

**The early evolution of eukaryotes  
with special reference to ribosome export factors**

**Hajime Ohyanagi**

**Doctor of Philosophy**

**Department of Genetics**

**School of Life Science**

**The Graduate University for Advanced Studies**

**2008 (School Year)**

## **Acknowledgments**

I wish to express my sincere gratitude to Professor Takashi Gojobori for his continuous guidance and encouragement during all the stages of this work. I thank Professors Nori Kurata, Toshihiko Shiroishi, Takehiko Kobayashi, Masami Hasegawa, Norihiro Okada and Associate Professor Toshiyuki Takano for their useful comments on my work, serving as the members of my supervisory committee. I wish to express my appreciation to Associate Professor Kazuho Ikeo for his valuable advices and discussions. I feel deep gratitude to the board members and my supervisors in Mitsubishi Space Software Co., Ltd., for giving me a great opportunity to spend the time in National Institute of Genetics in my career. I also appreciate the support and encouragement of all of my colleagues and friends in National Institute of Genetics and Mitsubishi Space Software Co., Ltd. Finally I would like to dedicate this thesis to my beloved wife, Yumi.

# Contents

<b>ACKNOWLEDGMENTS</b> .....	<b>2</b>
<b>CONTENTS</b> .....	<b>3</b>
<b>ABSTRACT</b> .....	<b>6</b>
<b>CHAPTER 1</b>	
<b>INTRODUCTION</b> .....	<b>11</b>
1.1. THE ORIGIN OF EUKARYOTIC NUCLEUS .....	11
1.2. RIBOSOME BIOGENESIS AND RIBOSOME EXPORT FACTORS (REFs).....	13
1.3. EUKARYOTIC DATABASES .....	14
<b>CHAPTER 2</b>	
<b>THE EARLY EVOLUTION OF EUKARYOTES REVEALED FROM</b>	
<b>THE EVOLUTIONARY RATES OF RIBOSOME EXPORT FACTORS</b> .....	<b>15</b>
2.1. INTRODUCTION.....	15
2.2. MATERIALS AND METHODS .....	18
2.2.1. Proteome data .....	18
2.2.2. Ortholog detection.....	18
2.2.3. Ribosomal proteins .....	18
2.2.4. Ribosome Export Factor proteins ( <i>REF</i> proteins).....	19
2.2.5. Other proteins .....	19
2.2.6. Estimation of relative evolutionary rates.....	20
2.2.7. Molecular phylogenetic tree construction.....	20
2.3. RESULTS AND DISCUSSION .....	21

<i>2.3.1. Comparison of evolutionary rates among REF proteins, ribosomal proteins, and other proteins</i> .....	21
<i>2.3.2. Comparison of evolutionary rates between non-mREF proteins and mREF proteins</i> .....	22
<i>2.3.3. Molecular phylogenetic trees of REFs</i> .....	23
<i>2.3.4. Conclusion</i> .....	24

## CHAPTER 3

### THE EARLY EVOLUTION OF EUKARYOTES AND

### THE EVOLUTIONARY ORIGIN OF RIBOSOME EXPORT FACTORS ..... 35

3.1. INTRODUCTION.....	35
3.2. MATERIALS AND METHODS .....	37
3.2.1. Protein sequences .....	37
3.2.2. Ribosome Export Factors (REFs).....	37
3.2.3. REF-ortholog detection by reciprocal BLAST best hit method .....	38
3.2.4. REF-ortholog detection by PSI-BLAST .....	38
3.2.5. Spread of homologous genes of the REF-orthologs in archaeal and eubacterial lineages.....	38
3.3. RESULTS AND DISCUSSION .....	39
3.3.1. REF-orthologs detection in archaeal and eubacterial lineages.....	39
3.3.2. Exclusion of horizontally transferred genes and determination of REF origins.	40
3.3.3. Eubacterial origin of nucleus .....	41
3.3.4. Conclusion .....	43

## CHAPTER 4

### CONCLUSION ..... 46

**APPENDIX**

**THE MOLECULAR DATABASE OF *HYDRA* CELLS,**

**THE RICE ANNOTATION PROJECT DATABASE (RAP-DB),**

**AND THE H-INVITATIONAL DATABASE (H-INVDB):**

**EXAMPLES OF BIOLOGICAL DATABASES FOR MODEL EUKARYOTES ..... 49**

APPENDIX 1.....50

MOLECULAR DATABASE OF *HYDRA* CELLS .....50

APPENDIX 2.....51

RAP (RICE ANNOTATION PROJECT) AND RAP-DB (RICE ANNOTATION PROJECT DATABASE) .51

APPENDIX 3.....53

THE H-INVITATIONAL DATABASE (H-INVDB) .....53

APPENDIX 4.....55

DISCUSSION AND FUTURE DIRECTION .....55

**REFERENCES..... 64**

## Abstract

It is believed that primordial eukaryotes were derived from prokaryotes, acquiring nucleus. A number of attempts have been made to reveal the early evolution of eukaryotes, and some hypotheses for the emergence of the early eukaryotes are proposed so far. However, the evolutionary process of early eukaryotes is still a controversial issue and remains one of the biggest questions in current biology. In this study, with the eventual goal toward elucidation of the evolutionary origin and process of early eukaryotes, I conducted molecular evolutionary analyses of transporter proteins of ribosomes between the nucleus and the cytoplasm, called ribosome export factors (REFs).

This thesis consists of four chapters and an appendix. In **Chapter 1**, I described the research background for this study, with particular emphasis on the molecular function of the REFs. The ribosome, one of the largest complexes in eukaryotic cells, is to be exported from the nucleus to the cytoplasm through nuclear pores. As discovered in recent years, the kinetic steps in this nucleocytoplasmic transport pathway are stimulated by the REFs. The REFs would be worth focusing on because they can be considered as one of the components in the eukaryotic core system, translation, and as one of the key genes in the evolutionary process of early eukaryotes for maintaining the mobility of the ribosomes under the existence of nuclear membrane in the then-emerging eukaryotic cells.

In **Chapter 2**, with the aim of revealing the functional significance of the REFs in

the process of eukaryotic evolution, I examined the functional constraints of the entire translation system, the ribosomal proteins and the REF proteins. Estimating the relative evolutionary rates of the yeast REF proteins, I found that, although not as much as the ribosomal proteins, the REF proteins do slowly evolve. More interestingly, the evolutionary rates of the REFs can be classified into two groups. In order to explain this difference in evolutionary rates between the two groups, I considered two subcategories for the REFs, according to the steps in which the REFs are involved. Those two subcategories are non-membranous REFs (non-mREFs) and membranous REFs (mREFs). Interestingly, this categorization was coincided with the evolutionary rate difference: Namely, the rapidly evolving REFs were the non-mREFs while the slowly evolving REFs were the mREFs. These results show that the mREF proteins evolve slower than the non-mREF proteins, suggesting the functional importance of mREFs in the evolutionary process of eukaryotes.

In **Chapter 3**, I examined the evolutionary origin of the eukaryotic nucleus by conducting the ortholog detection analysis of the REFs in prokaryotic lineages. The evolutionary origin of the nucleus is still unclear, although a number of hypotheses have been proposed so far. I searched for the origin of the REFs in archaeal and eubacterial lineages by the method of PSI-BLAST. The results obtained showed that the non-mREFs originated exclusively from eubacterial proteins whereas the mREFs were from both archaeal and eubacterial proteins. Thus, the REFs working inside the nuclear membrane (*i.e.* non-mREFs) are derived only from eubacteria, while alternatively, the REFs shuttling between the nucleus and the cytoplasm (*i.e.* mREFs) are from both

archaea and eubacteria. If we assume that the early nucleus has parsimoniously employed intranuclear proteins as the intranuclear transporters (*i.e.* non-mREFs), these data suggest that the structure of the nucleus may be a descendant of the eubacterial cell. At least, it is suggested that the nucleus arose in a cell that contained chromosomes possessing a substantial fraction of eubacterial genes. Therefore, from the viewpoint of ribosome transport, it is plausible that the nuclear structure is not originated from archaea, but from eubacteria.

Lastly, in **Chapter 4**, I provided a summary and conclusions for the present study. I have shown that the REFs evolve slowly, in addition, the mREFs evolve more slowly, suggesting that the entire eukaryotic translation system is under the functional constraints, and in particular, that the mREFs are functionally important in the process of eukaryotic evolution. Moreover, from the prokaryotic origin of the REFs, it is suggested that the nucleus is rather a descendant of the eubacterial cell, not the archaeal cell.

In **Appendix**, I made particular mention to the biological database projects for eukaryotes, in which I have been involved. Comprehensive annotations of model eukaryotes and integrated databases for such annotations are becoming more and more important in the current post-genome era. Moreover, such databases are useful for the study of early evolution of eukaryotes that is the main aim of the present study. Such databases are also invaluable for comprehensive access to the information resources, and will stimulate the comparative evolutionary genomics. With the eventual goal to know the early evolution of eukaryotes, here I refer to three eukaryotic database projects



in which I have been involved, the Molecular Database of *Hydra* Cells, the Rice Annotation Project Database (RAP-DB), and the H-Invitational Database (H-InvDB).

The Molecular Database of *Hydra* Cells includes the invaluable data of expression patterns of cell type-specific genes in *Hydra*, a member of phylum Cnidaria, which branched more than 500 million years ago from the main stem leading to all bilaterian animals. The database framework was developed by myself, and it serves a unique opportunity for graphically browsing more than 100 cell type-specific genes in *Hydra*. All of the resources can be accessed through <http://hydra.lab.nig.ac.jp/hydra/>.

The RAP-DB is a database for *Oryza sativa* ssp. *Japonica*, one of the model eukaryotes, and has been developed in order to comprehensively house all the annotations produced by the RAP (Rice Annotation Project), which is internationally organized with the aim of providing standardized and highly accurate annotations of the rice genome. The latest version of the RAP-DB contains 31,439 genes validated by cDNAs. The RAP-DB has been also developed by myself, and employed in the analyses within **Chapter 2**. The RAP-DB is available at <http://rapdb.lab.nig.ac.jp/>.

The H-Invitational Database (H-InvDB) was originally developed as an integrated database of the human transcriptome that was based on extensive annotation of large sets of full-length cDNA (FLcDNA) clone. I participated in the Annotation Meeting of Genome Information Integration Project for the further development of the human genome annotations. Now, the database provides annotation for 175,537 human transcripts and 120,558 human mRNAs extracted from the public DNA databank, in addition to 54,978 human FLcDNA, in the latest release, H-InvDB\_4.3. The H-InvDB

is available at <http://www.h-invitational.jp/>.

The three projects in which I have been involved produced comprehensive information for the model eukaryotes. Each database provides a nice implementation for each biological resource and will stimulate the further exploration in the early evolution of eukaryotes.

# Chapter 1

## Introduction

### 1.1. The origin of eukaryotic nucleus

All eukaryotic cells contain compartments that are bounded by biological membranes while prokaryotes possess only simple intracellular compartments, or none at all. Among these compartments, the most crucial structure of eukaryotes is the nucleus, which separates the biologically fundamental events; DNA replication and RNA transcription which take place in the nucleus, and protein synthesis which is done in the cytoplasm. If the traffic of molecules between the nucleus and cytoplasm is exquisitely controlled, this might provide the eukaryotic cell with a unique opportunity of evolutionary diversification of sophisticated molecular interactions and networks. Thus, it seems that evolutionary establishment of a transport system between the nucleus and the cytoplasm was crucial to evolution of eukaryotes.

Various hypotheses for the origin of the nucleus have been proposed (Lake, 1988; Searcy, 1992; Lake and Rivera, 1994; Martin and Muller, 1998; Moreira and Lopez-Garcia, 1998; Vellai et al., 1998; Lopez-Garcia and Moreira, 1999; Horiike et al., 2001; Cavalier-Smith, 2002; Horiike et al., 2002; Cavalier-Smith, 2004; Horiike et al., 2004; Mans et al., 2004; Martin, 2005; Embley and Martin, 2006; Martin and Koonin, 2006), with one of the strongest models suggesting that as a result of the endosymbiosis of an archaeal cell into a host eubacterial cell, the archaeal cell became the eukaryotic nucleus (Lake, 1988; Lake and Rivera, 1994; Moreira and Lopez-Garcia, 1998; Lopez-Garcia and Moreira, 1999; Horiike et al., 2001; Horiike et al., 2002; Horiike et

al., 2004). Because the nuclear membrane must have been a newly acquired device which appeared in the early stage of the eukaryote, besides from being a beneficial wall for the separation of the molecules in the cells, the nuclear membrane would have also been an “obstacle” for intracellular kinetics of the molecules. Hence the kinetic mechanisms beyond the nuclear membrane should have been acquired at the same time or just around the time of the appearance of the nucleus. However, the proposed hypotheses generally put no emphasis on the intracellular transport system, which should be the crucial mechanism in the eukaryotic evolution for keeping the mobility of ribosome. Therefore, from the viewpoint of intracellular kinetics, the origin of eukaryotic nucleus remains an open question.

## **1.2. Ribosome biogenesis and Ribosome Export Factors (REFs)**

From the viewpoint of the intracellular kinetics, one of the molecules most seriously affected by the appearance of the nuclear membrane may have been the ribosome, as it is assembled inside the nucleus, then exported to the cytoplasm (Tschochner and Hurt, 2003). Moreover, it is an essential, abundant, and huge molecule in living cells (Warner, 1999). To produce a mature ribosome, eukaryotic cells must assemble more than 70 ribosomal proteins along with four different ribosomal RNA (rRNA) species inside the nucleus, and then export them to cytoplasm (Tschochner and Hurt, 2003). As elucidated in recent years, this export process cannot occur spontaneously, but requires some factors that are not rRNAs, nor ribosomal proteins (Ho and Johnson, 1999; Ho et al., 2000; Gadal et al., 2001; Johnson et al., 2002; Nissan et al., 2002; Tschochner and Hurt, 2003; Mans et al., 2004). In this study, I termed such factors Ribosome Export Factors (REFs).

