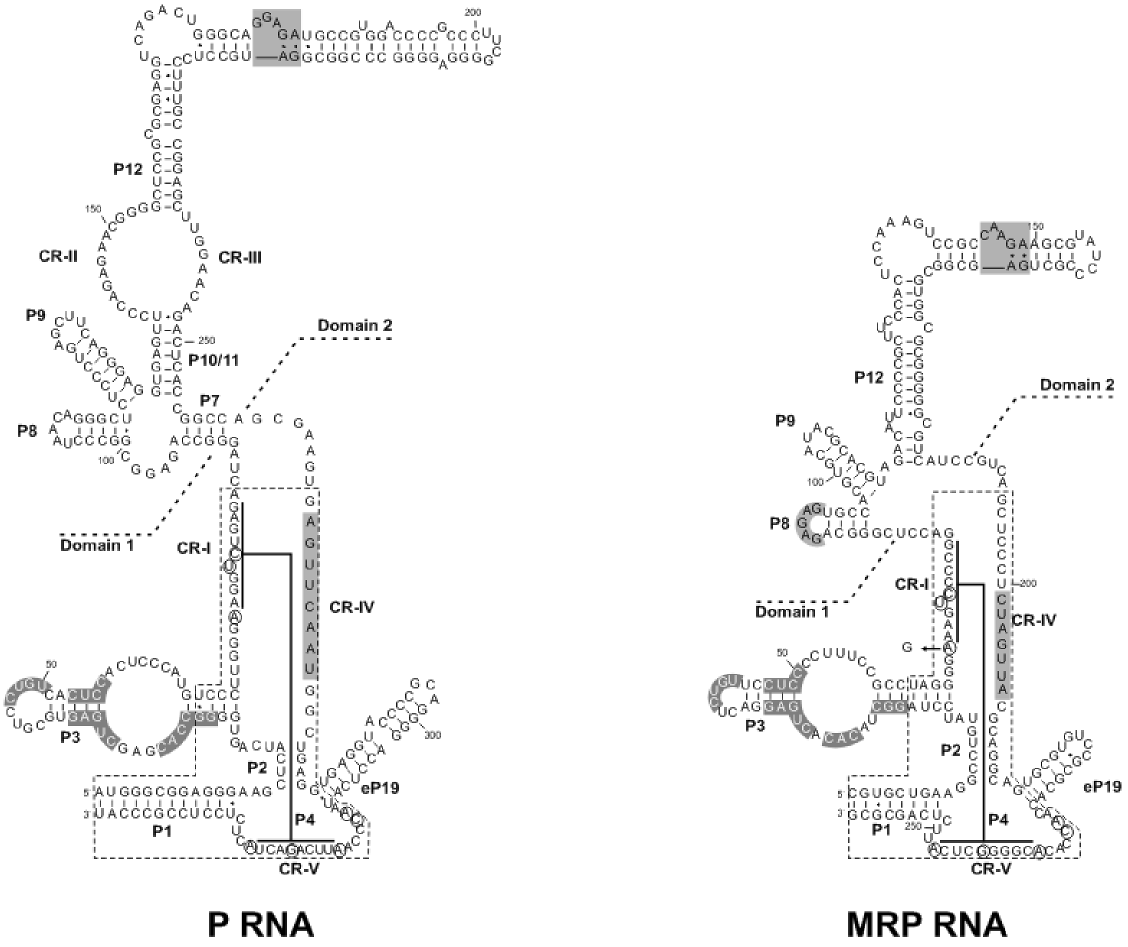

Figure 1. Secondary structure models of human RNase P and MRP RNAs. They are organized into two different domains, one catalytic domain and one specificity domain known to bind pre-tRNA substrates[38]. The universal consensus structure[39-40] comprises five critical regions, termed CR-I through CR-V. Both RNAs contain landmark helical elements (P1-4) and comprise a pseudoknot structure formed by base pairing of CR-I and CR-V[33]. The CR-IV region, the ‘GARAR’ sequence of P8, sequences in the P3 helix that are shared by P and MRP RNA and the K-turn motifs are shaded. The dotted lines highlight a characteristic region that is universally present.

Both RNases P and MRP have an RNA subunit and one or more protein subunits, which are required for catalysis in eukaryotes and Archaea[41]. P and MRP RNAs are similar in terms of sequence and secondary structure[33] (Fig. 1).

As opposed to bacterial RNase P, which has a single protein subunit, the eukaryal and archaeal RNase P contain multiple protein subunits that contribute to the stabilization and localization[42] of the RNA and to the architecture of the RNA active site[43].

Human RNase P and MRP contain at least 10 protein subunits[44-45], though not all of these subunits are shared between P and MRP (Table 3). In yeast at least nine protein subunits are part of the nuclear RNase P[46]. All but Rpr2, which is unique to RNase P[47], are present also in the RNase MRP as well as the MRP-specific proteins Snm1[48] and Rmp1[49]. The archaeal RNase P, from *Pyrococcus horikoshii*, has 5 protein subunits homologous to the eukaryotic counterpart[50-51].

Table 3. Protein subunits of RNases P and MRP. Proteins in the same row are homologous. Rpp14 and Pop5 are homologous. Ph1496 is the archaeal L7Ae protein. (a) Protein may be absent according to Welting et al[52].

H. sapiens			***S. cerevisiae***			***P. horikoshii***	
	MRP	**P**		**MRP**	**P**		**P**
hPOP1	+	+	POP1	+	+		
RPP38	+	+	POP3	+	+	Ph1496	+
RPP29	+a	+	POP4	+	+	Ph1771	+
hPOP5	+	+	POP5	+	+	Ph1481	+
RPP25	+	+	POP6	+	+		
RPP20	+	+	POP7	+	+		
RPP14	+a	+	POP8	+	+		
RPP30	+	+	RPP1	+	+	Ph1877	+
RPP21		+	RPR2		+	Ph1601	+
			SNM1	+			
			RMP1	+			
RPP40	+	+					

The spliceosome

Splicing is an essential step of gene expression in which introns are removed and exons are joined by two sequential trans-esterification reactions. These are catalyzed by a multicomponent complex, the spliceosome[53].

There are two known intron classes, the U2-type and U12-type, that are spliced by the U2-dependent (major) or U12-dependent (minor) spliceosome, respectively. The major spliceosome includes the U1, U2, U4, U5 and U6 spliceosomal RNAs (snRNAs) as well as multiple protein factors. The minor spliceosome contains several protein subunits and the U5 snRNA as well as the U11, U12, U4atac and U6atac snRNAs that are functionally and structurally related to the U1, U2, U4 and U6 RNAs of the major spliceosome[54].

For U2-type introns, spliceosome assembly is initiated by the interaction of U1 snRNP with the 5’ splice site and U2 snRNP with the branch site followed by the association of the U4–U5–U6 tri-snRNP complex. Structural rearrangements then take place where U6 separates from U4 and pairs to U2. U6 interacts with the 5´ splice site and U1 is displaced from the spliceosome. The U6/U2 complex plays an important role in the catalytic reaction[55]. The assembly of the minor spliceosome is similar to that of the major spliceosome; however a difference is that U11 and U12 snRNPs form a highly stable di-snRNP that binds cooperatively to the 5´ splice site and branch site[56].

U2-type introns are distributed across eukaryotes while U12-type introns have been demonstrated only in vertebrates, insects, cnidarians[57], *Rhizopus oryzae*, Phytophthora and *Acanthamoeba castellanii*[58]. U12-type introns are absent from the yeast *Saccharomyces cerevisiae* and from the nematode *Caenorhabditis elegans*[54].

Histone mRNA 3’ end processing

All eukaryotic mRNAs end with a poly(A) tail at their 3’ end, except for the metazoan replication-dependent histone mRNAs, which instead end with a highly conserved stem-loop structure (SL)[59], an example of an RNA regulatory element. These histone mRNAs encode the four core histones (H2A, H2B, H3 and H4) and a linker histone (H1). The corresponding genes are clustered in the genome and lack introns. Typically their transcription rate increases as cells approach S phase[60].

In the 3’ end processing of replication-dependent histone mRNAs the SL sequence binds to the stem-loop binding protein (SLBP)[61]. Then the U7 small nuclear RNA (U7 snRNA) base-pairs with the histone downstream element (HDE)[62]. A cleavage complex containing CPSF73[63], CPSF100 and symplekin[64] among other factors, is recruited to cleave the pre-mRNA. The cleavage occurs five nucleotides downstream of the SL. The U7 snRNA, a component of the U7 small nuclear ribonucleoprotein (U7 snRNP), contains an Sm protein-binding site, where five (B, D3, G, E, and F) of the seven Sm core proteins assemble together with two U7-specific proteins (Lsm10 and Lsm11)[65-66]. A zinc finger protein (ZFP100)[67] interacts with both Lsm11 and SLBP, stabilizing the U7 snRNP-pre-mRNA interaction (Fig. 2).