In **Chapter 2** and **Chapter 3**, with the purpose of understanding the evolutionary features of REFs in conjunction with an eventual goal toward elucidation of the evolutionary origin and process of early eukaryotes, I estimate the evolutionary rates of the REFs and search the evolutionary origin of the REFs in prokaryotic lineages.

### 1.3. Eukaryotic Databases

Many eukaryotic genome sequencing projects have finished, and it is time to annotate the genomes and utilize the information for evolutionary studies. Hence, comprehensive annotations of model eukaryotes and integrated databases for such annotations are becoming more and more important in the current post-genome era, and will stimulate comparative evolutionary genetics. The Molecular Database of *Hydra* Cells was constructed by myself in order to provide a unique opportunity for graphically browsing more than 100 cell type-specific genes in *Hydra*. The Rice Annotation Project Database (RAP-DB) is a database for one of the model eukaryotes, *Oryza sativa* ssp. *Japonica*, and has been developed in order to comprehensively house all the annotations produced by the RAP (Rice Annotation Project), which is internationally organized with the aim of providing standardized and highly accurate annotations of the rice genome. The RAP-DB was also constructed by myself. The H-Invitational Database (H-InvDB) provides annotation for 175,537 human transcripts and 120,558 human mRNAs extracted from the public DNA databank, in addition to 54,978 human FLCdNA, in the latest release, H-InvDB\_4.3. I participated the Annotation Meeting of Genome Information Integration Project as an annotator, in order to further development the H-InvDB annotations.

In the **Appendix**, I refer to these three eukaryotic database projects in which I have been involved.

## **Chapter 2**

### **The early evolution of eukaryotes revealed from the evolutionary rates of ribosome export factors**

#### **2.1. Introduction**

Eukaryotes are clearly differentiated from prokaryotes, possessing a definite intracellular structure, the nucleus, which is characterized particularly by a nuclear membrane. One of important roles of the nucleus in a eukaryotic cell is separation of biologically fundamental events; DNA replication and RNA transcription take place in the nucleus, whereas protein synthesis is done in the cytoplasm. If the traffic of molecules between the nucleus and the cytoplasm is exquisitely controlled, it might provide a eukaryotic cell with a unique opportunity of evolutionary diversification of sophisticated molecular interactions and networks. Thus, it seems that evolutionary establishment of a transport system between the nucleus and the cytoplasm was crucial to evolution of eukaryotes.

In a eukaryotic cell, more than 70 ribosomal proteins and 4 different rRNAs are to be assembled in the nucleus, and the ribosomal complex is to be exported to the cytoplasm (Tschochner and Hurt, 2003). As elucidated in recent years, this export process requires some adapter molecules that are not rRNAs, nor ribosomal proteins (Ho and Johnson, 1999; Ho et al., 2000; Gadal et al., 2001; Johnson et al., 2002; Nissan et al., 2002; Tschochner and Hurt, 2003; Mans et al., 2004). Such molecules are proteins which stimulate the kinetic steps in this nucleocytoplasmic transport pathway. In this study, I call those proteins Ribosome Export Factors (REFs). To let a ribosome pass

through the nuclear membrane, the REFs should have appeared in the early stage of eukaryotic evolution because the nucleus must have been a newly acquired structure which came out in the epoch of eukaryotes. Thus, it is of particular interest to understand the evolutionary features of REFs. However, even the evolutionary rates of REFs have not been well studied.

Ribosomal molecules that are to make up the tertiary complex are highly expressed, ubiquitously expressed, and highly indispensable. In particular, export of ribosomes, one of the largest complexes in a eukaryotic cell, from the nucleus to the cytoplasm must have been crucial for the translation system of proteins. The evolutionary rates of ribosomal proteins are well known to be very slow due to their functional constraints (Hori et al., 1977). Thus, it is of immediate interest to compare the evolutionary rates of REFs with those of ribosomal proteins.

In the present study, with the purpose of understanding the evolutionary features of REFs in conjunction with an eventual goal toward elucidation of the evolutionary origin and process of early eukaryotes, we estimated the rates of amino acid substitution for REFs and compared them with those of ribosomal proteins. The results obtained suggest that there are two classes of REFs; slowly and rapidly evolving REFs. I found that the slowly evolving REFs correspond to the membranous REFs (mREFs), which are committed to direct transport of proteins from the nucleus to the cytoplasm. From my observations and findings, I concluded that the mREFs might have played an important role when early eukaryotes acquired the nucleus and maintained the translation system by establishing transport pathways through the nuclear membrane between the nucleus



and the cytoplasm. I also discuss the evolutionary significance of the non-mREFs (REFs involved in the non-membranous protein transport from the nucleolus to nucleoplasm) in the process of eukaryotic evolution.

## 2.2. Materials and Methods

### 2.2.1. Proteome data

The list of finished genomes of eukaryotes was obtained from the Entrez Genome Project Database (<http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi>) (Wheeler et al., 2007). The genome projects that require the license agreements to end users were excluded from the list. The genomes that lack any REFs were also excluded from the list (but the lack of Mtr2p was allowed, see **section 2.2.7.**). The final list contains 16 species (Figure 1), and the protein sets were downloaded from the URLs. Among them, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus* were employed to estimate the evolutionary rates of the REFs (see **section 2.2.6.**) because these two species were closely related and well annotated.

### 2.2.2. Ortholog detection

Each protein set of 15 species, excluding *S. cerevisiae* (**Figure 2-1**), was subjected to the BLASTP search (Altschul et al., 1997) against the protein set of *S. cerevisiae*, and the reciprocal BLAST best hit with a bit score >50 was taken as a possible ortholog between the species and *S. cerevisiae*.

### 2.2.3. Ribosomal proteins

Among the 5,602 orthologous proteins detectable between *S. cerevisiae* and *S. paradoxus*, 64 orthologous proteins were taken as the ribosomal proteins in *S. cerevisiae*, according to the gene description in the protein set in the *Saccharomyces* Genome

Database (Nash et al., 2007).

#### 2.2.4. Ribosome Export Factor proteins (REF proteins)

Among the 5,602 orthologous proteins detectable between *S. cerevisiae* and *S. paradoxus*, 8 orthologous proteins (Mtr2p, Nmd3p, Noc1p, Noc2p, Noc3p, Noc4p, Nop14p, and Xpo1p) were taken as the REF proteins, according to the previous report (Tschochner and Hurt, 2003). The accession numbers of these REF protein sequences from *S. cerevisiae* are as follows; **NP 012735** (Mtr2p), **NP 012040** (Nmd3p), **NP 010345** (Noc1p), **NP 014849** (Noc2p), **NP 013102** (Noc3p), **NP 015470** (Noc4p), **NP 010133** (Nop14p), and **NP 011734** (Xpo1p). For the other 15 species (except for *S. cerevisiae*, see Figure 1) the orthologous proteins to the REFs of *S. cerevisiae* were taken as corresponding REFs. The REFs were subcategorized into two subgroups, non-membranous REFs (non-mREF) and membranous REFs (mREFs) according to the previous report (Tschochner and Hurt, 2003) (See **section 2.3.2.**).

#### 2.2.5. Other proteins

Among the 5,602 orthologous proteins detectable between *S. cerevisiae* and *S. paradoxus*, 5,530 proteins that were not ribosomal proteins nor REF proteins were taken as “other proteins” as the control fraction in *S. cerevisiae*. The proteins Q0105, YGR176W, and TDR193W were excluded from the present analysis because the ClustalW could not obtain the pairwise alignments of corresponding orthologous protein pairs between *S. cerevisiae* and *S. paradoxus* (see **section 2.2.6.**). Finally, 5,527 proteins

were taken as other proteins.

#### *2.2.6. Estimation of relative evolutionary rates*

I estimated the number of amino acid substitutions per site for the orthologous protein pairs of *S. cerevisiae* and *S. paradoxus*. Because the divergence time of the two species is still roughly estimated (~5-20 million years) (Kellis et al., 2003), the numbers were directly employed as relative evolutionary rates in this study, following the convention of molecular evolutionary studies. In particular, pairwise alignments were obtained by ClustalW (version 1.83) (Thompson et al., 1994) for each orthologous pair, and the numbers of amino acid substitutions were estimated by Kimura's empirical method (Kimura, 1983), which is implemented in PHYLIP (version 3.66) (Felsenstein, 2005).

#### *2.2.7. Molecular phylogenetic tree construction*

The multiple alignment of 112 REFs from a total of 16 eukaryotes was obtained by ClustalW. Next, the protein distance matrix was estimated by the Jones-Taylor-Thornton model (Jones et al., 1992). The phylogenetic tree was then constructed by the neighbor-joining method (Saitou and Nei, 1987). Finally, the figure of tree was drawn by MEGA3.1 (Kumar et al., 2004). Mtr2p was excluded from this analysis because it is a fungi lineage specific protein.

## 2.3. Results and Discussion

### *2.3.1. Comparison of evolutionary rates among REF proteins, ribosomal proteins, and other proteins*

In order to understand the evolutionary rates of the REFs, the evolutionary distances of orthologous pairs of proteins were estimated as relative evolutionary rates in fungi lineage (see **section 2.2.6**). As shown in **Figure 2-2**, although not as much as the 64 ribosomal proteins, the 8 REFs evolved slowly (although the rate difference between the REFs and other proteins was weakly statistically significant, Mann-Whitney *U* test,  $P = 0.052$ , this might be due to the small number of REFs). Moreover, a few of the REFs (Nmd3p and Xpo1p) exhibited much slower rates than the average rate of the ribosomal proteins. Thus, as for the fungi lineage, it was shown that the REF proteins are under strong functional constraints like ribosomal proteins. My results suggest that the eukaryotic translation system is under strong functional constraints as a whole.

A more interesting point is that the evolutionary rates of REFs seem to be separated into two classes (**Figure 2-2a**). While some REFs evolve at around a half of the grand average of other proteins, the other REFs evolve in a different order of the rate (about  $10^{-1}$ ). If the REFs are divided into two subgroups with the mean evolutionary rate of REFs themselves as the threshold (0.0508, **Figure 2-2**), the rapid group consists of Noc1p, Noc2p, Noc3p, Noc4p, and Nop14p, while the slow group consists of Mtr2p, Nmd3p, and Xpo1p (**Figure 2-2**). The reason for this rate difference is not clear. Thus, I focused on this problem in following sections (see **sections 2.3.2**. and **2.3.3**).

### *2.3.2. Comparison of evolutionary rates between non-mREF proteins and mREF proteins*

In order to understand the evolutionary rate difference of REFs, the two subgroups were employed in the REFs. As already discussed, the eukaryotic ribosome is to be transported from the nucleus to the cytoplasm. In terms of cell biology, the nucleus has an internal structure, the nucleolus, which is known as the venue of the initial ribosome assembly, and is not a membrane-bound structure (Raška et al., 2006) (**Figure 2-3**). Hence, the ribosome is considered to be exported from inside to outside of the nucleus with two completely different steps, non-membranous transport (from the nucleolus to the nucleoplasm) and membranous transport (from the nucleoplasm to the cytoplasm) (**Figure 2-3**). The REFs were then subcategorized into the two subgroups according to these steps (see **section 2.2.4.**), and these two subgroups were named as non-membranous-REFs (non-mREFs) and membranous-REFs (mREFs), respectively (**Figure 2-3**).

Surprisingly, this grouping completely coincides with the rate difference, namely, the rapidly evolving REFs are the non-mREFs, while the slowly evolving REFs are the mREFs (**Figure 2-2**). The rate difference between the non-mREFs and the mREFs was statistically significant (Mann-Whitney  $U$  test,  $P < 0.05$ ). Thus, in fungi lineage, the mREFs, which are involved in the membranous transport, were shown to be under much stronger functional constraints than the non-mREFs, which are involved in the non-membranous transport. This suggests that the evolutionary appearance of mREFs

may be one of crucial factors for ensuring the transport from the nucleus to the cytoplasm when the nucleus was formed in the process of eukaryotic evolution.

Because the grouping according to the eukaryotic nested structure of the nucleolus in the nucleus plausibly explains the evolutionary rate difference of the REFs, this suggests that at least two different factors are included in the functional constraints on the REFs. The non-mREFs conduct the intra-nuclear transport, whereas the mREFs stimulate the membrane-crossing transport. Therefore, one of the factors of the functional constraints might be “intra-nuclear transport” which has a weak affect on the non-mREFs, whereas the other might be “nuclear membrane crossing transport” which has a strong affect on the mREFs.

Mtr2p exhibited an intermediate evolutionary rate (**Figure 2-2**); however, the reason for this ambiguity is not clear. There may be another factor of the functional constraints on Mtr2p. Alternatively, Mtr2p is known to be fungi specific in eukaryote lineages (Kadowaki et al., 1994), implying that Mtr2p is an additional factor after the divergence of fungi, and exhibits the ambiguity of the evolutionary rate. Moreover, because Xpo1p is known to be responsible for multiple kinds of cargo (Maurer et al., 2001), it is highly conserved because of causes asides from the eukaryotic translation system.