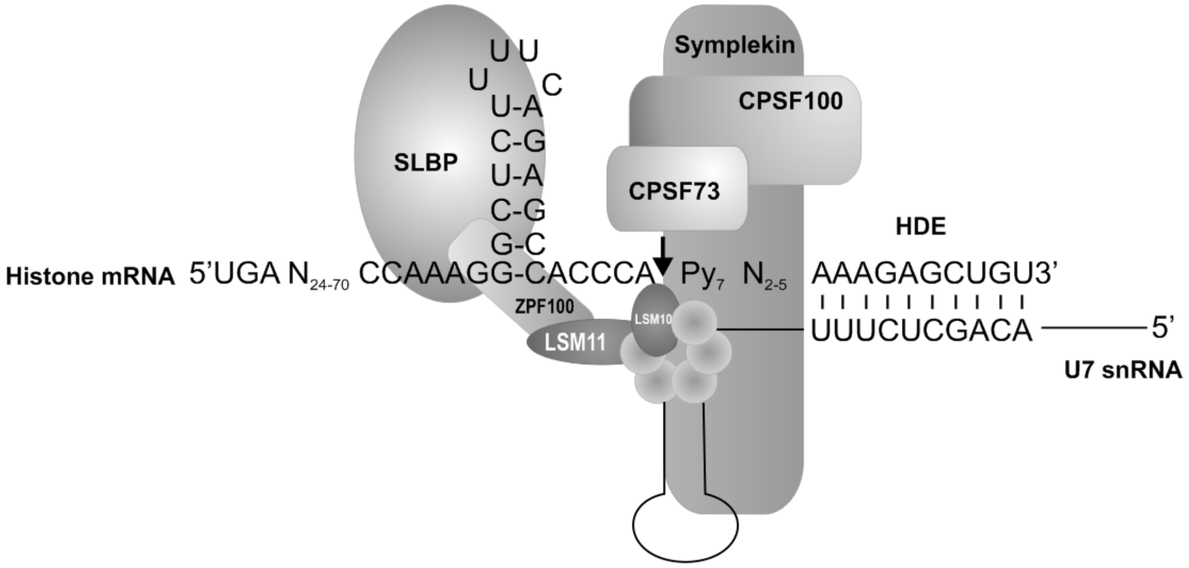

Figure 2. Processing of mammalian canonical histone pre-mRNA. Sm proteins are depicted as circles, while U7 snRNP-specific Sm-like proteins, Lsm11 and Lsm10 are dark ovals. The cleavage site is indicated with an arrow. Figure adapted from Marzluff et al[59].

Gene organization and gene order

Recombination processes result in part of a genome being restructured due to DNA segments crossing over between chromosomes. This has the effect that genes are extensively shuffled during evolution. In bacteria many genes are organized in operons where the genes are functionally related and so the shuffling of genes is constrained. As opposed to bacteria, eukaryotic genes are not restricted in this way.

However there is evidence that suggests that the order of genes in eukaryotes is not entirely random. There are examples where genes tend to cluster, as when they have similar expression[68-71] or are functionally related. Similarly, genes that encode subunits of stable complexes [72-74] or are involved in the same metabolic pathway also tend to cluster[75]. The intergenic distance between genes has been shown to be a strong predictor of gene order conservation in Fungi[76], while in mammals gene pairs that are divergently transcribed with a short intergenic distance are more abundant[73-74,77-80].

Thus, it seems likely that where strong gene order conservation is observed, the genes implicated could be related in terms of transcriptional control. In this section we examined non-random gene order in a wide range of eukaryotic organisms to identify the most important determinants of gene order conservation. In addition we focused on pairs of adjacent genes that are homologous to identify genes that may be transcriptionally linked.

Methodology aspects

Identification of protein homologues

Homologous proteins where identified with sequence similarity methods such as BLAST or FASTA. However since organisms from distant phylogenetic groups were analyzed, it was likely that the proteins under study have diverged significantly and thus searches based on pairwise alignments would not be enough. Therefore profile-based searches were applied (PSI-BLAST and HMMER either with Pfam or custom made models), which are more sensitive tools as they make use of information in a multiple sequence alignment rather than a single sequence.

In order to identify as many homologues as possible, it is important to implement an iterative process, where the resulting hits are used as queries in subsequent runs. Therefore, automation is a key step to perform this task.

In cases where protein sets were not available for a given organism, the corresponding genome sequences were scanned with TBLASTN. The resulting hits and their corresponding flanking regions were retrieved and trimmed with GeneWise[81], an homology-based algorithm that predicts gene structure using protein sequences.

Identification of RNA genes

As opposed to proteins, ncRNAs have limited sequence conservation; however they tend to maintain a consensus secondary structure. Sequence similarity searches aided in the identification of ncRNAs sequences from closely related species. In order to identify ncRNAs in more distantly related species, patterns and statistical models that incorporate primary sequence and secondary structure information such as RNABOB and covariance models were used. As an example, for the identification of snRNA homologues (Fig. 3), an iterative process was implemented, where a similarity search was run as a filtering step, since the use of CMs on genomic sequences is extremely computationally heavy. The resulting hits were then scanned with the corresponding CM and the hits above a specific threshold were checked for the presence of essential primary sequence motifs as well as its ability to fold into the typical secondary structure. In these protocols, the evaluation of essential elements in the candidates gives further support and makes our predictions more reliable.

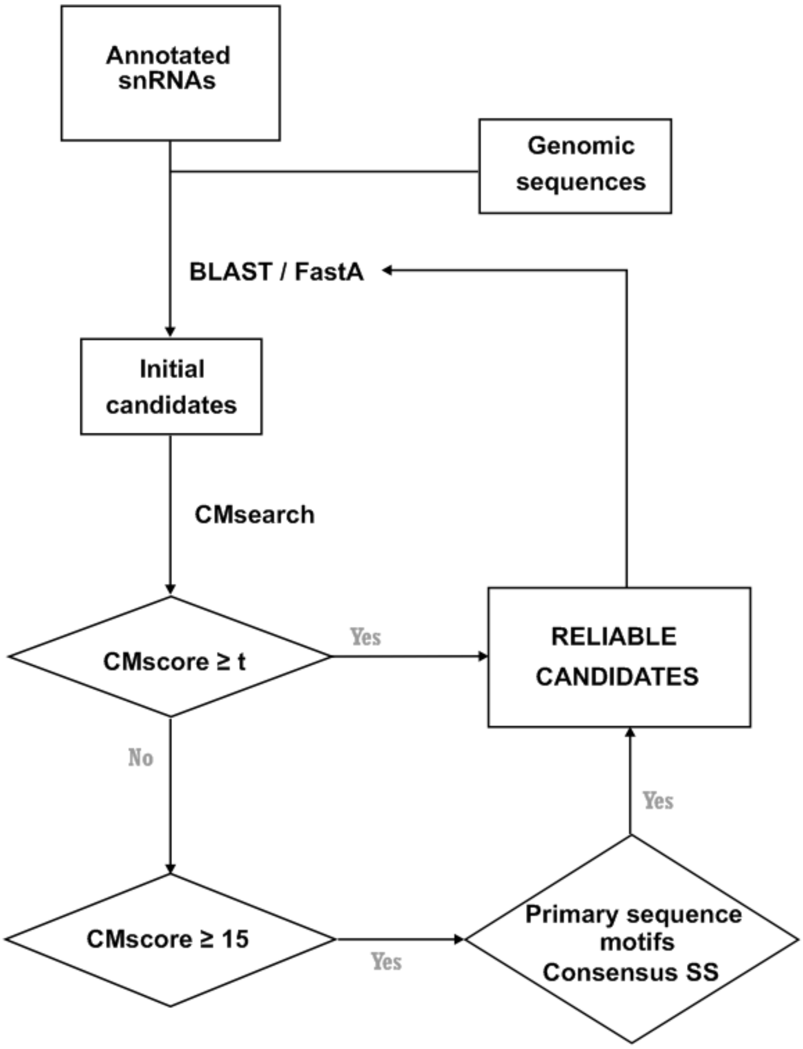

Figure 3. Protocol for snRNA detection. Annotated snRNAs (U1, U2, U4, U5, U6, U11, U12, U4atac and U6atac) from Rfam[82] were used as initial queries with BLAST or FASTA against genomic sequences. Significant hits were retrieved and analyzed with cmsearch and the corresponding CM. All sequences above a specific threshold *t* were considered as reliable candidates. Sequences with a lower score than the threshold but greater that 15 were examined with respect to primary sequence motifs and secondary structure characteristic of the specific snRNA. Sequences that meet the criteria were considered as reliable candidates. Predicted sequences were used as queries in a second round of searches to retrieve homologues in species where a particular snRNA was not yet identified. This was repeated until no more significant hits were retrieved.

It is important to mention that although the use of CMs is a sensitive method, ncRNAs from other phylogenetic groups can escape detection when they are not represented in the CM. As an example, most of the protozoan histone mRNA stem-loop structures remained undetected when the corresponding CM was used, due to a bias towards metazoan stem loop structures. To overcome this problem, we performed pattern searches with RNABOB, resulting in potential protozoan stem-loop structures. In this manner, the combination of different methods increases the sensitivity of ncRNA identification.

Phylogenetic analysis

Besides showing the probable evolution of various organisms through the comparison of homologous sequences, phylogenetic analysis aids in the correct classification of homologue proteins. As an example, two protein subunits of RNase P, Pop5 and Rpp14 contain the same protein domain, thus it is difficult to correctly classify them by profile-based searches alone. Here, phylogenetic analysis proved to be necessary in the classification of such homologues.

Orthologue identification using OrthoMCL

OrthoMCL[83] is a software that clusters proteins based on sequence similarity, using an all-against-all BLAST search of each species proteome, followed by normalization of inter-species differences, and Markov clustering. The aim of OrthoMCL is to generate clusters where the members of each cluster are orthologues. It uses a relatively non-stringent e-value (1e-^{5}) to include distantly diverged orthologues and several rules are applied during the process to eliminate poorly alignable sequences. We noted that the clustering is somewhat ambiguous since it is dependent on the actual set of protein sequences that is used. However, when more data is added to the pipeline, the clusters tend to be smaller and more consistent with respect to their protein domain architecture.

Gene Ontology assignment

Gene ontology terms are controlled vocabularies for describing gene product characteristics and thus give an insight of its function. In this work, a high number of proteins lack GO annotation, since we included organisms that are at different stages in the sequencing of their genome. To annotate these proteins, we performed stringent sequence similarity searches against proteins from UniProt Knowledgebase with GO annotation. There were cases where no term could be assigned to a protein. However since orthologous proteins are expected to retain similar function, we made use of the clusters obtained with OrthoMCL. Thus, if an unannotated protein belonged to a cluster where at least one of the proteins was annotated with a GO term, this term was assigned to the unannotated protein. A drawback of this method is that any error from the clustering will be propagated to the GO annotation.