### *2.3.3. Molecular phylogenetic trees of REFs*

In the fungi lineage, *i.e.* between *S. cerevisiae* and *S. paradoxus* (the two neighboring species diverged in just ~5-20 million years), the REFs were clearly classified into two classes; the slowly evolving mREFs and the rapidly evolving non-mREFs (see **section**

2.3.2.). With the aim of testing whether the rate difference is just fungi specific, or in a broad range of species, the molecular phylogenetic tree of 112 REF proteins from 16 eukaryotes (4 animals, 3 plants, and 9 fungi) were reconstructed. In a united tree, the 7 REFs were independently clustered with each other (**Figure 2-4**), showing that the common ancestor of these 16 eukaryotes had already gained the entire set of these 7 REFs. Therefore, the total branch length from the common ancestor (roots in each subtree, red points in **Figure 2-4**) to all the OTUs of each subtree could be considered as the index for evolutionary rate. As shown in **Figure 2-4**, in terms of total branch length the mREFs showed much shorter branches than the non-mREFs (Mann-Whitney  $U$  test,  $P < 0.05$ ). Thus, I concluded that the evolutionary rate difference is not only fungi specific, but also in a broad range of eukaryotic species, suggesting that the functional constraints are universal in the long period of eukaryotic evolution, and supporting the coincidence between the grouping and the rate difference. This suggests that the mREFs have been one of crucial factors for maintaining the transport from the nucleus to the cytoplasm in the time scale of eukaryotic evolution.

#### *2.3.4. Conclusion*

The ribosome, which consists of rRNAs and ribosomal proteins, has been one of the most crucial molecules in molecular evolutionary studies for years. Because it is highly conserved, the genetic information of the ribosome (especially that of the rRNAs) has been employed to construct the phylogenies of distantly related species. However, in the present study, with the aim of unveiling the whole figure of the functional



constraints on the eukaryotic translation system, I put particular emphasis not only on the ribosome itself, but also on the REFs, the helpers for the kinetics of the ribosome. Moreover, the REFs should have appeared in the early stage of eukaryotic evolution, to let the ribosome pass through the nuclear structure. Therefore, I believe that the REFs are worth further study as the components in the eukaryotic translation system of proteins, and also as the key genes in the eukaryotic evolution for keeping the mobility of ribosomes.

I first found that the REF proteins are slowly evolving, like the ribosomal proteins (**Figure 2-2**). However, more interestingly, the results suggested that there are two classes of REFs; slowly and rapidly evolving REFs. The two subgroups for the REFs were then employed according to the intranuclear structure (**Figure 2-3**), and the grouping coincided well with the rate difference, *i.e.* the rapidly evolving REFs are the non-mREFs, while the slowly evolving REFs are the mREFs. Further analysis with 112 REF protein sequences from 16 eukaryotic species showed that the rate difference of REFs is not fungi specific, but in a broad range of eukaryotic species, supporting the coincidence between the grouping and the rate difference (**Figure 2-4**).

My results suggest that the mREFs might have played an important role when eukaryotes acquired a nuclear structure. The mREFs might have been ensuring the translation system by establishing a transport pathway through the nuclear membrane between the nucleus and the cytoplasm. Although the epoch of the eukaryotic nuclear membrane is veiled yet (Martin, 2005; Embley and Martin, 2006), the mREFs might be the key factors to spatiotemporally separate the nucleus-specific biological reactions



from the cytoplasm's ones (Martin and Koonin, 2006). It also suggests that the non-mREFs might have been playing a particular role for facilitating the transport pathway inside the nucleus. The lethality of every single null mutant of the REF genes (not only the mREF genes, but also the non-mREF genes) in *S. cerevisiae* (Giaever et al., 2002) supports these suggestions.



These data suggest that the REFs have been under strong functional constraints as a part of the eukaryotic translation system. Among them, rather the mREFs might have been crucial factors for maintaining the eukaryotic translation system in the process of eukaryotic evolution as kinetic facilitators of the ribosome. In other words, the mREFs might have been providing the lifeline beyond the boundary to keep the mobility of the ribosome in the presence of the nuclear membrane.

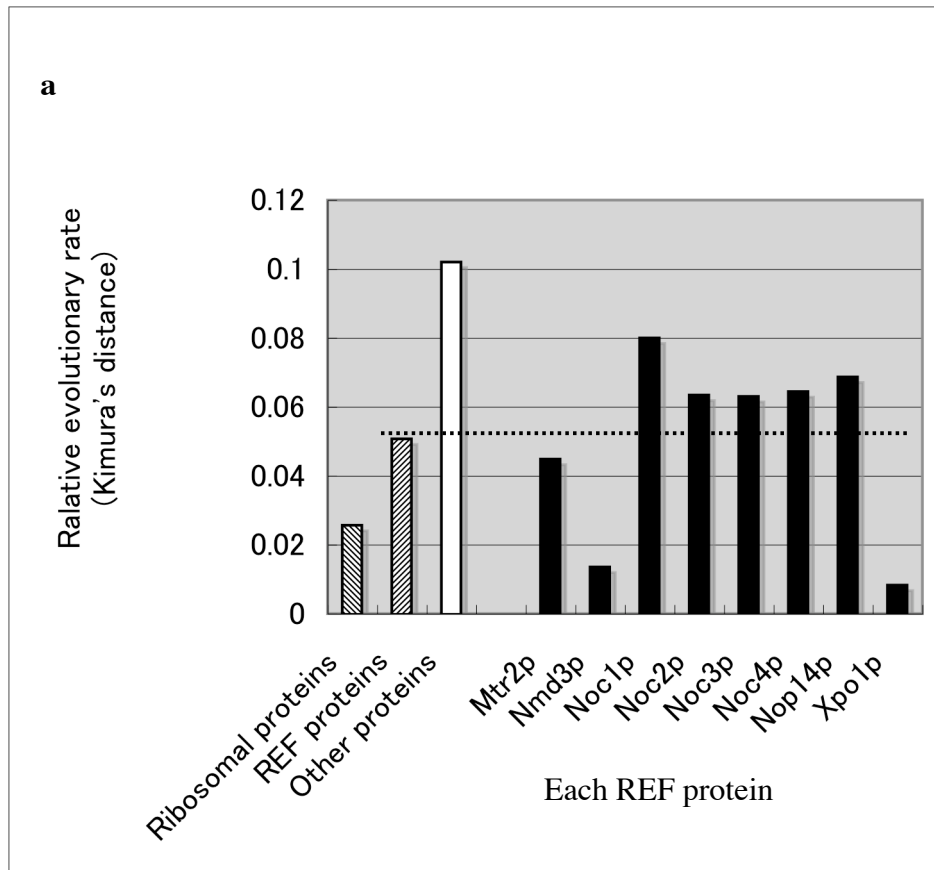
**Figure 2-1.**

GROUP	SPECIES	NUMBER OF PROTEINS	DATA SOURCE	URL	REFERENCE
Plants	<i>Arabidopsis thaliana</i>	26,719	MIPS <i>Arabidopsis thaliana</i> Database	ftp://ftpmips.gsf.de/cress/arabiprot/arabi_all_proteins_v090704.gz	(Schoof et al., 2004)
	<i>Oryza sativa</i> ssp. <i>japonica</i>	28,540	Rice Annotation Project Database	http://rapdb.dna.affrc.go.jp/rapdownload/rap1/rap1_rep.tar.gz	(Ohyanagi et al., 2006; Itoh et al., 2007)
	<i>Ostreococcus lucimarinus</i>	7,651	DOE JGI	ftp://ftp.jgi-psf.org/pub/JGI_data/Ostreococcus_lucimarinus/O.lucimarinus.FM.aafasta.gz	
Animals	<i>Caenorhabditis elegans</i>	22,844	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Caenorhabditis_elegans/CHR_*/*.faa	(Wheeler et al., 2007)
	<i>Drosophila melanogaster</i>	20,058	NCBI	ftp://ftp.ncbi.nih.gov/genomes/Drosophila_melanogaster/CHR_*/*.faa	(Wheeler et al., 2007)
	<i>Homo sapiens</i>	34,180	NCBI	ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/protein/protein.fa.gz	(Wheeler et al., 2007)
	<i>Mus musculus</i>	46,892	NCBI	ftp://ftp.ncbi.nih.gov/genomes/M_musculus/protein/protein.fa.gz	(Wheeler et al., 2007)
Fungi	<i>Candida glabrata</i>	5,215	Génolevures online database	http://cbi.labri.fr/Genolevures/raw/seq/annotation/Release2/Cagl-GL2r2.aa	(Sherman et al., 2006)
	<i>Debaryomyces hansenii</i>	6,319	Génolevures online database	http://cbi.labri.fr/Genolevures/raw/seq/annotation/Release2/Deha-GL2r2.aa	(Sherman et al., 2006)
	<i>Eremothecium gossypii</i>	4,720	Ashbya Genome Database	ftp://ftp.ebi.ac.uk/pub/databases/integr8/fasta/protomes/982.A_gossypii.fasta.gz	(Gattiker et al., 2007)
	<i>Kluyveromyces lactis</i>	5,327	Génolevures online database	http://cbi.labri.fr/Genolevures/raw/seq/annotation/Release2/K11a-GL2r2.aa	(Sherman et al., 2006)
	<i>Pichia stipitis</i>	5,841	DOE JGI	ftp://ftp.jgi-psf.org/pub/JGI_data/Pichia_stipitis/v2.0/FM1.aa.fasta.gz	
	<i>Saccharomyces cerevisiae</i>	6,719	<i>Saccharomyces</i> Genome Database	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/genomic_sequence/orf_protein/orf_trans_all.fasta.gz	(Nash et al., 2007)
	<i>Saccharomyces paradoxus</i>	8,955	<i>Saccharomyces</i> Genome Database	ftp://genome-ftp.stanford.edu/pub/yeast/data_download/sequence/fungal_genomes/S_paradoxus/MIT/orf_protein/orf_trans.fasta.gz	(Nash et al., 2007)
	<i>Schizosaccharomyces pombe</i>	5,004	Sanger Institute	ftp://ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/pompep	
	<i>Yarrowia lipolytica</i>	6,436	Génolevures online database	http://cbi.labri.fr/Genolevures/raw/seq/annotation/Release2/Yali-GL2r2.aa	(Sherman et al., 2006)

**Figure 2-1.** List of eukaryotic proteomes employed in this study. For detail, see **section 2.2.1.**

**Figure 2-2.**  Ribosomal proteins (mean value)  REF proteins (mean value)

 Other proteins (mean value)  REF proteins

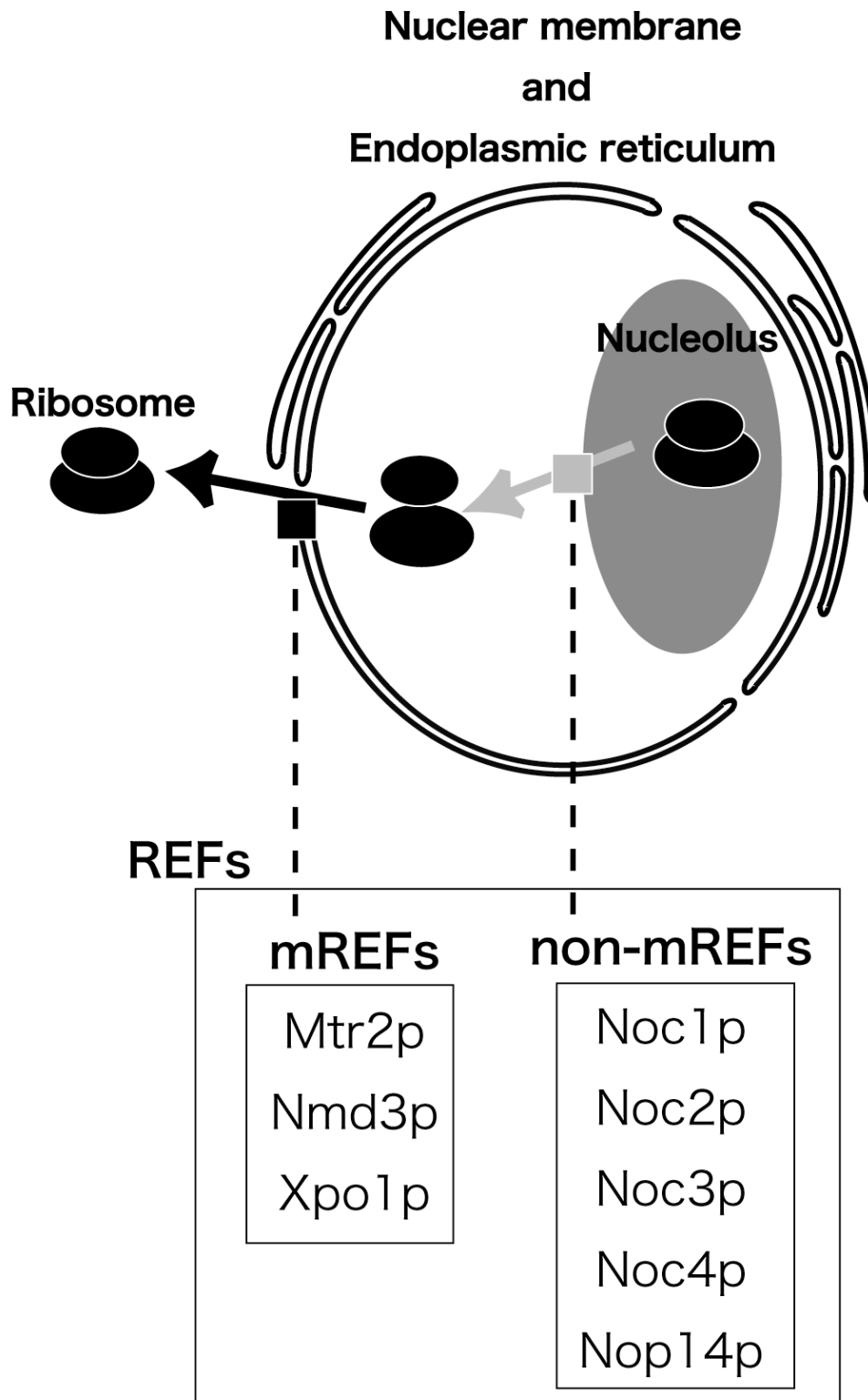


**b**

Name of protein	Relative evolutionary rate (Kimura's distance)
Ribosomal proteins (mean value)	0.0256
REF proteins (mean value)	0.0508
Other proteins (mean value)	0.102
Mtr2p	0.0448
Nmd3p	0.0136
Noc1p	0.0801
Noc2p	0.0635
Noc3p	0.0630
Noc4p	0.0644
Nop14p	0.0686
Xpo1p	0.00835

**Figure 2-2.** Estimated relative evolutionary rates of Ribosome Export Factors (REFs) in yeast lineage (a, between *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*). The striped bars show the arithmetic mean value of the evolutionary rates of 64 ribosomal proteins (see **section 2.2.3.**) and the arithmetic mean value of the evolutionary rates of 8 REF proteins (see **section 2.2.4.**). The open bar shows the arithmetic mean value of the evolutionary rates of other proteins (5,527 proteins, see **section 2.2.5.**). The solid bars show the evolutionary rates of each REF protein. The order of the REFs is alphabetical. The threshold for the subgroups is shown as the horizontal broken line (see **section 2.3.1.**). The numerical values are listed (b).