Results and Discussion

This work is about computational prediction for protein and non-coding RNA genes and is composed of two different parts. The first deals with the identification and analysis of protein and non-coding RNA genes from RNA-protein complexes that were previously not well explored with respect to their phylogenetic distribution:

- RNase P and MRP (Paper I)
- Spliceosome (Paper II)
- Histone 3’ end processing (Paper III)

In the second part we identified a number of factors associated with non-random gene order by examining a large number of eukaryotic species (Paper IV). Data related to this project may be accessed through the eukaryotic Gene Order Browser (eGOB) at http://egob.bioimedicine.gu.se (Paper V).

Ribonucleases P and MRP (Paper I)

RNases P and MRP are involved in tRNA and rRNA processing, respectively. A large number of RNases P and MRP RNAs had previously been predicted[84]. In order to further analyze the structure and evolution of RNases P and MRP their protein subunits were now identified and examined.

Inventory of the protein subunits lead to novel protein relationships

Novel homologues of the protein subunits were identified in a wide range of eukaryotic organisms. Their distribution show that proteins with essential roles such as Pop1, Pop4, Pop5 and Rpp1 are widely distributed which is consistent with the distribution of P and MRP RNA. In addition, interesting novel protein relationships were also found as described below (Fig. 4).

Rpp20/Pop7/Rpp25 family. These proteins belong to the ALBA superfamily[85], a family of DNA/RNA-binding proteins. Homology between Rpp20 and Pop7 was proposed previously[86]. Our results lend further support to this notion.

		PROTEIN SUBUNITS								
		Pop3	**Rpp38**	**Pop6**	**Rpp25**	**Pop7**	**Rpp20**	**Pop8**	**Rpp14**	**Pop5**
Fungi	Ascomycota	▨		▨		▨		▨		▨
	Basidiomycota	▨								
	Microsporidia									▨
Metazoa	Vertebrates		▨		▨		▨		▨	▨
	Nematoda				▨		▨		▨	▨
	Insects				▨		▨		▨	▨

Figure 4. Phylogenetic distribution of selected RNase P and MRP protein subunits. Inferred homologies between fungal and metazoan homologues are paired next to each other. Boxes with shaded background represent organisms where a protein homologue was identified with profile-based searches such as PSI-BLAST.

Rpp14/Pop5 and Pop8. Rpp14 and Pop5 have been previously noted to be homologues[87]. Rpp14 is found only in metazoa while Pop5 is more widely distributed. Here, Pop8 was found to share an evolutionary relationship with these two proteins. A possibility is that Pop8 is the fungal orthologue of Rpp14.

Rpp25/Pop6. This relationship was identified by profile-based searches and is consistent with protein-protein interaction data since in human, Rpp25 interacts with Rpp20 and binds to Rpp29[45], while in yeast, Pop6 interacts with Pop7 (the Rpp20 homologue) and binds to Pop4 (the Rpp29 homologue)[46].

A K-turn motif in P and MRP RNAs might interact with Rpp38

P and MRP RNA homologues were identified in most of the major phylogenetic groups. Plants and heterokonts lack P RNA while Euglenozoa lacks both, P and MRP RNA, which is consistent with the absence of their protein counterparts. However, an MRP RNA candidate has been presented in *Trypanosoma* *brucei*[88]. This RNA as well as homologous sequences in closely related species lack some consensus features expected of MRP RNA. Thus, if MRP RNAs are present in Euglenozoa they are clearly different from previously known members of this RNA family.

Further analysis of P and MRP RNA sequences revealed an RNA secondary structure motif, the K-turn motif, which is important for protein recognition and stabilization of RNA tertiary structure (Fig. 1). This motif was previously described for MRP RNA but not for P RNA. It is known that K-turns interact with ribosomal proteins, including the archaeal L7Ae protein, a subunit of the archaeal RNase P. Since Rpp38 and Pop3 are homologues to this archaeal protein and share the same L7Ae domain, it is likely that these proteins bind to the K-turns of P and MRP RNA.

Summary

RNases P and MRP from yeast and human have previously been experimentally characterized. A number of proteins were shown to be shared by the yeast and human enzymes, such as Pop1 (human hPop1), Pop4 (human Rpp29), Pop5 (human hPop5) , Rpp1 (human Rpp30) and Rpr2 (human Rpp21)[89]. Here we have demonstrated that the homology between the yeast and human enzymes is even more extensive. Thus, we present evidence for an evolutionary relationship between yeast Pop7/human Rpp20, yeast Pop3/human Rpp38, yeast Pop8/human Rpp14 and yeast Pop6/human Rpp25. These results provide further evidence of homology between the human and the yeast RNPs.

Spliceosome (Paper II)

In order to analyze the evolution and the phylogenetic distribution of the RNA molecules of the spliceosome, which are critical for specificity and catalysis during splicing of eukaryotic pre-mRNA, a computational screen was performed to predict such RNAs in an extensive set of eukaryotic organisms.

U2-type spliceosomal snRNAs are widely distributed

In virtually every species examined, U2-type spliceosomal snRNAs were identified. We showed that U2-type snRNAs are ubiquitous in the Basidiomycota lineage and presented evidence for the first time that they are present in the Zygomycota and in the Chytridiomycota branches. We identified novel snRNAs in Microsporidia, showing that U2-type snRNAs are present in all fungal groups. Even in the smallest eukaryotic genomes known to date, the nucleomorphs of *Guillardia theta* and *Bigelowiella natans*, U2-type snRNAs were found. However, despite the presence of three introns[90-91] and approximately 27 spliceosomal proteins[92] in *Giardia lamblia* and several introns as well as U2 and U5 snRNP-protein specific subunits in *Cyanidioschyzon merolae* [93] no spliceosomal RNAs were found in these organisms, suggesting that spliceosomal RNAs are lacking and have been replaced by protein functions or alternatively, that the spliceosomal RNAs are very different from those in other species.

Identification of novel U12-type spliceosomal snRNAs confirms an early origin of the U12-type spliceosome

U12-type introns were previously identified in plants, most metazoan taxa[57], *R. oryzae, A. castellanii* and Phytophthora[58] while minor spliceosomal RNAs have also been found in these organisms, except for *R. oryzae*. In contrast, organisms such as *S. cerevisiae* and *C.elegans* seem to lack the U12-type splicing[54]. Here, several novel minor spliceosomal RNA orthologues have been identified in organisms where these snRNAs have not been previously reported, such as the nematode *Trichinella spiralis*, the amoeba *A. castellanii* and the slime mold *Physarum polycephalum*, as well as organisms from the Zygomycota and

Chytridiomycota fungal lineages. These results, together with the identification of U11/U12 specific proteins in these organisms provide further evidence of the presence of a minor spliceosome. Moreover, the phylogenetic distribution of U12 introns has shown to be entirely consistent with the distribution of U12 snRNAs[94]. Therefore, our current knowledge of the phylogenetic distribution of U12-type snRNAs points to several instances where the minor spliceosome was lost during evolution (Fig. 5). For instance, in the branch of nematodes, the deeply branching *T. spiralis* has U12-dependent splicing, whereas it was lost in the branch leading to Caenorhabditis.

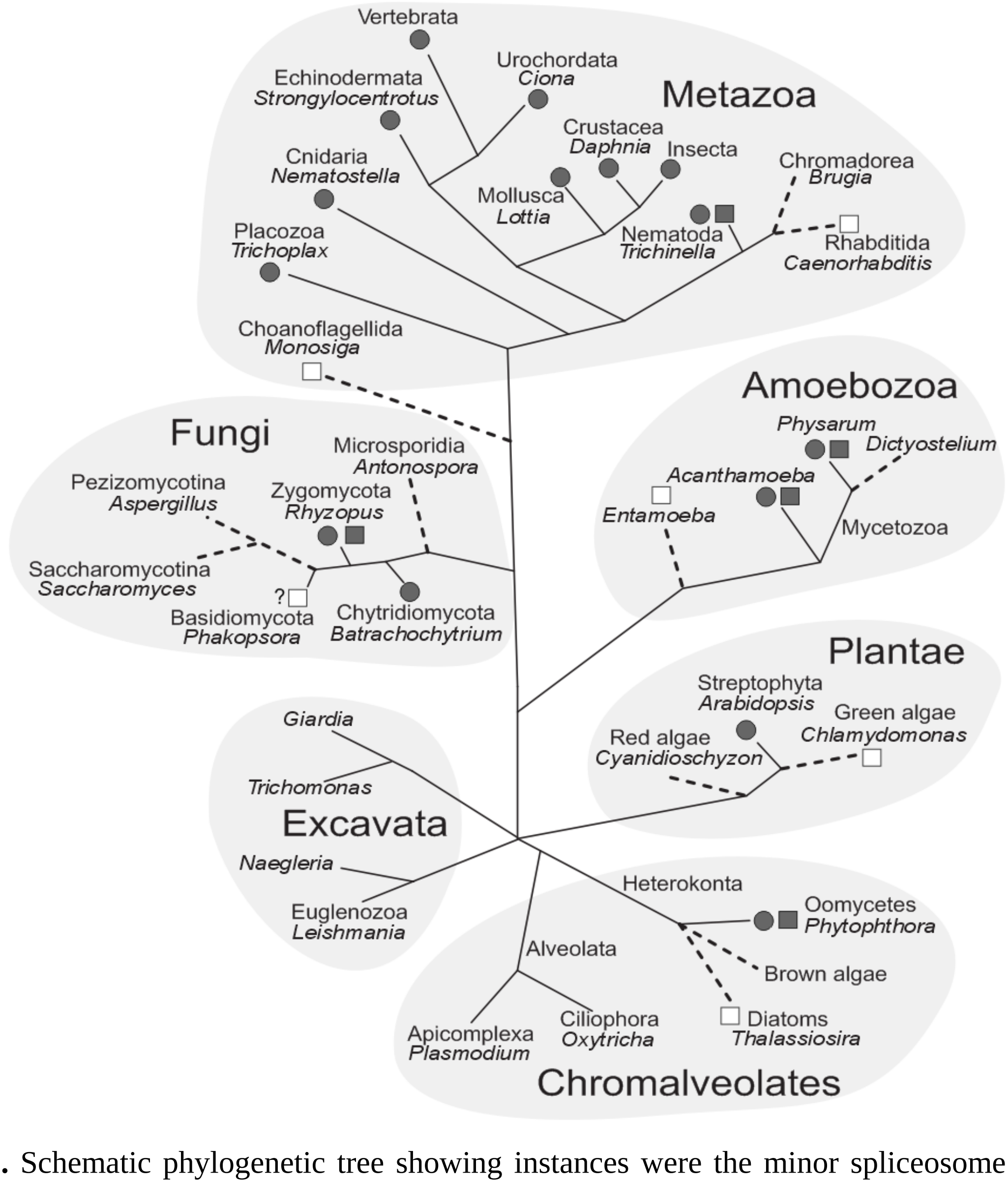

Figure 5. Schematic phylogenetic tree showing instances were the minor spliceosome was lost. Species where one or more U12-type spliceosomal RNAs were found are depicted as filled circles. Presence or absence of U12 introns are shown as filled or empty squares, respectively[94]. Dashed lines indicate branches where minor snRNAs seem to have been lost.

Summary

To examine the phylogenetic distribution and evolution of the RNA molecules of the spliceosome a systematic analysis was performed on a broad range of eukaryotic genomes. Several U2 and U12-type snRNAs were predicted in many phylogenetic groups where these RNAs were not previously reported, such as in all fungal lineages, lower metazoa and many protozoa. The detailed map on the phylogenetic distribution of U12-type RNA genes supports an early origin of the minor spliceosome and points to several occasions where it was lost during evolution.

Histone mRNA 3' end processing (Paper III)

Histone pre-mRNAs are not polyadenylated and contain a unique 3' end stem loop (SL) structure. Formation of the 3' end of histone mRNAs is generated by a specialized machinery that involves the U7 snRNP and protein factors[59]. With the aim of examining the phylogenetic distribution and evolution of the histone mRNA 3' end processing machinery a systematic search was carried out with respect to the molecular components of this machinery.

Identification and distribution of cis and trans-acting elements of the histone 3' end processing machinery

It was previously assumed that the machinery of histone 3' end processing was specific to metazoa although components had also been identified in green algae[95]. In order to more closely examine the phylogenetic distribution of this mechanism we searched for novel homologues of the *cis* and *trans*-acting elements.

The 3' end stem-loop is also found in protozoa. A large collection of histone mRNAs from a variety of eukaryotic species were analyzed with respect to the histone mRNA stem-loop. SL motifs were found in most metazoa, but interestingly also in a number of protozoa. Unlike metazoan SL motifs, protozoan SL motifs are less frequent and the distance between the stop codon and the SL is variable. The purine-rich histone downstream element (HDE) that pairs with a region of U7 RNA is present in all metazoan and absent in most of the protozoan sequences. In Fungi, SL motifs are completely absent while a few were identified in plant species. Histone genes without an SL, could correspond to the histone variants[96] that as opposed to the replication-dependent histone mRNAs, are polyadenylated and expressed throughout the cell cycle. These two types of histone genes could not be distinguished in this work.

Stem-loop binding protein homologues are found in organisms where the SL motif is present. Known homologues of SLBP contain a strongly conserved RNA-binding domain[97-98] that interacts with the SL structure. Here we identified metazoan and possible protozoan homologues from most of the organisms where a

SL was found, providing stronger evidence that in these organisms there is a histone mRNA 3' end processing machinery.

Novel U7 snRNA and U7 snRNP-specific proteins are identified in metazoa. In addition to the previously known homologues to U7 RNA (mammals, sea urchin and insects[99]) novel homologues of U7 snRNA were identified in teleosts. All U7 RNA share a high degree of sequence similarity and have the typical Sm site and hairpin as shown in Figure 6. The U7 snRNP-specific proteins, Lsm10 and Lsm11, were identified in metazoa but not in protozoa, except for *Dictyostelium discoideum*. These findings together with the fact that no metazoan-like HDE was identified in protozoa suggest that if a U7 RNA is present in protozoa and lower metazoa it may have properties distinct form those of previously known members of this RNA family.

```
H. sapiens         CUUUUAG-AAUUUGUCUAG-UA-GGCUUUCUGG.....CUUUUCA......CCGGAAAGCC-CCU
M. musculus        CUUUUAG-AAUUUGUCUAG-CA-GGUUUUCUGAC.....UUCG........UCGGAAAACC-CCU
G. gallus          CUUUUAG-UAUUUGUCCAG-CA-GGUUUCCCG.......CCCCG........CGGGAAGCC-CAA
X. borealis        CUUUUAC-UAUUUGUCUAG-CA-GGUUCUUAC........UCU.........GUAGGAGCC-ACA
D. rerio           CUUUUAG-UAUUUGUCUA--CA-GGCUUCCUU........UAA.........AAGGAAGCC-CAC
G. acuelatus       CUUUAGA-UAUUUUUCUAG-UA-GGUUUCUC........GUAAA.........GAGAAGCC-CUC
O. latipes         CUGAAGA-UAUUUGUCUAG-CA-GGUUUCUC........AUAAA.........GAGAAGCC-CCU
P. marinus         UUUUUAU-UAUUUGUCUAG-UA-GGUCUGUC.........UCU.........GACGGACCG-CAC
F. rubripes        CUUUAGA-UAUUUCUCUAG-UA-GGCUUUUC........AUACA.........GAGAAGCC-CCC
U. nigroviridis    CUUUAGA-UAUUUCUCUAG-AA-GGCUUCUC........AUAAU.........GCGAAGCC-CCC
S. purpuratus      UCUUUCA-AGUUUCUCUAG-CA-GGGUCUCGCAUCCG..AAGU...CGGACGCGAGUGCCC-ACC
P. miliaris        UCUUUCA-AGUUUCUCUAG-AA-GGGUCUCGCGUCCG..AAGU...CGGAGGCGAGUGCCC-ACC
B. floridae        UGUUGGU-UAUUUCUCUAA-UC-GGUUCUU........CAUACUC.........AAAAGCC-ACA
D. melanogaster    UCUUUGA-AAUUUGUCUUG-GU-GGGACCCUUUGU....CUAG....GCAUUGAGUGUUCC-CGU
                                          <<<<<<<<<<<<<<.........>>>>>>>>>>>>>>>
```

Figure 6. Alignment of U7 snRNA.

The histone 3’ end processing developed early in the evolution of eukaryotes

Important components of histone 3’ end processing are present in many different protozoa (Fig. 7). In the light of these results, the histone 3’ end processing developed very early, but was partially or completely lost in the development of protozoa, plants and fungi. It was previously thought that the histone 3’ end processing was developed from the polyadenylation machinery[95], but it now seems that the machineries of polyadenylation and histone mRNA processing diverged very early; however we cannot reach a conclusion as to the ancestral version of these processing pathways.

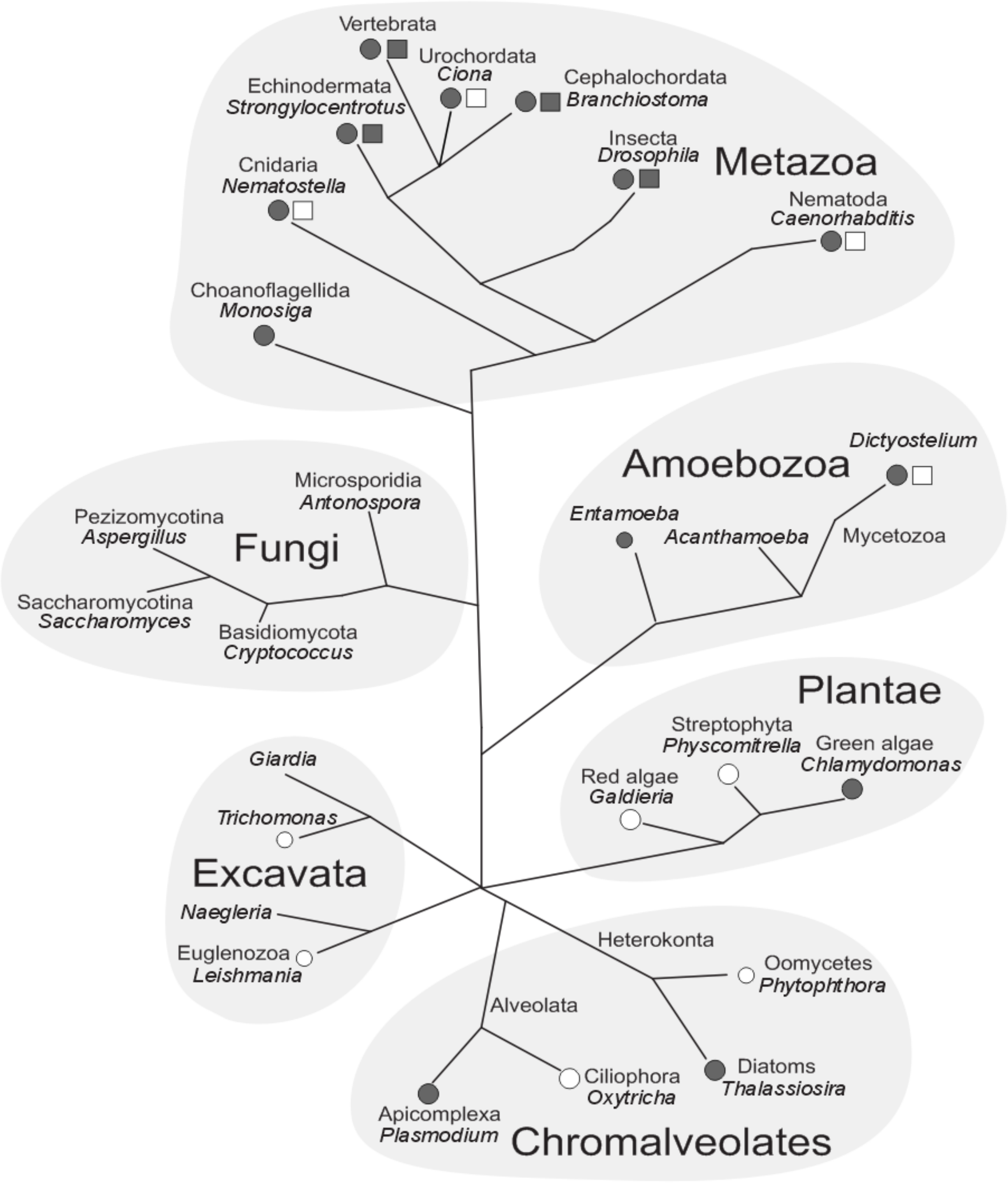

Figure 7. Schematic phylogenetic tree showing the distribution of components involved in histone 3’ end formation. Species where SL motifs (filled circles) and U7 snRNA (filled squares) were identified are shown. Lack of SL motif and U7 snRNA is depicted as empty circles and squares, respectively.