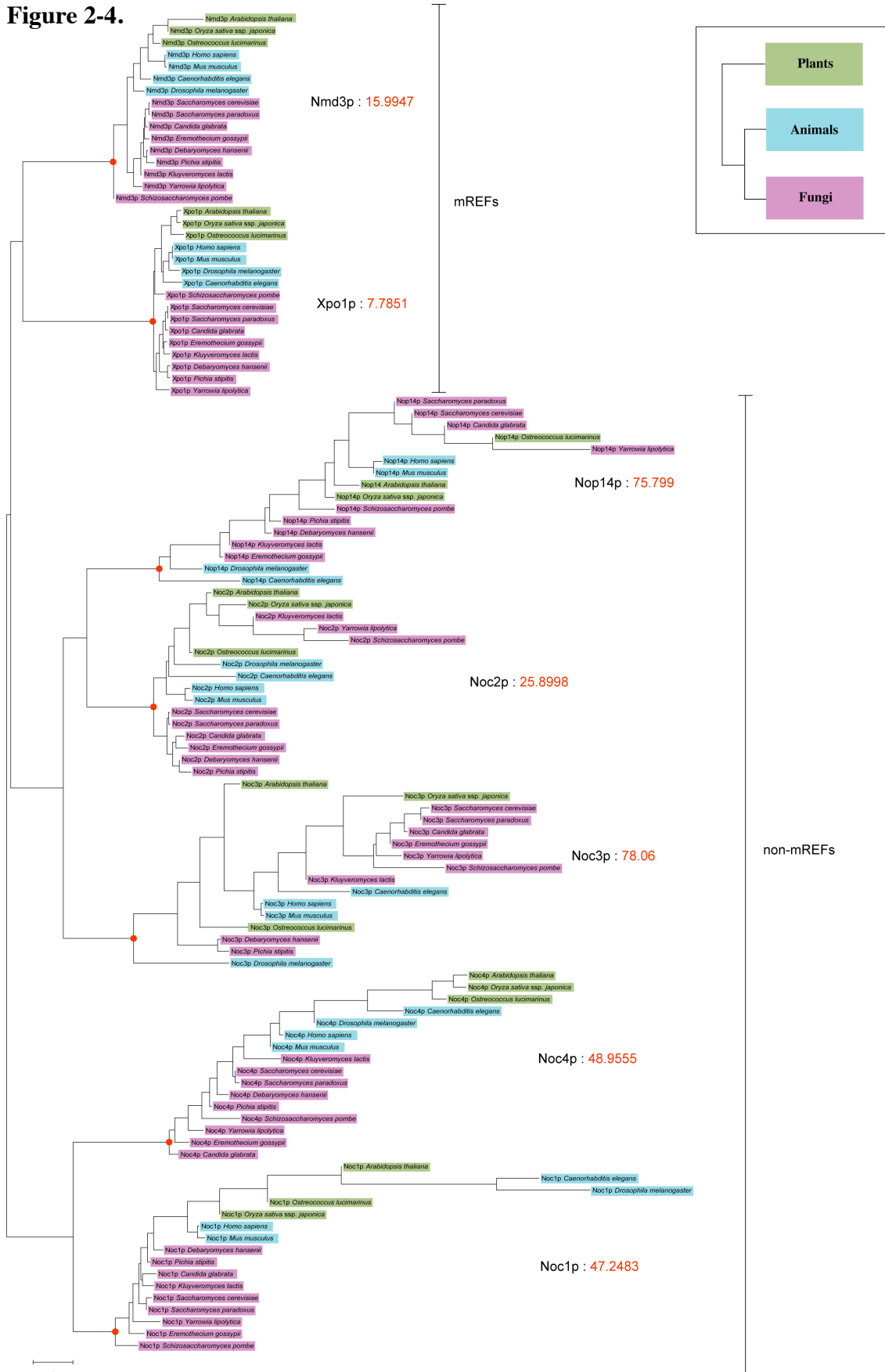
Figure 2-3.



**Figure 2-3.** Eukaryotic ribosome is exported in binary steps. Eukaryotic ribosome (paired solid ovals) is initially assembled in nucleolus, then exported via the nucleoplasm (non-membranous transport, grey arrow) to the cytoplasm (membranous transport, solid arrow). Some REFs are known to stimulate each step (listed gene products in rectangles). The grey oval inside the nuclear membrane stands for nucleolus.



**Figure 2-4.**



**Figure 2-4.** Molecular phylogenetic tree of 112 REFs from 16 eukaryotes. The tree was generated by the neighbor-joining method (see **section 2.2.7.**). Every subtree basically resembles the known species tree (inset) (Hedges, 2002), although some obvious discrepancies were observed. The numbers aside the subtrees are the total branch lengths from the point of common ancestor (red point) to all the OTUs of each subtree. The number below the scale is the number of amino acid substitutions per site.

## **Chapter 3**

### **The early evolution of eukaryotes and the evolutionary origin of ribosome export factors**

#### **3.1. Introduction**

Whereas prokaryotes possess only simple intracellular compartments or none at all, all eukaryotic cells contain compartments that are bounded by biological membranes. The crucial structure of eukaryotes is the nucleus, which encloses the genetic material in the cell. Although various hypotheses have been proposed (Lake, 1988; Searcy, 1992; Lake and Rivera, 1994; Martin and Muller, 1998; Moreira and Lopez-Garcia, 1998; Vellai et al., 1998; Lopez-Garcia and Moreira, 1999; Horiike et al., 2001; Cavalier-Smith, 2002; Horiike et al., 2002; Cavalier-Smith, 2004; Horiike et al., 2004; Mans et al., 2004; Martin, 2005; Embley and Martin, 2006; Martin and Koonin, 2006), any conclusive ideas have not been established yet. Thus, the origin of eukaryotic nucleus remains an open question.

Because the nuclear membrane must have been a newly acquired structure which appeared in the early stage of the eukaryote, being a beneficial wall for the separation of the molecules in the cells, the nuclear membrane would have also been an obstacle for intracellular kinetics of the molecules. Hence the kinetic mechanisms through the nuclear membrane such as the protein transport system should have been acquired at the same time or just around the time as the appearance of the nucleus. One of the molecules most seriously affected by the appearance of the nuclear membrane may have been the ribosome, because it is assembled in a very much complex process through the

nuclear membrane. After the ribosomal proteins are synthesized in the cytoplasm and brought back to the nucleus, the ribosome is to be assembled inside the nucleus, then exported to the cytoplasm again (Tschochner and Hurt, 2003). Moreover, it is a quite essential, abundant, and huge molecule in living cells (Warner, 1999). As already discussed in **Chapter 2**, to produce a mature ribosome which is a quite essential, abundant, and huge molecule in living cells (Warner, 1999), eukaryotic cells must assemble more than 70 ribosomal proteins along with four different rRNA species inside the nucleus, and export them to the cytoplasm (Tschochner and Hurt, 2003). This export process requires some factors that are not rRNAs, nor ribosomal proteins. In **Chapter 2**, I termed such factors Ribosome Export Factors (REFs) (**Figure 2-3**), and revealed their evolutionary account in the process of eukaryotic evolution.

One of the models for the origin of the nucleus suggests that as a result of the endosymbiosis of an archaeal cell into a host eubacterial cell, the archaeal cell became the eukaryotic nucleus. In other words, the nucleus is a descendant of the archaeal cell (Lake, 1988; Lake and Rivera, 1994; Moreira and Lopez-Garcia, 1998; Lopez-Garcia and Moreira, 1999; Horiike et al., 2001; Horiike et al., 2002; Horiike et al., 2004). However, the REFs, one of the components of the eukaryotic translation systems and one of the key family of genes in the eukaryotic evolution for maintaining the mobility of the ribosome, have not been taken into account in the models.

In order to understand the evolutionary origin of the nucleus, I compare the evolutionary histories of the non-mREFs (the REFs working only inside the nucleus) and the mREF (the REFs shuttling between the nucleus and the cytoplasm) (**Figure 2-3**).

If the nucleus is derived from archaea, the non-mREFs (the REFs working only inside the nucleus) are expected to be of archaeal origin. Alternatively, if the nucleus is derived from eubacteria, then a eubacterial origin of the non-mREFs is expected. I searched for the evolutionary origin of the eukaryotic non-mREFs and mREFs in archaeal and eubacterial lineages by the BLAST search.

## **3.2. Materials and Methods**

### *3.2.1. Protein sequences*

The protein set of *Saccharomyces cerevisiae* was downloaded from *Saccharomyces* Genome Database (**Figure 2-1**). The protein sets of all archaea and eubacteria were downloaded from Genome Information Broker in DDBJ (GIB, <http://gib.genes.nig.ac.jp/>) (Fumoto et al., 2002; Sugawara et al., 2007). The status of the GIB dataset was; 40 archaeal genomes and 469 eubacterial genomes (as of May 17, 2007). The non-redundant protein sequences dataset (nr) was downloaded from the NCBI ftp server (<ftp://ftp.ncbi.nih.gov/>) (Wheeler et al., 2007). The status of the nr dataset was; 4,878,246 sequences and 1,686,729,293 total letters (as of April 21, 2007).

### *3.2.2. Ribosome Export Factors (REFs)*

Among the protein set of *S. cerevisiae*, three proteins (Mtr2p, Nmd3p, and Xpo1p) were taken as the mREF proteins, and five proteins (Noc1p, Noc2p, Noc3p, Noc4p, and Nop14p) were taken as the non-mREF proteins, as mentioned in **section 2.2.4**.

### *3.2.3. REF-ortholog detection by reciprocal BLAST best hit method*

Each REF protein was subjected to the BLASTP search (Altschul et al., 1997) against the protein data set of the all archaea and eubacteria (see **section 3.2.1.**), and the reciprocal BLAST best hits was taken as a REF-ortholog.

### *3.2.4. REF-ortholog detection by PSI-BLAST*

The REFs were then subjected to the PSI-BLAST search (Altschul et al., 1997; Altschul and Koonin, 1998; Schäffer et al., 2001) against the protein sequences dataset of the all archaea, eubacteria, and nr datasets (see **section 3.2.1.**). The nr dataset was included in the subject dataset in order to optimize the Position-Specific Scoring Matrix (PSSM) for eukaryotic REFs. The threshold of PSI-BLAST E-value for inclusion in PSSM was set to 0.005 (default value). Then the hits with E-value  $< 10^{-4}$  were considered statistically significant, which were taken as REF-orthologs. The hits in the nr dataset were excluded from the final result. The iterations of PSI-BLAST were continued until the search gave at least one statistically significant hit in the archaea or eubacteria datasets, or the search converged.

### *3.2.5. Spread of homologous genes of the REF-orthologs in archaeal and eubacterial lineages*

In order to take the horizontal gene transfer events into account (see **section 3.3.2.**), I tested whether or not the REF-orthologs were horizontally transferred genes. The spread of homologous proteins of each REF-ortholog was estimated in archaeal and eubacterial lineages. The REF-orthologs were subjected to BLASTP search against the protein

dataset of the all archaea and eubacteria (see **section 3.2.1.**). The low-complexity subsequence filter for query sequence was turned off. The hits with E-value  $< 10^{-4}$  were considered statistically significant, which were taken as REF-homologs. The numbers of archaeal or eubacterial species which had statistically significant BLASTP hits in their genomes were taken as indices for the spread of homologous genes in both lineages.