Summary

With the aim of examining the phylogenetic distribution and evolution of the histone mRNA 3’ end processing machinery a systematic search was carried out to analyze genomic data of a broad range of organisms. The SL motif and the SLBP protein were identified in metazoa as well as in many protozoa. Novel homologues of U7 snRNA and U7 snRNP-specific proteins where identified in teleosts. These results suggest that the processing of histone mRNAs is an early invention in the evolution of eukaryotes.

Gene Order (Paper IV and Paper V)

As of today, a limited number of eukaryotic species have been studied in relation with non-random gene order. We took here a more systematic comparative genomics approach, where organisms representing all important eukaryotic phylogenetic groups were considered.

Analysis of factors that are associated to non-random gene order

We focused on pairs of neighboring genes and analyzed parameters suspected to be related to gene order such as the relative transcription direction, intergenic distance and functional relationships as inferred from gene ontology; all in relation to evolutionary conservation.

Divergently and co-directional gene pairs are more common among evolutionarily conserved gene pairs. It has been shown that in prokaryotes, divergently and co-directional transcribed gene pairs are conserved across distant species[100]. Here we showed that also in eukaryotes, conserved gene pairs arranged in a divergent and co-directional fashion are much more common than convergent pairs. The high frequency of co-directional gene pairs in bacteria is due to the presence of polycistronic operons, while in eukaryotes this might be due to gene duplication.

Divergently transcribed gene pairs with short intergenic regions are enriched in mammals. There are several reports where divergently transcribed gene pairs with a short intergenic distance are enriched in the human genome[73-74,77-80]. By examining the distribution of intergenic distance sizes, we also observed this enrichment in mammals as well as in birds and frogs, suggesting that short intergenic regions of divergently transcribed genes developed during the evolution of terrestrial vertebrates, and are more significant in mammals.

Divergently transcribed genes that are evolutionarily conserved tend to be functionally related. Prediction of functional associations from conserved divergently transcribed genes has been demonstrated for prokaryotes[100-101]. Here we observed that for metazoa, co-directional and divergently transcribed genes are likely to be related by function when the pair is strongly conserved. In fungi a similar trend is observed only for divergently transcribed gene pairs.

Interesting gene pairs that were not previously recognized include two fungal pairs of ribosomal protein genes (Table 4). One of the most conserved gene pairs, L13/S16 is also present in apicomplexans. In Bacteria[102] and Archaea[103], L13 and S9 (homologue of S16) are transcriptionally linked. This gene pair is then present in all kingdoms of life suggesting a strong functional relationship between the two proteins.

Table 4. Five most evolutionarily conserved gene pairs in Fungi.

Protein gene A	Protein gene B
60S ribosomal protein L13	40S ribosomal protein S16
Pirodoxine biosynthesis protein SNZ1	Glutamine amidotransferase SNO1
Histone H2A	Histone H2B
60S ribosomal protein L21	40S ribosomal protein S9
DNA replication licensing factor MCM2	Protein mlo2

Human gene pairs that are likely to be regulated by bidirectional promoters

Divergently transcribed genes with an intergenic region less than 1000 base pairs are assumed to have a "bidirectional promoter"[77]. In the human genome, we found that 8% of the divergently transcribed genes could harbor a bidirectional promoter. Only about 0.34% of these gene pairs have previously been shown to be regulated by bidirectional promoters (Fig. 8).

Analysis of these bidirectional gene pairs showed that highly evolutionarily conserved gene pairs are likely to involve bidirectional promoters and to have genes that are functionally related. Examples of protein-protein gene pairs that were identified in this work and that have been previously characterized as having bidirectional promoters include 60 kDa/10 kDA shock proteins[104], H2A/H2B histone proteins[105] and collagen type IV alpha 1 and 2[106]. Besides, we were able to identify gene pairs that may be of interest for studies of transcriptional control, such as three olfactory receptor gene pairs, two subunits of the ligand gated ion channel and two heat shock proteins.

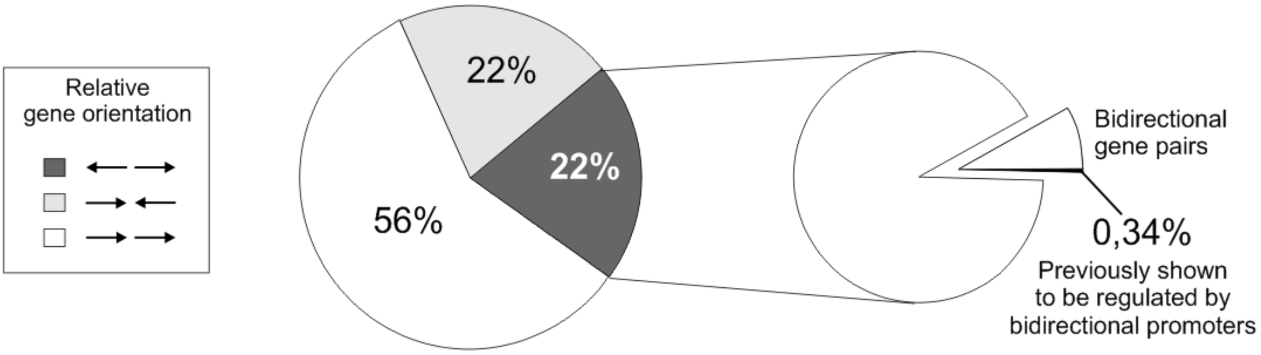

Figure 8. Relative orientation of human gene pairs and the distribution of bidirectional gene pairs.

There are only a few known cases of divergently transcribed gene pairs that involve ncRNA genes where a bidirectional promoter is suggested to regulate transcription. Examples are PARP2/RPPH1, an RNA pol II gene and the RNase P RNA gene sharing the same promoter region[107]; and HAND2/DEIN where their similar expression profiles in primary neuroblastoma are corregulated by asymmetrical activity of the promoter[108]. In this work we were able to identify novel RNA-protein gene pairs that may be regulated by a bidirectional promoter as shown in Table 5.

As for RNA-RNA bidirectional gene pairs, those that are evolutionarily conserved are mainly pairs of tRNA genes.

It is important to mention that transcriptional control by bidirectional promoters is not limited to gene pairs less than 1 kb apart. There are examples where the distance is greater than 1 kb, e.g. proteins involved in the metabolism of xenobiotics (CYP1A1/CYP1A2[109]), proteins associated with hypotonia-cystinuria syndrome (PREPL/C2ORF34[110]) and proteins involved in Fanconi anemia (FANCA/SPIRE, FANCF/GAS2[74]).

Table 5. Conserved pairs of divergently transcribed genes from human that may be regulated by a bidirectional promoter.

ncRNA gene	Protein gene
Ser tRNA	N-acetyllactosaminide beta-1,3-N-acetylglucosaminyltransferase
U12 minor spliceosomal RNA	Polymerase delta-interacting protein 3
Arg tRNA	Ornithine cyclodeaminase
tRNA	Cytochrome b
microRNA MIR533	Zinc finger protein
Gln tRNA	Ras-like GTP-binding protein
SRP RNA	Ribosomal protein S29
U6 spliceosomal RNA	Mediator of RNA POLII transcription subunit 16
Small nucleolar RNA SNORD74	Zinc finger and BTB domain-containing protein 37

eGOB: eukaryotic Gene Order Browser

With the aim of making publically available all data generated in this project, we created the eukaryotic Gene Order Browser (eGOB). eGOB is useful for comparing and displaying a gene of interest together with its genomic context in all species where that gene is present. Therefore, questions related to genome evolution and gene organization may be examined.

The graphical view. As shown in figures 9 and 10, genes are represented by arrows, which denote the relative direction of transcription. Thick and thin arrows denote protein and ncRNA genes, respectively. Each gene is color-coded according to the clustering method and cluster id to which it belongs. Thus, genes with identical color indicate an orthology relationship. It is possible to toggle between coloring schemes to emphasize either the sequence (OrthoMCL/Rfam clustering) or functional similarity (Pfam grouping) among the displayed genes.

Visualizing the genomic context. In Figure 9, the genomic context of the H4/H3 histone gene pair is shown as well as its conservation across several organisms. In this case the coloring scheme is based on sequence similarity (OrthoMCL clustering, Fig. 9A) that allows us to easily identify H4 and H3 protein homologues. It is possible to analyze such pair with respect to a functional environment (Pfam grouping, Fig. 9B) by switching between coloring schemes. In this case, we can observe how the H4/H3 gene pair is part of a larger histone cluster.

Identifying genes that could be transcriptionally linked. The browser is also useful for exploring conserved gene pairs as a way of predicting genes that may be transcriptionally linked. An example is shown in Figure 10 where the 40S ribosomal protein S29 and the SRP RNA gene pair is conserved from human to the chicken and the intergenic distance between each pair less than 1 kb, suggesting that these genes are likely to be transcriptionally linked.

Summary

To examine non-random gene order we have carried out ncRNA gene prediction in a genome wide fashion. We have combined this information with data on protein gene localization. The results showed that factors such as relative gene orientation, intergenic distance and functional relationships are associated with non-random gene order. We have presented a list of conserved gene pairs that are of interest for further studies regarding transcriptional control. Information on gene order is presented in the eukaryotic Genome Browser (eGOB) where gene order may be displayed and compared between species. A gene of interest may be then studied in order to learn its genomic context or neighboring genes that are likely to be transcriptionally linked may be identified.

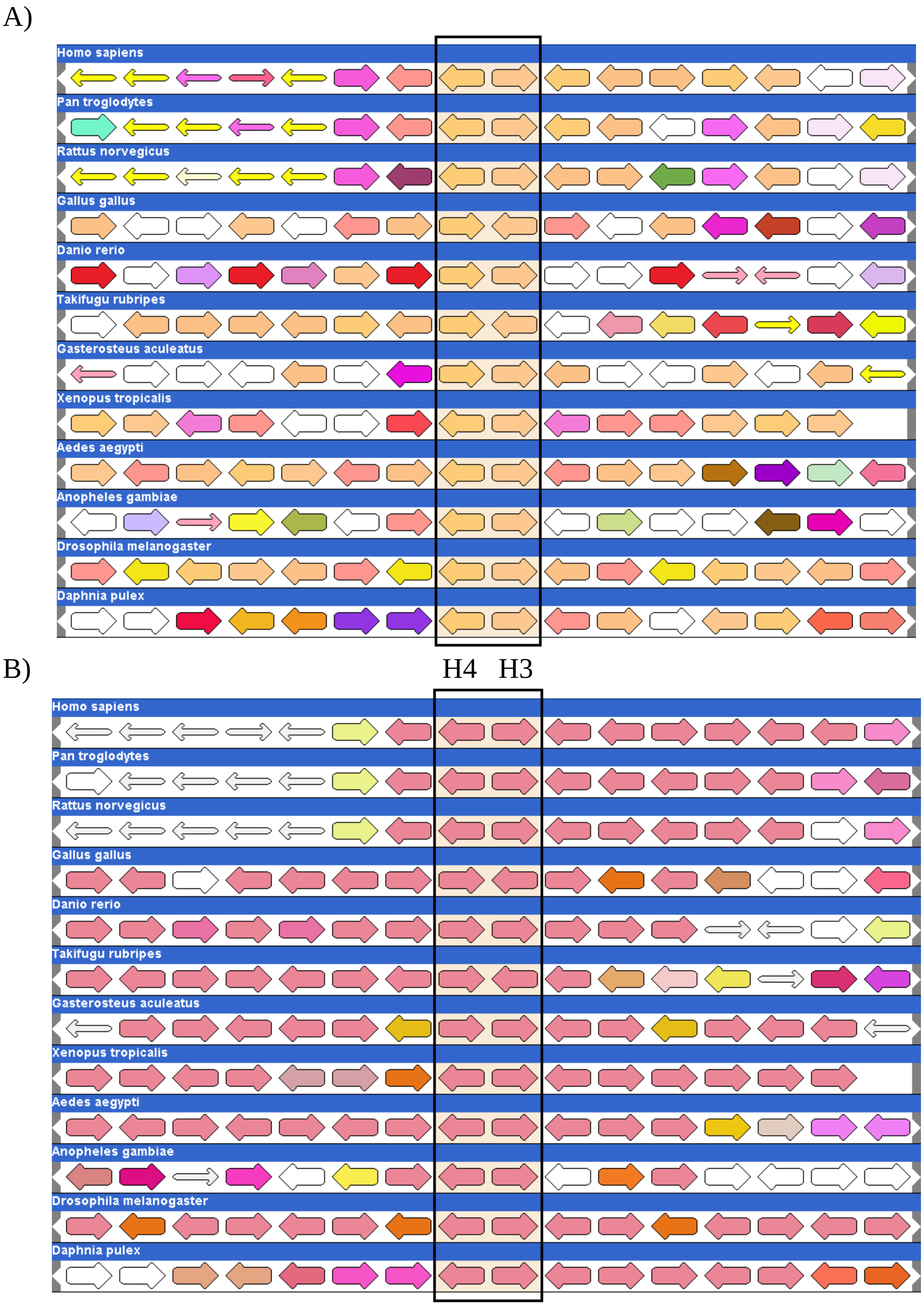

Figure 9. Genomic context of the H4 and H3 histone gene pair. Genes are presented by arrows, which denote the relative direction of transcription. Thick and thin arrows denote protein and ncRNA genes, respectively. Each gene is color-coded according to either the OrthoMCL/Rfam clustering (A) or the Pfam grouping (B).

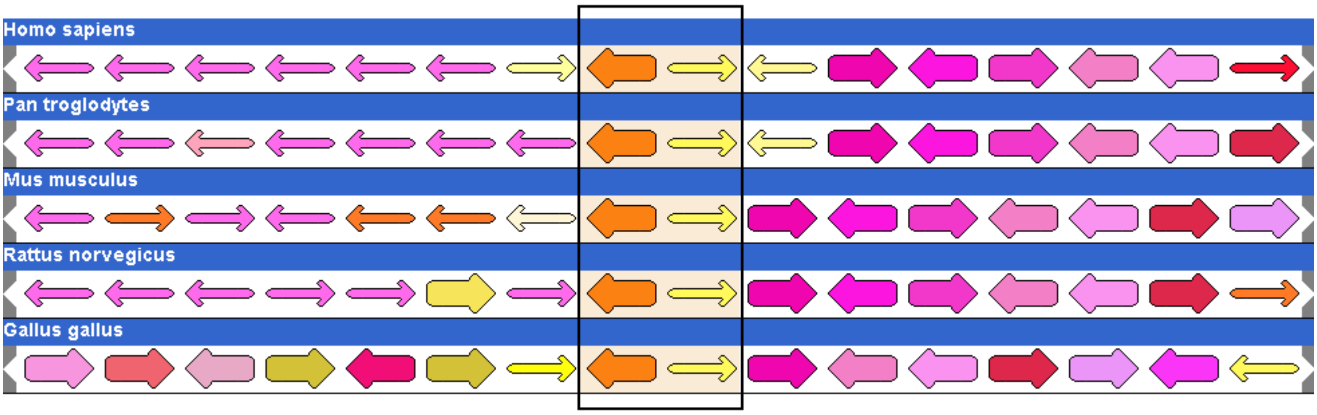

Figure 10. Genomic context and evolutionary conservation of the 40S ribosomal protein S29 and the SRP RNA gene pair. Genes are presented by arrows, which denote the relative direction of transcription. Thick and thin arrows denote protein and ncRNA genes, respectively. Each gene is color-coded according to the OrthoMCL/Rfam clustering.

Conclusions

From our work on protein and ncRNA genes, we are able to reach a number of conclusions as to the methodology:

- Our protocols for ncRNA homologue prediction proved to be sensitive as we combined different methods, such as covariance models and sequence similarity-based methods. As an example, we were able to identify the stem loop structure of the histone mRNA in protozoa, where their primary sequences are very different to the metazoan consensus and thus were previously missed.
- Filtering steps prior to the computationally demanding use of CMs improved on running time.
- Our predictions proved to be reliable since there was also experimental support to them, like in the case of some snRNAs[111-113].

Furthermore, with our computational methods, we reached biologically significant conclusions as to the structure, function and evolution of the proteins and ncRNAs studied in this work:

- We were able to identify protein homology in cases where the primary sequence is poorly conserved. Thus, homology relationships of the RNases P and MRP were identified with the aid of profile-based searches.
- The histone 3' end processing machinery is more ancient than previously anticipated and can be traced to the root of the eukaryotic phylogenetic tree.
- The detailed map of the distribution of the U12-type RNA genes supports an early origin of the minor spliceosome and points to a number of occasions where it was lost during evolution.

Finally, as for our examination of eukaryotic gene order:

- We identified pairs of genes that might be of interest for further studies of transcriptional control.
- We presented eGOB, a eukaryotic Gene Order Browser with information on the order of protein and ncRNA genes of different eukaryotic species, where questions related to evolution of gene organization may be examined. The browser also provides information on pairs of adjacent genes that are evolutionarily conserved.

Acknowledgments

Jag är djupt tacksam till min handledare Tore Samuelsson som har gett mig villkorslöst stöd, som har varit så engagerad i mitt arbete och som har gett mig möjligheten att utvecklas på så många möjliga sätt.

Thanks to my co-supervisor Graham Kemp, who introduced me to the programming world, without it I'd still analyzing the data from five years ago. Thanks for always being available.

Ett särskilt tack till Magnus Alm Rosenblad, vars entusiasm och passion för bioinformatickens världen som helt enkelt får en själv att älska området. Tack för att ni alltid har haft tid för att diskutera vetenskap.

I also want to thank Paul Piccinelli and Sebastian Bartschat, for their contribution in the P/MRP and snRNA projects. It was really nice to have you around to talk about bioinformatics issues, among other things, grazie and danke!

A big thanks to all the people at the department, for all those relaxing chats.

Thanks to CONACyT for giving people worthy opportunities of self improvement.

Muchísimas gracias a todos aquellos amigos que hicieron y hacen de nuestra estancia en este país una muy agradable experiencia, compartiendo desde carnes asadas y molletes hasta viajes y tardes de videojuegos y películas.

Mats y Tere, la razón por la que estamos en este país. Gracias a ambos por todo el cariño y apoyo que nos han dado desde que llegamos. Gran parte de esta tésis es gracias a uds.