### **3.3. Results and Discussion**

#### *3.3.1. REF-orthologs detection in archaeal and eubacterial lineages*

In order to detect the quite sensitive homologies between the yeast REFs and the corresponding archaeal or eubacterial orthologs, I employed the double homology detection method by the reciprocal-BLASTP and the PSI-BLAST (see **section 3.2.3., 3.2.4.**). First, the eight *S. cerevisiae* REF proteins (non-mREFs: Noc1p, Noc2p, Noc3p, Noc4p, and Nop14p; mREFs: Mtr2p, Nmd3p, and Xpo1p) were subjected to the reciprocal-BLASTP search against the protein dataset of all archaea and eubacteria (see **section 3.2.3.**). Only Nmd3p had a detectable archaeal ortholog at this first stage (data not shown). Second, the same eight REF proteins were subjected to the PSI-BLAST search against the entire dataset of archaea, eubacteria and nr datasets. The homologous proteins of Noc1p, Noc2p, Noc3p, Nop14p, and Xpo1p were detected with statistical significance in the eubacterial lineage, and the homologous proteins of Nmd3p were detected with statistical significance in the archaeal lineage, all of which were taken as REF-orthologs (**Figure 3-1**, second column). On the other hand, Noc4p and Mtr2p had

no detectable homolog in archaeal nor eubacterial lineages (**Figure 3-1**, second column). If the horizontal gene transfer events of prokaryotes did not take place (see **section 3.3.2.**), my results show that the non-mREFs (Noc1p, Noc2p, Noc3p, and Nop14p) were exclusively originated from eubacterial proteins (**Figure 3-1**, the upper portion), while the mREFs (Nmd3p and Xpo1p) were from both an archaeal protein and a eubacterial protein (**Figure 3-1**, the lower portion). This strongly suggests that the origin of the nucleus is occurred in a cell that harbored a source of eubacterial genes. This is because the REFs working inside the nucleus (*i.e.* non-mREF) were only from eubacteria whereas the REFs shuttling between the nucleus and the cytoplasm (*i.e.* mREFs) were from both archaea and eubacteria.

### *3.3.2. Exclusion of horizontally transferred genes and determination of REF origins*

Although I identified the REF-orthologs (see **section 3.3.1.**), the species phylogeny of the REFs cannot be inferred only from the gene phylogeny of the REFs, because archaea and eubacteria may have horizontally exchanged genetic information (Nakamura et al., 2004). The REF-orthologs have the possibility to be horizontally transferred genes from the opposite domain. Taking the horizontal gene transfer events into account, I examined whether or not each REF-ortholog is the exclave of archaea / eubacteria. As for the non-mREFs (**Figure 3-1**, the upper portion), Noc1p was obviously of eubacterial origin, as the homolog spread was limited only to the eubacterial domain (**Figure 3-1**, fifth column). Noc2p and Noc3p are considered possibly of eubacterial origin, as the spread of homolog was mainly, but not exclusively,



in eubacteria (**Figure 3-1**, fifth column), which is supported by the fact that the original PSI-BLAST hits were in eubacterial lineage (**Figure 3-1**, third column). Nop14p is also considered of eubacterial origin, as the spread of homolog of NP\_693412 was mainly in eubacteria (**Figure 3-1**, fifth column), and the original PSI-BLAST hit was in eubacterial lineage (**Figure 3-1**, third column). Another hit (YP\_001208967) was an ultra-conservative gene, “translation initiation factor IF-2” (**Figure 3-1**, fourth column) among all the prokaryotic species (100% archaea and 100% eubacteria, **Figure 3-1**, fifth column), providing no information for the decision of horizontal gene transfer event, but the original PSI-BLAST hit was in eubacterial lineage. As for the mREFs (**Figure 3-1**, the lower portion), Nmd3p was obviously of archaeal origin, as the homolog spread was limited only to the archaeal domain (**Figure 3-1**, fifth column). The reciprocal-BLASTP search also showed the archaeal origin of Nmd3p (data not shown). Xpo1p was obviously of eubacterial origin, as the homolog spread was limited only to the eubacterial domain (**Figure 3-1**, fifth column). Overall, none of the REF-orthologs have horizontally transferred between domains, hence the non-mREFs (Noc1p, Noc2p, Noc3 and Nop14p) were suggested to be exclusively from eubacterial proteins, while the mREFs were suggested to be from both an archaeal protein (for Nmd3p) and a eubacterial protein (for Xpo1p).

### *3.3.3. Eubacterial origin of nucleus*

From **Figure 3-1**, I concluded that the non-mREFs (Noc1p, Noc2p, Noc3, and Nop14p) originated exclusively from eubacterial proteins, while the mREFs (Nmd3p

and Xpo1p) were from both archaeal and eubacterial proteins. In other words, the REFs working inside the nucleus (*i.e.* non-mREFs) were only from eubacteria, whereas the REFs shuttling between the nucleus and the cytoplasm (*i.e.* mREFs) were from both archaea and eubacteria. If I assume that the early nucleus has parsimoniously employed intranuclear proteins as the intranuclear transporters (*i.e.* non-mREFs), my results imply that the nucleus arose *de novo* around chimeric chromosomes containing eubacterial genes. If I assume that those chromosomes stem from an archaeal host cell, as informational genes would suggest, then these findings could be taken as support for theories that do not explain the origin of eukaryotic nucleus through an archaeobacterial endosymbiont, but posit a eubacterial endosymbiont in an archaeal host instead (Searcy, 1992; Martin and Muller, 1998; Vellai et al., 1998). From the standpoint of the hydrogen hypothesis, each REF could stem from archaeal and eubacterial proteomes in a random manner. If the mitochondrial endosymbiont repeatedly donated genes to its archaeal host, as some models suggest (Martin and Koonin, 2006), then the predominance of eubacterial genes in REFs would not be surprising. While no strong argument in favor of any particular theory can currently be drawn from the present data, it is clear that the data are not compatible with the predictions of theories suggesting the origin of nucleus from an archaeal endosymbiont (Lake, 1988; Lake and Rivera, 1994; Moreira and Lopez-Garcia, 1998; Lopez-Garcia and Moreira, 1999; Horiike et al., 2001; Horiike et al., 2002; Horiike et al., 2004).

#### *3.3.4. Conclusion*

In searching for the evolutionary origin of the nucleus, I put particular emphasis on the REFs. I found a surprisingly strong signal linking the evolutionary origin of the nucleus to the existence of eubacterial genes in the eukaryotic lineage. I also know the particular eukaryotic cellular features that have not explained in my hypothesis yet; the existence of mitochondria, the single-bounded nuclear membrane, the nuclear pore complex, the linear chromosomes, the RNA-world relics, the splicing processes, and so on. However my hypothesis looks sound enough from the viewpoint of intracellular transport, while I also know that it is simply based on the intracellular kinetics of the ribosome. In conclusion, my results showed that the non-mREFs originated exclusively from eubacterial proteins, whereas the mREFs were from both archaeal and eubacterial proteins. These data suggest that the nucleus might be more readily understood as a descendant of eubacterial genes than as a descendant of an archaeal cell. Further precise annotations of comprehensive proteins, especially of the intracellular transporter proteins, on the available eukaryotic genome, will help the study of the early evolution of eukaryotes.

Figure 3-1

REFs	Accession numbers of PSI-BLAST hits (E-value)	Species name of the PSI-BLAST hit	Annotation of the PSI-BLAST hit	Homolog spread of the PSI-BLAST hit (in 40 archaea / in 469 eubacteria)	Decision	
non-mREF	YP_165878 (1e-09)	<i>Silicibacter pomeroyi</i> DSS-3 (Eubacteria)	sterol desaturase family protein	0 archaea (0%) / 48 eubacteria (10.2%)	Eubacteria	
	YP_459411 (2e-09)	<i>Erythrobacter litoralis</i> HFOCC2594 (Eubacteria)	sterol desaturase family protein	0 archaea (0%) / 55 eubacteria (11.7%)		
	YP_798571 (8e-09)	<i>Leptospira borgpetersenii</i> serovar Hardjo-ovis L550 (Eubacteria)	Sterol desaturase	0 archaea (0%) / 56 eubacteria (11.9%)		
	YP_619172 (2e-08)	<i>Spingopyxis alaskensis</i> RB2256 (Eubacteria)	Sterol desaturase	0 archaea (0%) / 55 eubacteria (11.7%)		
	NP_682707 (4e-08)	<i>Thermosynechococcus elongatus</i> BP-1 (Eubacteria)	sterol desaturase family protein	0 archaea (0%) / 63 eubacteria (13.4%)		
	YP_495518 (1e-07)	<i>Novosphingobium aromaticivorans</i> DSM 12444 (Eubacteria)	Sterol desaturase	0 archaea (0%) / 59 eubacteria (12.6%)		
	NP_420481 (2e-07)	<i>Caulobacter crescentus</i> CB15 (Eubacteria)	sterol desaturase family protein	0 archaea (0%) / 35 eubacteria (7.5%)		
	YP_000990 (5e-07)	<i>Leptospira interrogans</i> serovar Copenhageni str. Fiocruz LI-130 (Eubacteria)	sterol desaturase	0 archaea (0%) / 39 eubacteria (8.3%)		
	NP_713258 (8e-07)	<i>Leptospira interrogans</i> serovar Lai str. 56601 (Eubacteria)	sterol desaturase family protein	0 archaea (0%) / 40 eubacteria (8.5%)		
	NP_171559 (2e-06)	<i>Synechococcus elongatus</i> PCC 6301 (Eubacteria)	similar to sterol C5-desaturase	0 archaea (0%) / 30 eubacteria (6.4%)		
	YP_169036 (3e-06)	<i>Silicibacter pomeroyi</i> DSS-3 (Eubacteria)	sterol desaturase family protein	0 archaea (0%) / 43 eubacteria (9.2%)		
	NP_979308 (3e-07)	<i>Bacillus cereus</i> ATCC 10987 (Eubacteria)	hypothetical protein	2 archaea (5.0%) / 65 eubacteria (13.9%)		
	Noc2p					Eubacteria
	Noc3p	YP_944121 (1e-10)	<i>Psychromonas ingrahamii</i> 37 (Eubacteria)	protein containing tetratricopeptide (TPR) repeat		2 archaea (5.0%) / 121 eubacteria (25.8%)
Noc4p	no hit found				not determined	
Nop14p	NP_693412 (1e-18)	<i>Oceanobacillus ihayensis</i> HTE831 (Eubacteria)	hypothetical protein	6 archaea (15%) / 385 eubacteria (82.1%)	Eubacteria	
	YP_001208967 (1e-05)	<i>Dichelobacter nodosus</i> VCS1703A (Eubacteria)	translation initiation factor IF-2	40 archaea (100%) / 469 eubacteria (100%)		
mREF	Ntx2p	no hit found			not determined	
	Nmd3p	NP_343741 (5e-08)	<i>Sulfolobus solfataricus</i> F2 (Archaea)	hypothetical protein	34 archaea (85.0%) / 0 eubacteria (0%)	Archaea
		NP_110997 (2e-06)	<i>Thermoplasma volcanium</i> G951 (Archaea)	NMD protein (ribosome stability and mRNA decay protein)	31 archaea (77.5%) / 0 eubacteria (0%)	
		NP_148547 (4e-06)	<i>Aeropyrum pernix</i> K1 (Archaea)	hypothetical protein	11 archaea (27.5%) / 0 eubacteria (0%)	
		NP_559451 (2e-05)	<i>Pyrobaculum aerophilum</i> str. IM2 (Archaea)	hypothetical protein	38 archaea (95.0%) / 0 eubacteria (0%)	
		YP_001055470 (3e-05)	<i>Pyrobaculum calidifontis</i> JCM 11548 (Archaea)	NMD3	38 archaea (95.0%) / 0 eubacteria (0%)	
		YP_919866 (5e-05)	<i>Thermofilum pendens</i> Hrk 5 (Archaea)	NMD3	36 archaea (90.0%) / 0 eubacteria (0%)	
	Xpo1p	YP_040849 (3e-06)	<i>Staphylococcus aureus</i> subsp. aureus MRSA252 (Eubacteria)	very large surface anchored protein	0 archaea (0%) / 27 eubacteria (5.8%)	Eubacteria
		YP_186319 (7e-05)	<i>Staphylococcus aureus</i> subsp. aureus COL (Eubacteria)	Cell wall associated fibronectin-binding protein	0 archaea (0%) / 26 eubacteria (5.5%)	
		YP_499969 (9e-05)	<i>Staphylococcus aureus</i> subsp. aureus NCTC 8325 (Eubacteria)	hypothetical protein	0 archaea (0%) / 28 eubacteria (6.0%)	

**Figure 3-1.** Ortholog detection analysis. Five non-mREFs and three mREFs were subjected to the double homology detection method by the reciprocal BLASTP and the PSI-BLAST (see **section 3.2.3., 3.2.4.**). The result of reciprocal BLASTP search is not shown. All results are by the PSI-BLAST search. Blue cells are of eubacterial origin, and red cells of archaeal origin.

## Chapter 4

### Conclusion

Because the nucleus is the cellular structure essentially differentiating eukaryotes from prokaryotes, the formation of the nuclear membrane should have been a key event in eukaryotic evolution. Although an evolutionary cause of formation of the nucleus such as symbiosis has been speculated but still not clarified, evolutionary establishment of the transport system across the nuclear membrane must have been prerequisite for survival of the then-emerging eukaryotes. In particular, export of ribosomes, one of the largest complexes in a eukaryotic cell, from the nucleus to the cytoplasm must have been crucial for retaining the translation system of proteins. As discovered in recent years, kinetic steps in this nucleocytoplasmic transport pathway are stimulated by proteins called Ribosome Export Factors (REFs). Hence, evolution of REFs is of particular interest, and description of the evolutionary features of REFs is of immediate value. However, even the evolutionary rates of REFs are still poorly understood compared with those of ribosomal components. With the aim of understanding the evolutionary features of REFs, I estimated the rates of amino acid substitutions of REFs for two related species of yeast, *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*, and compared them with those of ribosomal components. I found that the average rate of amino acid substitutions for REFs was somewhat higher than that of ribosomal components, although the former was much lower than the grand average of other 5,527 proteins. Moreover, I also found that the REFs were clearly classified into

two classes; the slowly evolving REFs and the rapidly evolving REFs. Interestingly, I found that the slowly and rapidly evolving REFs correspond to the membranous-REFs (mREFs) and the non-membranous REFs (non-mREFs), respectively. Further analyses with 112 REFs from 16 eukaryotic species also showed clear differences in the evolutionary rates between these two classes. Because the non-mREFs are involved in non-membranous transport from the nucleolus to nucleoplasm and the mREFs are committed to membranous transport from the nucleoplasm to the cytoplasm, the mREFs appear to have much stronger functional constraints than the non-mREFs. Thus, I conclude that the evolutionary appearance of mREFs may be one of the crucial factors for ensuring the transport from the nucleus to the cytoplasm when the nucleus was formed in the process of eukaryotic evolution (Ohyanagi et al., 2008a).