A mis papás, no puedo mas que volverles a agradecer el haberme formado como soy, gracias a ustedes estoy donde estoy, los quiero mucho. Y claro, a mi hermana, a quien solo puedo decir: AAAHH! AAAHH! AAAHH!. Gracias también a mis suegros y cuñados, por apoyarnos tanto.

A mi pequeña le agradezco su comprensión en días de mucho trabajo, tqm preciosa.

Y finalmente, gracias por todo tu apoyo y comprensión, por todo el trabajo extra que has invertido, por tus trenes de napalm y por tu "un paso a la vez". Y bueno, tú conoces la letra: *"My dearest friend, if you don't mind …"*

Gott und Leben, noch einmal.

References

1. Lander, E.S., *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Fickett, J.W. Finding genes by computer: the state of the art. *Trends Genet* **12**, 316-320 (1996).
3. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**, 78-94 (1997).
4. Eddy, S.R. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* **2**, 919-929 (2001).
5. Erdmann, V.A., Szymanski, M., Hochberg, A., Groot, N. & Barciszewski, J. Non-coding, mRNA-like RNAs database Y2K. *Nucleic acids research* **28**, 197-200 (2000).
6. Rivas, E. & Eddy, S.R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).
7. Washietl, S., Hofacker, I.L. & Stadler, P.F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**, 2454-2459 (2005).
8. Pedersen, J.S., *et al.* Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* **2**, e33 (2006).
9. Washietl, S., *et al.* Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* **17**, 852-864 (2007).
10. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
11. Lipman, D.J. & Pearson, W.R. Rapid and sensitive protein similarity searches. *Science* **227**, 1435-1441 (1985).
12. Altschul, S.F., *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389-3402 (1997).
13. Eddy, S.R. Profile hidden Markov models. *Bioinformatics (Oxford, England)* **14**, 755-763 (1998).
14. Grillo, G., Licciulli, F., Liuni, S., Sbisa, E. & Pesole, G. PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences. *Nucleic Acids Res* **31**, 3608-3612 (2003).
15. Macke, T.J., *et al.* RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res* **29**, 4724-4735 (2001).
16. Eddy, S.R. & Durbin, R. RNA sequence analysis using covariance models. *Nucleic acids research* **22**, 2079-2088 (1994).
17. Eddy, S.R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC bioinformatics* **3**, 18 (2002).
18. Lai, E.C., Tomancak, P., Williams, R.W. & Rubin, G.M. Computational identification of Drosophila microRNA genes. *Genome biology* **4**, R42 (2003).
19. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955-964 (1997).
20. Lowe, T.M. & Eddy, S.R. A computational screen for methylation guide snoRNAs in yeast. *Science (New York, N.Y* **283**, 1168-1171 (1999).
21. Regalia, M., Rosenblad, M.A. & Samuelsson, T. Prediction of signal recognition particle RNA genes. *Nucleic acids research* **30**, 3368-3377 (2002).
22. Maniatis, T. & Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416**, 499-506 (2002).
23. Bentley, D. Coupling RNA polymerase II transcription with pre-mRNA processing. *Current opinion in cell biology* **11**, 347-351 (1999).
24. Hirose, Y. & Manley, J.L. RNA polymerase II and the integration of nuclear events. *Genes & development* **14**, 1415-1429 (2000).
25. Proudfoot, N. Connecting transcription to messenger RNA processing. *Trends in biochemical sciences* **25**, 290-293 (2000).
26. Shatkin, A.J. & Manley, J.L. The ends of the affair: capping and polyadenylation. *Nat Struct Biol* **7**, 838-842 (2000).
27. Takagaki, Y. & Manley, J.L. Complex protein interactions within the human polyadenylation machinery identify a novel component. *Molecular and cellular biology* **20**, 1515-1525 (2000).
28. Lodish, H., *et al. Molecular Cell Biology*, (W H FREEMAN AND COMPANY, 2000).

29. Frank, D.N. & Pace, N.R. Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annual review of biochemistry* **67**, 153-180 (1998).

30. Jarrous, N. & Reiner, R. Human RNase P: a tRNA-processing enzyme and transcription factor. *Nucleic Acids Res* **35**, 3519-3524 (2007).

31. Xiao, S., Scott, F., Fierke, C.A. & Engelke, D.R. Eukaryotic ribonuclease P: a plurality of ribonucleoprotein enzymes. *Annual review of biochemistry* **71**, 165-189 (2002).

32. Topper, J.N., Bennett, J.L. & Clayton, D.A. A role for RNAase MRP in mitochondrial RNA processing. *Cell* **70**, 16-20 (1992).

33. Forster, A.C. & Altman, S. Similar cage-shaped structures for the RNA components of all ribonuclease P and ribonuclease MRP enzymes. *Cell* **62**, 407-409 (1990).

34. Lygerou, Z., Allmang, C., Tollervey, D. & Seraphin, B. Accurate processing of a eukaryotic precursor ribosomal RNA by ribonuclease MRP in vitro. *Science (New York, N.Y* **272**, 268-270 (1996).

35. Schmitt, M.E. & Clayton, D.A. Nuclear RNase MRP is required for correct processing of pre-5.8S rRNA in Saccharomyces cerevisiae. *Molecular and cellular biology* **13**, 7935-7941 (1993).

36. Martin, A.N. & Li, Y. RNase MRP RNA and human genetic diseases. *Cell Res* **17**, 219-226 (2007).

37. Ridanpaa, M., *et al.* Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell* **104**, 195-203 (2001).

38. Loria, A. & Pan, T. Domain structure of the ribozyme from eubacterial ribonuclease P. *RNA (New York, N.Y* **2**, 551-563 (1996).

39. Frank, D.N., Adamidi, C., Ehringer, M.A., Pitulle, C. & Pace, N.R. Phylogenetic-comparative analysis of the eukaryal ribonuclease P RNA. *RNA (New York, N.Y* **6**, 1895-1904 (2000).

40. Pitulle, C., Garcia-Paris, M., Zamudio, K.R. & Pace, N.R. Comparative structure analysis of vertebrate ribonuclease P RNA. *Nucleic acids research* **26**, 3333-3339 (1998).

41. Pace, N.R. & Brown, J.W. Evolutionary perspective on the structure and function of ribonuclease P, a ribozyme. *J Bacteriol* **177**, 1919-1928 (1995).

42. Jarrous, N., Wolenski, J.S., Wesolowski, D., Lee, C. & Altman, S. Localization in the nucleolus and coiled bodies of protein subunits of the ribonucleoprotein ribonuclease P. *J Cell Biol* **146**, 559-572 (1999).

43. True, H.L. & Celander, D.W. Protein components contribute to active site architecture for eukaryotic ribonuclease P. *The Journal of biological chemistry* **273**, 7193-7196 (1998).

44. Jiang, T. & Altman, S. Protein-protein interactions with subunits of human nuclear RNase P. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 920-925 (2001).

45. Welting, T.J., van Venrooij, W.J. & Pruijn, G.J. Mutual interactions between subunits of the human RNase MRP ribonucleoprotein complex. *Nucleic acids research* **32**, 2138-2146 (2004).

46. Houser-Scott, F., *et al.* Interactions among the protein and RNA subunits of Saccharomyces cerevisiae nuclear RNase P. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 2684-2689 (2002).

47. Salinas, K., Wierzbicki, S., Zhou, L. & Schmitt, M.E. Characterization and purification of Saccharomyces cerevisiae RNase MRP reveals a new unique protein component. *J Biol Chem* **280**, 11352-11360 (2005).

48. Schmitt, M.E. & Clayton, D.A. Characterization of a unique protein component of yeast RNase MRP: an RNA-binding protein with a zinc-cluster domain. *Genes Dev* **8**, 2617-2628 (1994).

49. Chamberlain, J.R., Lee, Y., Lane, W.S. & Engelke, D.R. Purification and characterization of the nuclear RNase P holoenzyme complex reveals extensive subunit overlap with RNase MRP. *Genes & development* **12**, 1678-1690 (1998).

50. Hall, T.A. & Brown, J.W. Archaeal RNase P has multiple protein subunits homologous to eukaryotic nuclear RNase P proteins. *RNA (New York, N.Y* **8**, 296-306 (2002).

51. Terada, A., Honda, T., Fukuhara, H., Hada, K. & Kimura, M. Characterization of the archaeal ribonuclease P proteins from Pyrococcus horikoshii OT3. *J Biochem* **140**, 293-298 (2006).

52. Welting, T.J., Kikkert, B.J., van Venrooij, W.J. & Pruijn, G.J. Differential association of protein subunits with the human RNase MRP and RNase P complexes. *RNA* **12**, 1373-1382 (2006).

53. Nilsen, T.W. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays* **25**, 1147-1149 (2003).

54. Patel, A.A. & Steitz, J.A. Splicing double: insights from the second spliceosome. *Nature reviews* **4**, 960-970 (2003).
55. Madhani, H.D. & Guthrie, C. A novel base-pairing interaction between U2 and U6 snRNAs suggests a mechanism for the catalytic activation of the spliceosome. *Cell* **71**, 803-817 (1992).
56. Frilander, M.J. & Steitz, J.A. Initial recognition of U12-dependent introns requires both U11/5' splice-site and U12/branchpoint interactions. *Genes & development* **13**, 851-863 (1999).
57. Burge, C.B., Padgett, R.A. & Sharp, P.A. Evolutionary fates and origins of U12-type introns. *Molecular cell* **2**, 773-785 (1998).
58. Russell, A.G., Charette, J.M., Spencer, D.F. & Gray, M.W. An early evolutionary origin for the minor spliceosome. *Nature* **443**, 863-866 (2006).
59. Marzluff, W.F., Wagner, E.J. & Duronio, R.J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet* **9**, 843-854 (2008).
60. DeLisle, A.J., Graves, R.A., Marzluff, W.F. & Johnson, L.F. Regulation of histone mRNA production and stability in serum-stimulated mouse 3T6 fibroblasts. *Mol Cell Biol* **3**, 1920-1929 (1983).
61. Dominski, Z., Zheng, L.X., Sanchez, R. & Marzluff, W.F. Stem-loop binding protein facilitates 3'-end formation by stabilizing U7 snRNP binding to histone pre-mRNA. *Molecular and cellular biology* **19**, 3561-3570 (1999).
62. Spycher, C., *et al.* 3' end processing of mouse histone pre-mRNA: evidence for additional base-pairing between U7 snRNA and pre-mRNA. *Nucleic Acids Res* **22**, 4023-4030 (1994).
63. Dominski, Z., Yang, X.C. & Marzluff, W.F. The polyadenylation factor CPSF-73 is involved in histone-pre-mRNA processing. *Cell* **123**, 37-48 (2005).
64. Kolev, N.G. & Steitz, J.A. Symplekin and multiple other polyadenylation factors participate in 3'-end maturation of histone mRNAs. *Genes & development* **19**, 2583-2592 (2005).
65. Pillai, R.S., *et al.* Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes & development* **17**, 2321-2333 (2003).
66. Pillai, R.S., Will, C.L., Luhrmann, R., Schumperli, D. & Muller, B. Purified U7 snRNPs lack the Sm proteins D1 and D2 but contain Lsm10, a new 14 kDa Sm D1-like protein. *The EMBO journal* **20**, 5470-5479 (2001).
67. Dominski, Z., Erkmann, J.A., Yang, X., Sanchez, R. & Marzluff, W.F. A novel zinc finger protein is associated with U7 snRNP and interacts with the stem-loop binding protein in the histone pre-mRNP to stimulate 3'-end processing. *Genes & development* **16**, 58-71 (2002).
68. Boutanaev, A.M., Kalmykova, A.I., Shevelyov, Y.Y. & Nurminsky, D.I. Large clusters of co-expressed genes in the Drosophila genome. *Nature* **420**, 666-669 (2002).
69. Cohen, B.A., Mitra, R.D., Hughes, J.D. & Church, G.M. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nat Genet* **26**, 183-186 (2000).
70. Lercher, M.J., Urrutia, A.O. & Hurst, L.D. Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nat Genet* **31**, 180-183 (2002).
71. Schmid, M., *et al.* A gene expression map of Arabidopsis thaliana development. *Nat Genet* **37**, 501-506 (2005).
72. Kleinjan, D.A. & Lettice, L.A. Long-range gene control and genetic disease. *Adv Genet* **61**, 339-388 (2008).
73. Trinklein, N.D., *et al.* An abundance of bidirectional promoters in the human genome. *Genome Res* **14**, 62-66 (2004).
74. Yang, M.Q., Koehly, L.M. & Elnitski, L.L. Comprehensive annotation of bidirectional promoters identifies co-regulation among breast and ovarian cancer genes. *PLoS Comput Biol* **3**, e72 (2007).
75. Lee, J.M. & Sonnhammer, E.L. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* **13**, 875-882 (2003).
76. Poyatos, J.F. & Hurst, L.D. The determinants of gene order conservation in yeasts. *Genome Biol* **8**, R233 (2007).
77. Adachi, N. & Lieber, M.R. Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109**, 807-809 (2002).
78. Yang, M.Q. & Elnitski, L.L. Prediction-based approaches to characterize bidirectional promoters in the mammalian genome. *BMC Genomics* **9 Suppl 1**, S2 (2008).

79. Yang, M.Q., Taylor, J. & Elnitski, L. Comparative analyses of bidirectional promoters in vertebrates. *BMC Bioinformatics* **9 Suppl 6**, S9 (2008).

80. Li, Y.Y., *et al.* Systematic analysis of head-to-head gene organization: evolutionary conservation and potential biological relevance. *PLoS Comput Biol* **2**, e74 (2006).

81. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988-995 (2004).

82. Gardner, P.P., *et al.* Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**, D136-140 (2009).

83. Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**, D363-368 (2006).

84. Piccinelli, P., Rosenblad, M.A. & Samuelsson, T. Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res* **33**, 4485-4495 (2005).

85. Aravind, L., Iyer, L.M. & Anantharaman, V. The two faces of Alba: the evolutionary connection between proteins participating in chromatin structure and RNA metabolism. *Genome biology* **4**, R64 (2003).

86. Stolc, V., Katz, A. & Altman, S. Rpp2, an essential protein subunit of nuclear RNase P, is required for processing of precursor tRNAs and 35S precursor rRNA in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 6716-6721 (1998).

87. Koonin, E.V., Wolf, Y.I. & Aravind, L. Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. *Genome Res* **11**, 240-252 (2001).

88. Barth, S., *et al.* Elucidating the role of C/D snoRNA in rRNA processing and modification in Trypanosoma brucei. *Eukaryot Cell* **7**, 86-101 (2008).

89. Hartmann, E. & Hartmann, R.K. The enigma of ribonuclease P evolution. *Trends Genet* **19**, 561-569 (2003).

90. Russell, A.G., Shutt, T.E., Watkins, R.F. & Gray, M.W. An ancient spliceosomal intron in the ribosomal protein L7a gene (Rpl7a) of Giardia lamblia. *BMC evolutionary biology* **5**, 45 (2005).

91. Nixon, J.E., *et al.* A spliceosomal intron in Giardia lamblia. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 3701-3705 (2002).

92. Collins, L. & Penny, D. Complex spliceosomal organization ancestral to extant eukaryotes. *Molecular biology and evolution* **22**, 1053-1066 (2005).

93. Misumi, O., *et al.* Cyanidioschyzon merolae genome. A tool for facilitating comparable studies on organelle biogenesis in photosynthetic eukaryotes. *Plant physiology* **137**, 567-585 (2005).

94. Bartschat, S. & Samuelsson, T. U12 type introns were lost at multiple occasions during evolution. *BMC Genomics* **11**, 106 (2010).

95. Dominski, Z. & Marzluff, W.F. Formation of the 3' end of histone mRNA: getting closer to the end. *Gene* **396**, 373-390 (2007).

96. Ausio, J. Histone variants--the structure behind the function. *Briefings in functional genomics & proteomics* **5**, 228-243 (2006).

97. Michel, F., Schumperli, D. & Muller, B. Specificities of Caenorhabditis elegans and human hairpin binding proteins for the first nucleotide in the histone mRNA hairpin loop. *RNA (New York, N.Y* **6**, 1539-1550 (2000).

98. Sullivan, E., *et al.* Drosophila stem loop binding protein coordinates accumulation of mature histone mRNA with cell cycle progression. *Genes & development* **15**, 173-187 (2001).

99. Dominski, Z. & Marzluff, W.F. Formation of the 3' end of histone mRNA. *Gene* **239**, 1-14 (1999).

100. Korbel, J.O., Jensen, L.J., von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nat Biotechnol* **22**, 911-917 (2004).

101. Moreno-Hagelsieb, G. Inferring functional relationships from conservation of gene order. *Methods Mol Biol* **453**, 181-199 (2008).

102. Kaczanowska, M. & Ryden-Aulin, M. Temperature sensitivity caused by mutant release factor 1 is suppressed by mutations that affect 16S rRNA maturation. *J Bacteriol* **186**, 3046-3055 (2004).

103. Kromer, W.J. & Arndt, E. Halobacterial S9 operon. Three ribosomal protein genes are cotranscribed with genes encoding a tRNA(Leu), the enolase, and a putative membrane protein in

the archaebacterium Haloarcula (Halobacterium) marismortui. *J Biol Chem* **266**, 24573-24579 (1991).

104. Hansen, J.J., *et al.* Genomic structure of the human mitochondrial chaperonin genes: HSP60 and HSP10 are localised head to head on chromosome 2 separated by a bidirectional promoter. *Hum Genet* **112**, 71-77 (2003).

105. Albig, W., Trappe, R., Kardalinou, E., Eick, S. & Doenecke, D. The human H2A and H2B histone gene complement. *Biological chemistry* **380**, 7-18 (1999).

106. Burbelo, P.D., Martin, G.R. & Yamada, Y. Alpha 1(IV) and alpha 2(IV) collagen genes are regulated by a bidirectional promoter and a shared enhancer. *Proc Natl Acad Sci U S A* **85**, 9679-9682 (1988).

107. Ame, J.C., *et al.* A bidirectional promoter connects the poly(ADP-ribose) polymerase 2 (PARP-2) gene to the gene for RNase P RNA. structure and expression of the mouse PARP-2 gene. *J Biol Chem* **276**, 11092-11099 (2001).

108. Voth, H., *et al.* Co-regulated expression of HAND2 and DEIN by a bidirectional promoter with asymmetrical activity in neuroblastoma. *BMC Mol Biol* **10**, 28 (2009).

109. Jorge-Nebert, L.F., *et al.* Analysis of human CYP1A1 and CYP1A2 genes and their shared bidirectional promoter in eight world populations. *Hum Mutat* (2009).

110. Huang, C.C. & Chang, W.S. Cooperation between NRF-2 and YY-1 transcription factors is essential for triggering the expression of the PREPL-C2ORF34 bidirectional gene pair. *BMC Mol Biol* **10**, 67 (2009).

111. Chakrabarti, K., *et al.* Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA (New York, N.Y* **13**, 1923-1939 (2007).

112. Davis, C.A., Brown, M.P. & Singh, U. Functional characterization of spliceosomal introns and identification of U2, U4, and U5 snRNAs in the deep-branching eukaryote Entamoeba histolytica. *Eukaryotic cell* **6**, 940-948 (2007).

113. Mitrovich, Q.M. & Guthrie, C. Evolution of small nuclear RNAs in S. cerevisiae, C. albicans, and other hemiascomycetous yeasts. *RNA (New York, N.Y* (2007).