Various hypotheses have been proposed on the evolutionary origin of eukaryotic nucleus. One of the strongest suggests that as a result of the endosymbiosis of an archaeal cell into a host eubacterial cell, the archaeal cell became the eukaryotic nucleus, suggesting that the nucleus is a descendant of the archaeal cell. Because one of the major cargoes in the nucleocytoplasmic export in the eukaryotic cell is the ribosome, its stimulating proteins, REFs, might have an evolutionary history of inscribing the origin of eukaryotic nucleus. With the aim of understanding the evolutionary origin of the nucleus, I employed the yeast REFs and searched for their evolutionary origin in more than 500 genomes of archaea and eubacteria by the PSI-BLAST search. My results showed that the non-mREFs originated exclusively from eubacterial proteins, whereas the mREFs are from both archaeal and eubacterial proteins. Since the non-mREFs just

work inside the nucleus while the mREFs shuttle between the nucleus and the cytoplasm, these results suggest that the extant REFs working inside the nucleus have derived exclusively from eubacterial proteins, implying that the nucleus arose in a cell that contained chromosomes possessing a substantial fraction of eubacterial genes, in line with the predictions of several models entailing endosymbiosis at eukaryote origins. In other words, the structure of nucleus might be rather a descendant of eubacterial cell (Ohyanagi et al., 2008b).



**Appendix**  
**The Molecular Database of *Hydra* Cells,**  
**The Rice Annotation Project Database (RAP-DB),**  
**and The H-Invitational Database (H-InvDB):**  
**Examples of biological databases for model eukaryotes**

In **Chapter 2** and **Chapter 3**, with an eventual goal toward elucidation of the evolutionary origin and process of early eukaryotes, I revealed the rate of amino acid substitution for the REFs and the prokaryotic origin of the REFs. These results provide interesting clues to the evolution of eukaryotic cells, but further analyses with comprehensive information of a greater number of eukaryotic species will stimulate the studies on eukaryotic evolution. In this post-genome sequencing era, it is of immediate value to produce large annotations on the genomes, and to construct biological databases for the genomes and annotations.

With the aim of facilitating the studies on eukaryotic evolution, I took part in the annotation projects and database projects. In this **Appendix**, I summarize my activities on these projects.

## **Appendix 1.**

### **Molecular Database of *Hydra* Cells**

Cell lineages of cnidarians including *Hydra* represent the fundamental cell types of metazoans and provide important insights into the evolution of cell diversification in the animal kingdom. *Hydra* contains a multipotent interstitial cell (I-cell) that gives rise to nematocytes, nerve cells, gland cells and germ cells. This I-cell lineage is not essential for survival, since animals lacking the I-cell lineage can be maintained indefinitely in culture. Such I-cell free *hydra* are referred to as epithelial *hydra*. In the present study, a 6.6 thousand cDNA microarray was constructed, and competitive hybridization by using probes from epithelial *Hydra* and normal *Hydra* was performed to compare gene expression in epithelial *hydra* with normal *hydra*, and thus, to identify genes specific for the I-cell lineage. 151 genes were identified, which were differentially expressed in normal *hydra* and not in I-cell free animals. In situ hybridization showed that 86 of these genes were expressed in specific cell types of the I-cell lineage. An additional 29 genes were expressed in epithelial cells and were down-regulated in epithelial animals lacking I-cells. Based on the above information, I have constructed a database (<http://hydra.lab.nig.ac.jp/hydra/>) which describes the expression patterns of cell type specific genes in *hydra* (Hwang et al., 2007). Currently this database contains more than 100 cell type specific genes and their sequences, UniGENE identities, homologue information, gene ontology, and whole mount in situ hybridization images (**Appendix Figure 1**). All of the resources can be accessed through <http://hydra.lab.nig.ac.jp/hydra/>.

## **Appendix 2.**

### **RAP (Rice Annotation Project) and RAP-DB (Rice Annotation project Database)**

Rice is considered a model cereal plant because of its small genome size and high degree of chromosomal co-linearity with other major cereal crops such as maize, wheat, barley, and sorghum (Moore et al., 1995; Sasaki and Burr, 2000). The International Rice Genome Sequencing Project (IRGSP), a consortium of publicly-funded laboratories from 10 countries, initiated the sequencing of *Oryza sativa* ssp. *japonica* cultivar Nipponbare in 1998 using the clone-by-clone sequencing strategy (Sasaki and Burr, 2000). In 2004, the finished-quality sequence of the entire genome was completed and is now available in the public domain (Matsumoto, 2005).

The annotation of the sequence is indispensable in understanding the overall structure and function of the rice genome. However, most of the annotations of the rice genome sequences were obtained by automated methods. Although this provides an overview of the composition of the genes that comprise the genome, limitations in prediction programs often result in probable errors and artifacts among predicted genes. Therefore, in concordance with the completion of the rice genome sequence, the Rice Annotation Project (RAP) was organized in 2004 (Itoh et al., 2007) with the aim of providing standardized and highly accurate annotations of the rice genome.

To facilitate efficient management of the results of annotation and to establish a platform for integrating the data with other rice resources, I developed an annotation database called the Rice Annotation Project Database (RAP-DB) as first version

(Ohyanagi et al., 2006; Tanaka et al., 2008). The RAP-DB integrates the IRGSP genome sequence and the RAP annotations with other data on rice researches. The latest version of the RAP-DB contains a variety of annotation data as follows: clone positions, structures, and functions of 31,439 genes validated by cDNAs, RNA genes detected by massive parallel signature sequencing (MPSS) technology and sequence similarity, flanking sequences of mutant lines, transposable element, etc. The RAP-DB is available at <http://rapdb.lab.nig.ac.jp/> (**Appendix Figure 2**).

### **Appendix 3.**

#### **The H-Invitational Database (H-InvDB)**

Human transcripts represent very useful resources for examining the structure of human genes and alternative splicing isoforms. In particular, cloning and sequencing of full-length cDNAs (FLcDNAs) that cover all exons but no introns can facilitate the precise determination of human gene structure (Ota et al., 1997). Studies on human transcripts have thus been systematically and extensively carried out to draw the outline of the human transcriptome (Hu, 2000; Wiemann et al., 2001; Yodate, 2001; Kikuno et al., 2002; Strausberg et al., 2002). The human transcriptome consists of protein-coding mRNAs and non-coding functional RNAs. Analysis of those sequences will provide insights into how genomic information is transformed into higher-order biological phenomena. By comparative analysis of the transcriptome with the human genome, I will be able to determine the transcribed regions of the genome and to know the regulatory machinery of transcription. It is therefore of great significance to collect information about human transcripts as well as their annotations. Thus, the first international workshop entitled "Human Full-length cDNA Annotation Invitational" (abbreviated as H-Invitational or H-Inv) was held in Tokyo, Japan from August 25<sup>th</sup> to September 3<sup>rd</sup>, 2002, and constructed a novel, integrative database of the human transcriptome, called H-InvDB (Imanishi et al., 2004; Yamasaki et al., 2005; Yamasaki et al., 2008). This database consists of the annotation of 42,421 human FLcDNAs, collected from six high-throughput producers of human FLcDNAs in the world human gene collections.

To cover the increased number of human FLcDNAs since the initial release of H-InvDB, the second international annotation meeting entitled “H-Invitational 2 Functional Annotation Jamboree” (abbreviated as H-Invitational 2 or H-Inv2) was held in Tokyo, Japan from November 15<sup>th</sup> to 20<sup>th</sup>, 2003. The second major release of H-InvDB was released as release 2.0 based on the annotation at the H-Inv2 annotation jamboree. This consists of the annotation of 56,419 human FLcDNAs, collected from the same six high-throughput producers of human FLcDNAs as in the H-Inv1. After H-Inv2, the Genome Information Integration Project (GIIP) was started for further development and held the third and fourth annotation meeting in Oct., 2005 and Oct., 2006. I participated in the meeting as an annotator to contribute the further development of human genome annotation. The products of those two annotation meetings comprise release 3.0 and 4.0 of H-InvDB. The increase in the entries of H-InvDB are summarized in **Appendix Table 1**. The release H-InvDB\_4.3, provides annotation for 175,537 human transcripts and 120,558 human mRNAs extracted from the public DNA databank, in addition to 54,978 human FLcDNA. The H-InvDB is available at <http://www.h-invitational.jp/> (**Appendix Figure 3**).

## **Appendix 4.**

### **Discussion and Future Direction**

The three eukaryotic database projects in which I have been involved provide comprehensive information for the model eukaryotes. The molecular Database of *Hydra* cell is not a large-scale database, but serves as a unique opportunity for the cell type-specific gene expression. In conjunction with the comparative genomic methods, it will facilitate the evolutionary study of the nervous system. The RAP-DB is now one of the most major databases in plant biology, updating the database contents frequently (Tanaka, T., personal communication). It will stimulate the evolutionary studies of plants, as a reference dataset for cereal crops. The H-InvDB is now a quite integrated database, and one of the most important biological databases for human genome-related resource, accelerating the evolutionary studies of humans and mammals.

Each database provides a nice implementation for each biological resource. However, one of the problems of the current biological database issue is a lack of coordination and integration among published databases. They were not developed in shared framework. Each of them has original data structure. Hence they are potentially isolated resources. It would be of immediate value to take the database connection systems (for example, Distributed Annotation System, DAS) into consideration for future development of the biological resources (Dowell et al., 2001).

Appendix Figure 1.

Molecular Database of Hydra Cells

http://hydra.lab.nig.ac.jp/hydra/

Laboratory for DNA Data Analysis (DDA)  
Center for Information Biology and DNA Data Bank of Japan  
National Institute of Genetics, Mishima, Japan

Molecular Database of Hydra Cells

Main Nematocyte Nerve cell Stem cell and Germ line cell Gland cell Ecto. epi. cell Endo. epi. cell

AAATATACAACAACAACCTCCCTTTGATAAATGATGGAAAAGTTGCAACTACCCAAGC  
TATTCCAATAGAACTACTACATTGAAAAATAA

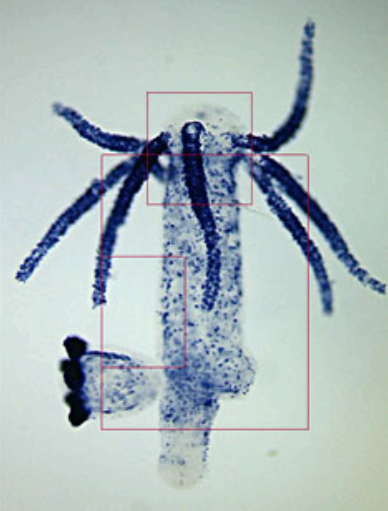
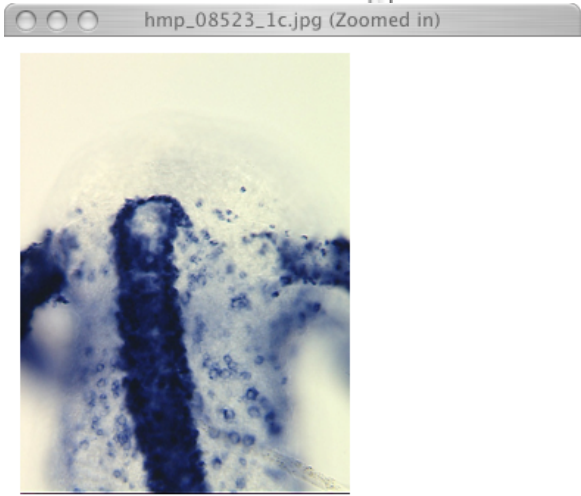
Amino acid sequence (deduced from full length or longer cDNA)

```
MSVPSPLSHTHRQLEKQNI AQLTMKFKEYTDQVRQMRDHCKKTENN VFLSHIAALDNEIK
DLQSIYERELESVRGQLDACIAERNQLHLDASKYGALSKELQDKYNDEKTRTKLENALA
DAHRVLSEKDLLIKELRISIAQHNAHLDTAKERDELQSTLTLTRVTCEGETKMRLDLEA
YVQKLTEQINFERDIHEKDIIDLNRNAAAERTIEIADQKLRHDLVDEKLQOQIENIKR
QTTYDFVQYQEASENSYQLQLEHKNRMAKETQALAQKKEENIHLKAIIEEMNAKIYKLD
GKVSSYHEQNTILIHTEVERRAAAATCHELEKKLQELQEHYNTKVRELNIVSSVNIPID
LELESLSQLIEAEAKRLDVALSNPSSSELVSTVRGELVSNRSHYVHNASPRKASSNAPLKR
QKSPAALVDTTTELPPLAIPKYTTTNLPLINDGKVRNYPYSYNNRYIEK
```

Result of homology search (Blastp)	Description	Lamin C CG10119-PA [Drosophila melanogaster]
	Accession number of the homolog	<a href="#">NP_523742.2</a>
	E-value	2.00E-21
Gene Ontology	Molecular function	-
	Biological process	<a href="#">GO:0007084 "mitotic nuclear envelope reassembly"</a>
	Cellular component	<a href="#">GO:0005638 "lamin filament"</a>

Expression pattern (whole body)

Click on the red box to view the enlarged region.

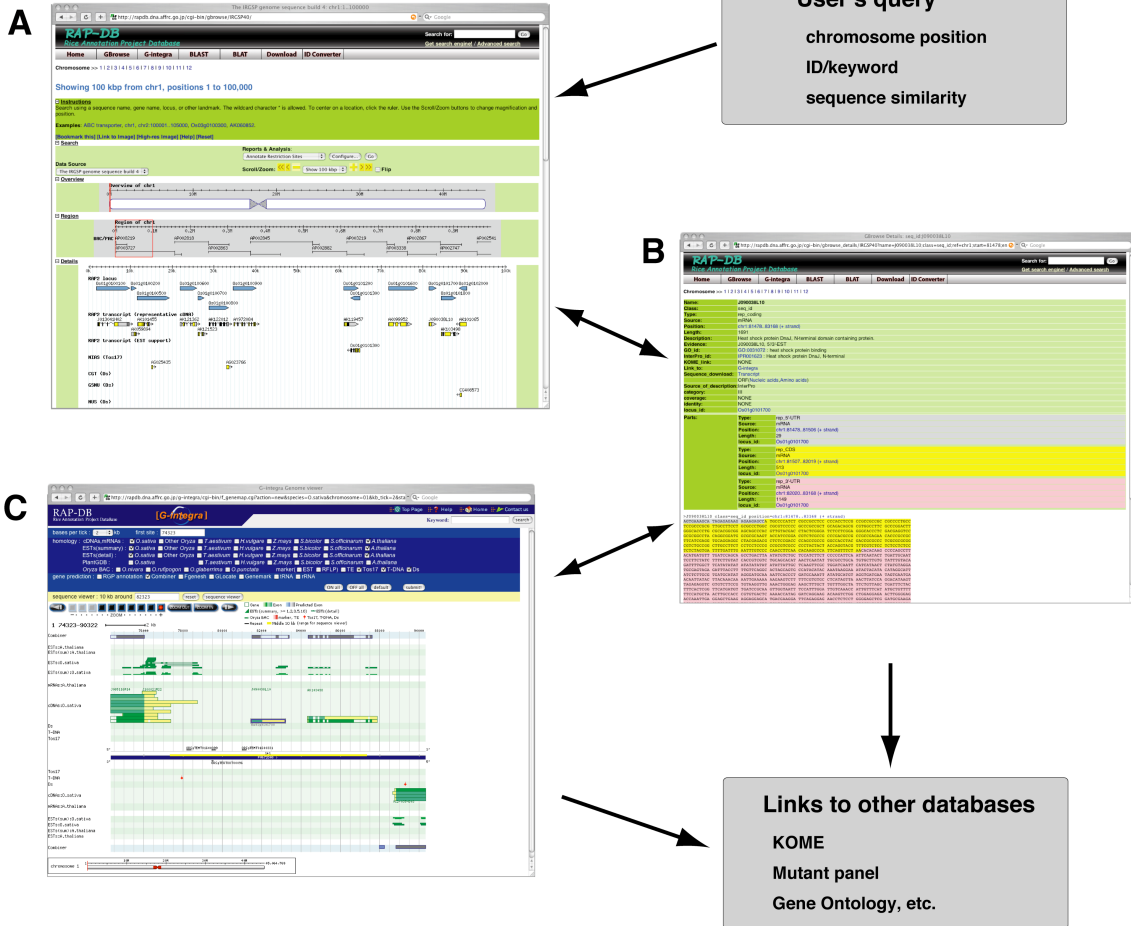



hmp\_08523\_1c.jpg (Zoomed in)



**Appendix Figure 1.** Screenshots of the Molecular Database of *Hydra* Cells. The main window shows gene name and their sequences, UniGENE identities, homologue information, gene ontologies, and whole mount in situ hybridization images. The magnified images of in situ images are also available (sub window).

## Appendix Figure 2.



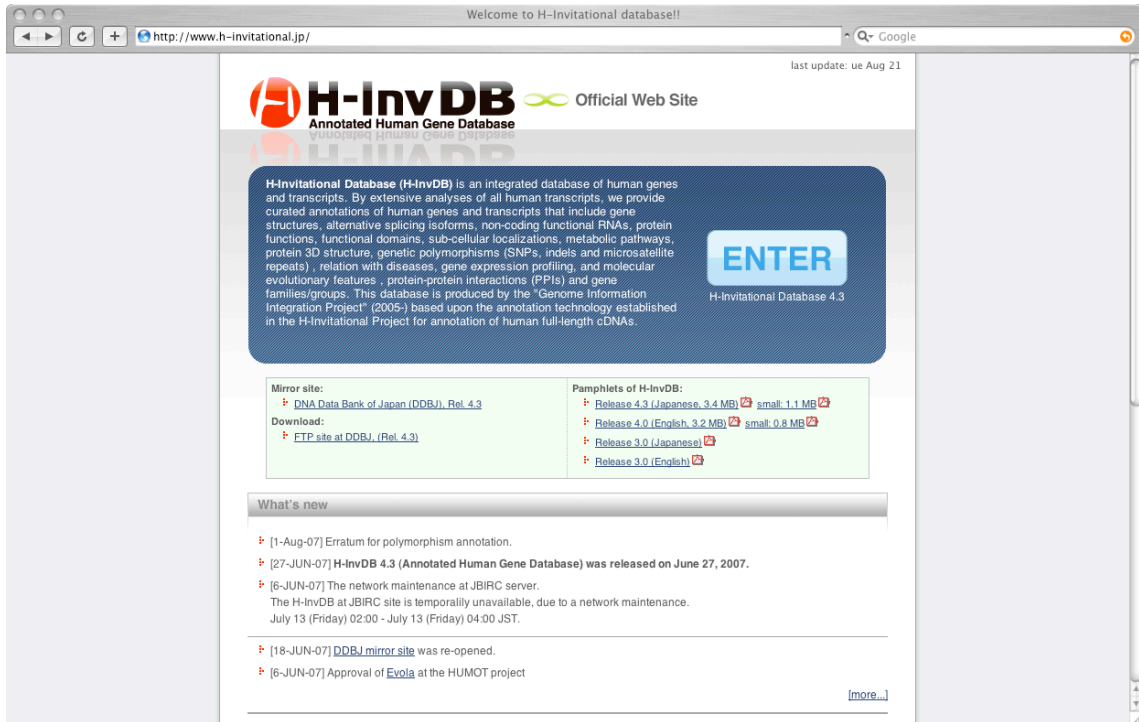
**Appendix Figure 2.** Flowchart of RAP-DB browsing. Users can search the rice genome annotations by chromosomal positions, IDs or keywords. Sequence similarity search by RAP-BLAST or RAP-BLAT is also available. (A) A graphical view of the RAP annotated loci and sequences, *Tos17*-flanking positions, and other tracks illustrated by GBrowse. (B) An annotation table corresponding to the sequence. Several items are hyperlinked to other databases. (C) Browsing a precise genomic view by G-integra.

**Appendix Table 1.**

H-InvDB release	Date of release	Number of transcripts (HIT)	Number of gene clusters (HIX)	Number of proteins (HIP)	Human genome	Date of sequence data-fix
1.0	2003/4/20	41,118	21,037	-	NCBI build 34.1	2002/7/15
2.0	2005/8/31	56,419	25,585	-	NCBI build 34.1	2003/9/1
3.0	2006/3/31	167,992	35,005	-	NCBI build 35.1	2005/3/1
4.0	2007/3/30	175,542	34,701	116,228	NCBI build 36.1	2006/6/15
4.3	2007/6/27	175,536	34,699	116,142	NCBI build 36.1	2006/6/15

**Appendix Table 1. Increase in H-InvDB entries**

Appendix Figure 3.



**Appendix Figure 3. The official web site of H-InvDB**

**(<http://www.h-invitational.jp/>).**

## References

- Altschul, S.F. and Koonin, E.V. Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23** (1998), pp. 444-7.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25** (1997), pp. 3389-402.
- Cavalier-Smith, T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *Int J Syst Evol Microbiol* **52** (2002), pp. 297-354.
- Cavalier-Smith, T. Only six kingdoms of life. *Proc Biol Sci* **271** (2004), pp. 1251-62.
- Dowell, R.D., Jokerst, R.M., Day, A., Eddy, S.R. and Stein, L. The distributed annotation system. *BMC Bioinformatics* **2** (2001), p. 7.
- Embley, T.M. and Martin, W. Eukaryotic evolution, changes and challenges. *Nature* **440** (2006), pp. 623-30.
- Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle (2005), pp. Distributed by the author.
- Fumoto, M., Miyazaki, S. and Sugawara, H. Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. *Nucleic Acids Res* **30** (2002), pp. 66-8.
- Gadal, O., Strauss, D., Kessl, J., Trumpower, B., Tollervey, D. and Hurt, E. Nuclear export of 60s ribosomal subunits depends on Xpo1p and requires a nuclear export sequence-containing factor, Nmd3p, that associates with the large subunit protein Rpl10p. *Mol Cell Biol* **21** (2001), pp. 3405-15.
- Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.D., Flaherty, P., Foury, F., Garfinkel, D.J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kotter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D.,



- Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.Y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W. and Johnston, M. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418** (2002), pp. 387-91.
- Hedges, S.B. The origin and evolution of model organisms. *Nat Rev Genet* **3** (2002), pp. 838-49.
- Ho, J.H. and Johnson, A.W. NMD3 encodes an essential cytoplasmic protein required for stable 60S ribosomal subunits in *Saccharomyces cerevisiae*. *Mol Cell Biol* **19** (1999), pp. 2389-99.
- Ho, J.H., Kallstrom, G. and Johnson, A.W. Nmd3p is a Crm1p-dependent adapter protein for nuclear export of the large ribosomal subunit. *J Cell Biol* **151** (2000), pp. 1057-66.
- Hori, H., Higo, K. and Osawa, S. The rates of evolution in some ribosomal components. *J Mol Evol* **9** (1977), pp. 191-201.
- Horiike, T., Hamada, K., Kanaya, S. and Shinozawa, T. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. *Nat Cell Biol* **3** (2001), pp. 210-4.
- Horiike, T., Hamada, K., Miyata, D. and Shinozawa, T. The origin of eukaryotes is suggested as the symbiosis of pyrococcus into gamma-proteobacteria by phylogenetic tree based on gene content. *J Mol Evol* **59** (2004), pp. 606-19.
- Horiike, T., Hamada, K. and Shinozawa, T. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria supported by the newly clarified origin of functional genes. *Genes Genet Syst* **77** (2002), pp. 369-76.
- Hu, R.-M., et al. Gene expression profiling in the human hypothalamus-pituitary-adrenal axis and full-length cDNA cloning. *Proc. Natl. Acad. U.S.A.* **97** (2000), pp. 9543-9548.
- Hwang, J.S., Ohyanagi, H., Hayakawa, S., Osato, N., Nishimiya-Fujisawa, C., Ikeo, K., David, C.N., Fujisawa, T. and Gojobori, T. The evolutionary emergence of cell type-specific genes inferred from the gene expression analysis of Hydra. *Proc Natl Acad Sci U S A* **104** (2007), pp. 14735-40.
- Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Yura, K., Miyazaki, S.,

Ikeo, K., Homma, K., Kasprzyk, A., Nishikawa, T., Hirakawa, M., Thierry-Mieg, J., Thierry-Mieg, D., Ashurst, J., Jia, L., Nakao, M., Thomas, M.A., Mulder, N., Karavidopoulou, Y., Jin, L., Kim, S., Yasuda, T., Lenhard, B., Eveno, E., Suzuki, Y., Yamasaki, C., Takeda, J., Gough, C., Hilton, P., Fujii, Y., Sakai, H., Tanaka, S., Amid, C., Bellgard, M., Bonaldo Mde, F., Bono, H., Bromberg, S.K., Brookes, A.J., Bruford, E., Carninci, P., Chelala, C., Couillault, C., de Souza, S.J., Debily, M.A., Devignes, M.D., Dubchak, I., Endo, T., Estreicher, A., Eyraas, E., Fukami-Kobayashi, K., Gopinath, G.R., Graudens, E., Hahn, Y., Han, M., Han, Z.G., Hanada, K., Hanaoka, H., Harada, E., Hashimoto, K., Hinz, U., Hirai, M., Hishiki, T., Hopkinson, I., Imbeaud, S., Inoko, H., Kanapin, A., Kaneko, Y., Kasukawa, T., Kelso, J., Kersey, P., Kikuno, R., Kimura, K., Korn, B., Kuryshv, V., Makalowska, I., Makino, T., Mano, S., Mariage-Samson, R., Mashima, J., Matsuda, H., Mewes, H.W., Minoshima, S., Nagai, K., Nagasaki, H., Nagata, N., Nigam, R., Ogasawara, O., Ohara, O., Ohtsubo, M., Okada, N., Okido, T., Oota, S., Ota, M., Ota, T., et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* **2** (2004), p. e162.

Itoh, T., Tanaka, T., Barrero, R.A., Yamasaki, C., Fujii, Y., Hilton, P.B., Antonio, B.A., Aono, H., Apweiler, R., Bruskiwich, R., Bureau, T., Burr, F., Costa de Oliveira, A., Fuks, G., Habara, T., Haberer, G., Han, B., Harada, E., Hiraki, A.T., Hirochika, H., Hoen, D., Hokari, H., Hosokawa, S., Hsing, Y.I., Ikawa, H., Ikeo, K., Imanishi, T., Ito, Y., Jaiswal, P., Kanno, M., Kawahara, Y., Kawamura, T., Kawashima, H., Khurana, J.P., Kikuchi, S., Komatsu, S., Koyanagi, K.O., Kubooka, H., Lieberherr, D., Lin, Y.C., Lonsdale, D., Matsumoto, T., Matsuya, A., McCombie, W.R., Messing, J., Miyao, A., Mulder, N., Nagamura, Y., Nam, J., Namiki, N., Numa, H., Nurimoto, S., O'Donovan, C., Ohyanagi, H., Okido, T., Oota, S., Osato, N., Palmer, L.E., Quetier, F., Raghuvanshi, S., Saichi, N., Sakai, H., Sakai, Y., Sakata, K., Sakurai, T., Sato, F., Sato, Y., Schoof, H., Seki, M., Shibata, M., Shimizu, Y., Shinozaki, K., Shinso, Y., Singh, N.K., Smith-White, B., Takeda, J., Tanino, M., Tatusova, T., Thongjuea, S., Todokoro, F., Tsugane, M., Tyagi, A.K., Vanavichit, A., Wang, A., Wing, R.A., Yamaguchi, K., Yamamoto, M., Yamamoto, N., Yu, Y., Zhang, H., Zhao, Q., Higo, K., Burr, B., Gojobori, T. and Sasaki, T. Curated genome annotation of *Oryza sativa* ssp. *japonica* and comparative genome analysis with *Arabidopsis thaliana*. *Genome*

- Res* **17** (2007), pp. 175-83.
- Johnson, A.W., Lund, E. and Dahlberg, J. Nuclear export of ribosomal subunits. *Trends Biochem Sci* **27** (2002), pp. 580-5.
- Jones, D.T., Taylor, W.R. and Thornton, J.M. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8** (1992), pp. 275-82.
- Kadowaki, T., Hitomi, M., Chen, S. and Tartakoff, A.M. Nuclear mRNA accumulation causes nucleolar fragmentation in yeast mtr2 mutant. *Mol Biol Cell* **5** (1994), pp. 1253-63.
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423** (2003), pp. 241-54.
- Kikuno, R., Nagase, T., Waki, M. and Ohara, O. HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res* **30** (2002), pp. 166-168.
- Kimura, M., The neutral theory of molecular evolution. Cambridge University Press, Cambridge (1983).
- Kumar, S., Tamura, K. and Nei, M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* **5** (2004), pp. 150-63.
- Lake, J.A. Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequences. *Nature* **331** (1988), pp. 184-6.
- Lake, J.A. and Rivera, M.C. Was the nucleus the first endosymbiont? *Proc Natl Acad Sci U S A* **91** (1994), pp. 2880-1.
- Lopez-Garcia, P. and Moreira, D. Metabolic symbiosis at the origin of eukaryotes. *Trends Biochem Sci* **24** (1999), pp. 88-93.
- Mans, B.J., Anantharaman, V., Aravind, L. and Koonin, E.V. Comparative genomics, evolution and origins of the nuclear envelope and nuclear pore complex. *Cell Cycle* **3** (2004), pp. 1612-37.
- Martin, W. Archaeobacteria (Archaea) and the origin of the eukaryotic nucleus. *Curr Opin Microbiol* **8** (2005), pp. 630-7.
- Martin, W. and Koonin, E.V. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* **440** (2006), pp. 41-5.
- Martin, W. and Muller, M. The hydrogen hypothesis for the first eukaryote. *Nature* **392**

- (1998), pp. 37-41.
- Matsumoto, T. The map-based sequence of the rice genome. *Nature* **436** (2005), pp. 793-800.
- Maurer, P., Redd, M., Solsbacher, J., Bischoff, F.R., Greiner, M., Podtelejnikov, A.V., Mann, M., Stade, K., Weis, K. and Schlenstedt, G. The nuclear export receptor Xpo1p forms distinct complexes with NES transport substrates and the yeast Ran binding protein 1 (Yrb1p). *Mol Biol Cell* **12** (2001), pp. 539-49.
- Moore, G., Devos, K.M., Wang, Z. and Gale, M.D. Cereal genome evolution. Grasses, line up and form a circle. *Curr Biol* **5** (1995), pp. 737-9.
- Moreira, D. and Lopez-Garcia, P. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *J Mol Evol* **47** (1998), pp. 517-30.
- Nakamura, Y., Itoh, T., Matsuda, H. and Gojobori, T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36** (2004), pp. 760-6.
- Nash, R., Weng, S., Hitz, B., Balakrishnan, R., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., Fisk, D.G., Hirschman, J.E., Hong, E.L., Livstone, M.S., Oughtred, R., Park, J., Skrzypek, M., Theesfeld, C.L., Binkley, G., Dong, Q., Lane, C., Miyasato, S., Sethuraman, A., Schroeder, M., Dolinski, K., Botstein, D. and Cherry, J.M. Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res* **35** (2007), pp. D468-71.
- Nissan, T.A., Bassler, J., Petfalski, E., Tollervey, D. and Hurt, E. 60S pre-ribosome formation viewed from assembly in the nucleolus until export to the cytoplasm. *Embo J* **21** (2002), pp. 5539-47.
- Ohyanagi, H., Ikeo, K. and Gojobori, T. Eukaryotic nuclear structure explains the evolutionary rate difference of ribosome export factors. *Gene* **421** (2008a), pp. 7-13.
- Ohyanagi, H., Ikeo, K. and Gojobori, T. The origin of nucleus: Rebuild from the prokaryotic ancestors of ribosome export factors. *Gene* **423** (2008b), pp. 149-152.
- Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., Fujii, Y., Antonio, B.A., Nagamura, Y., Imanishi, T., Ikeo, K., Itoh, T., Gojobori, T. and Sasaki, T. The Rice Annotation Project Database (RAP-DB): hub for Oryza

- sativa ssp. japonica genome information. *Nucleic Acids Res* **34** (2006), pp. D741-4.
- Ota, T., Nishikawa, T., Suzuki, Y., Maruyama, K., Sugano, S. and Isogai, T. Full-length cDNA project toward a high throughput functional analysis. *Microb Comp Genomics* **2** (1997), pp. 204-205.
- Raška, I., Shaw, P.J. and Cmarko, D. Structure and function of the nucleolus in the spotlight. *Curr Opin Cell Biol* **18** (2006), pp. 325-34.
- Saitou, N. and Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4** (1987), pp. 406-25.
- Sasaki, T. and Burr, B. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr Opin Plant Biol* **3** (2000), pp. 138-41.
- Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res* **29** (2001), pp. 2994-3005.
- Searcy, D.G., The Origin and Evolution of the Cell (eds Matsuno, H.H. & Matsuno, K.). World Scientific, Singapore (1992).
- Strausberg, R.L., EA, F., LH, G., JG, D., RD, K., FS, C., L, W., CM, S., GD, S., SF, A., B, Z., KH, B., CF, S., NK, B., RF, H., H, J., T, M., SI, M., J, W., F, H., L, D., K, M., AA, F., GM, R., L, H., M, S., MB, S., MF, B., TL, C., TE, S., MJ, B., TB, U., S, T., P, C., C, P., SS, R., NA, L., GJ, P., RD, A., SJ, M., SA, B., PJ, M., KJ, M., JA, M., PH, G., S, R., KC, W., S, H., AM, G., LJ, G., SW, H., DK, V., DM, M., EJ, S., X, L., RA, G., J, F., E, H., M, K., A, M., S, R., A, S., M, W., A, M., AC, Y., Y, S., GG, B., RW, B., JW, T., ED, G., MC, D., AC, R., J, G., J, S., RM, M., YS, B., MI, K., U, S., DE, S., A, S., JE, S., SJ, J., MA., M. and Team, T.M.G.C.P. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl. Acad. U.S.A.* **99** (2002), pp. 16899-16903.
- Sugawara, H., Abe, T., Gojobori, T. and Tateno, Y. DDBJ working on evaluation and classification of bacterial genes in INSDC. *Nucleic Acids Res* **35** (2007), pp. D13-5.
- Tanaka, T., Antonio, B.A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., Sakai, H., Wu, J., Itoh, T., Sasaki, T., Aono, R., Fujii, Y., Habara, T., Harada, E., Kanno,

- M., Kawahara, Y., Kawashima, H., Kubooka, H., Matsuya, A., Nakaoka, H., Saichi, N., Sanbonmatsu, R., Sato, Y., Shinso, Y., Suzuki, M., Takeda, J., Tanino, M., Todokoro, F., Yamaguchi, K., Yamamoto, N., Yamasaki, C., Imanishi, T., Okido, T., Tada, M., Ikeo, K., Tateno, Y., Gojobori, T., Lin, Y.C., Wei, F.J., Hsing, Y.I., Zhao, Q., Han, B., Kramer, M.R., McCombie, R.W., Lonsdale, D., O'Donovan, C.C., Whitfield, E.J., Apweiler, R., Koyanagi, K.O., Khurana, J.P., Raghuvanshi, S., Singh, N.K., Tyagi, A.K., Haberer, G., Fujisawa, M., Hosokawa, S., Ito, Y., Ikawa, H., Shibata, M., Yamamoto, M., Bruskiwich, R.M., Hoen, D.R., Bureau, T.E., Namiki, N., Ohyanagi, H., Sakai, Y., Nobushima, S., Sakata, K., Barrero, R.A., Souvorov, A., Smith-White, B., Tatusova, T., An, S., An, G., S, O.O., Fuks, G., Messing, J., Christie, K.R., Lieberherr, D., Kim, H., Zuccolo, A., Wing, R.A., Nobuta, K., Green, P.J., Lu, C., Meyers, B.C., Chaparro, C., Piegue, B., Panaud, O. and Echeverria, M. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36** (2008), pp. D1028-33.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22** (1994), pp. 4673-80.
- Tschochner, H. and Hurt, E. Pre-ribosomes on the road from the nucleolus to the cytoplasm. *Trends Cell Biol* **13** (2003), pp. 255-63.
- Vellai, T., Takacs, K. and Vida, G. A new aspect to the origin and evolution of eukaryotes. *J Mol Evol* **46** (1998), pp. 499-507.
- Warner, J.R. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* **24** (1999), pp. 437-40.
- Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **35** (2007), pp. D5-12.
- Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W.,

- Bocher, M., Blocker, H., Bauersachs, S., Blum, H., Lauber, J., Dusterhoft, A., Beyer, A., Kohrer, K., Strack, N., Mewes, H.W., Ottenwalder, B., Obermaier, B., Tampe, J., Heubner, D., Wambutt, R., Korn, B., Klein, M. and Poustka, A. Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res* **11** (2001), pp. 422-435.
- Yamasaki, C., Koyanagi, K.O., Fujii, Y., Itoh, T., Barrero, R., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M., Takeda, J., Fukuchi, S., Miyazaki, S., Nomura, N., Sugano, S., Imanishi, T. and Gojobori, T. Investigation of protein functions through data-mining on integrated human transcriptome database, H-Invitational database (H-InvDB). *Gene* **364** (2005), pp. 99-107.
- Yamasaki, C., Murakami, K., Fujii, Y., Sato, Y., Harada, E., Takeda, J., Taniya, T., Sakate, R., Kikugawa, S., Shimada, M., Tanino, M., Koyanagi, K.O., Barrero, R.A., Gough, C., Chun, H.W., Habara, T., Hanaoka, H., Hayakawa, Y., Hilton, P.B., Kaneko, Y., Kanno, M., Kawahara, Y., Kawamura, T., Matsuya, A., Nagata, N., Nishikata, K., Noda, A.O., Nurimoto, S., Saichi, N., Sakai, H., Sanbonmatsu, R., Shiba, R., Suzuki, M., Takabayashi, K., Takahashi, A., Tamura, T., Tanaka, M., Tanaka, S., Todokoro, F., Yamaguchi, K., Yamamoto, N., Okido, T., Mashima, J., Hashizume, A., Jin, L., Lee, K.B., Lin, Y.C., Nozaki, A., Sakai, K., Tada, M., Miyazaki, S., Makino, T., Ohyanagi, H., Osato, N., Tanaka, N., Suzuki, Y., Ikeo, K., Saitou, N., Sugawara, H., O'Donovan, C., Kulikova, T., Whitfield, E., Halligan, B., Shimoyama, M., Twigger, S., Yura, K., Kimura, K., Yasuda, T., Nishikawa, T., Akiyama, Y., Motonon, C., Mukai, Y., Nagasaki, H., Suwa, M., Horton, P., Kikuno, R., Ohara, O., Lancet, D., Eveno, E., Graudens, E., Imbeaud, S., Debily, M.A., Hayashizaki, Y., Amid, C., Han, M., Osanger, A., Endo, T., Thomas, M.A., Hirakawa, M., Makalowski, W., Nakao, M., Kim, N.S., Yoo, H.S., De Souza, S.J., Bonaldo Mde, F., Niimura, Y., Kuryshev, V., Schupp, I., Wiemann, S., Bellgard, M., et al. The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Res* **36** (2008), pp. D793-9.
- Yudate, H.T., et al. HUNT: launch of a full-length cDNA database from the Helix Research Institute. *Nucleic Acids Res* **29** (2001), pp. 185-188.