Research

An inventory of yeast proteins associated with nucleolar and ribosomal components

Eike Staub, Sebastian Mackowiak and Martin Vingron

Address: Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany.

Correspondence: Eike Staub. Email: eike.staub@nucleolus.net

Published: 26 October 2006

Genome ***Biology*** 2006, **7:**R98 (doi:10.1186/gb-2006-7-10-r98)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/10/R98

Received: 18 May 2006
Revised: 26 July 2006
Accepted: 26 October 2006

Abstract

Background: Although baker's yeast is a primary model organism for research on eukaryotic ribosome assembly and nucleoli, the list of its proteins that are functionally associated with nucleoli or ribosomes is still incomplete. We trained a naïve Bayesian classifier to predict novel proteins that are associated with yeast nucleoli or ribosomes based on parts lists of nucleoli in model organisms and large-scale protein interaction data sets. Phylogenetic profiling and gene expression analysis were carried out to shed light on evolutionary and regulatory aspects of nucleoli and ribosome assembly.

Results: We predict that, in addition to 439 known proteins, a further 62 yeast proteins are associated with components of the nucleolus or the ribosome. The complete set comprises a large core of archaeal-type proteins, several bacterial-type proteins, but mostly eukaryote-specific inventions. Expression of nucleolar and ribosomal genes tends to be strongly co-regulated compared to other yeast genes.

Conclusion: The number of proteins associated with nucleolar or ribosomal components in yeast is at least 14% higher than known before. The nucleolus probably evolved from an archaeal-type ribosome maturation machinery by recruitment of several bacterial-type and mostly eukaryote-specific factors. Not only expression of ribosomal protein genes, but also expression of genes encoding the 90S processosome, are strongly co-regulated and both regulatory programs are distinct from each other.

Background

In prokaryotes, heat and distinct ionic conditions are sufficient to assemble a ribosome from its building blocks *in vitro* [1]. In comparison, the biosynthesis of eukaryotic ribosomes is a complicated procedure. Eukaryotic ribosomes are made in the nucleolus, the ribosome factory of a eukaroytic cell. The nucleolus is a dense compartment in the nucleus of eukaryotes where freshly transcribed ribosomal RNA (rRNA) and ribosomal proteins imported from the cytosol meet complex machinery for ribosome maturation and assembly. Ribosomal subunits leave the nucleolus in a state in which the majority of their building blocks are already incorporated [2,3].

Several lines of evidence suggest that ribosome biosynthesis is not the sole function of nucleoli. They have been linked to

cell growth control, sequestering of regulatory molecules (for example, of the cell cycle), modification of small RNAs, mitotic spindle positioning, assembly of non-ribosomal ribonucleoprotein (RNP) particles, nuclear export, and DNA repair [2,4-6]. The wide range of different functions linked to the nucleolus is not surprising when considering the prominent position of ribosome biosynthesis with respect to cellular economy [7]. It seems as if the regulation of a broad range of cellular mechanisms related to cell growth and division is linked to the ribosome biosynthesis machinery through nucleoli. The full range of molecules involved in this cross-talk is only beginning to emerge. Large scale proteomic analyses of nucleolar constituents [8,9] and a survey of the human nucleolar protein network [10] have recently provided a first global picture of the functional network of human nucleoli.

The baker's yeast *Saccharomyces cerevisiae* is a favorite eukaryotic model organism for ribosome-related research. However, knowledge about the set of proteins associated with ribosomes or their nucleolar precursors in yeast is fragmentary. There are currently 439 yeast proteins annotated as ribosomal, ribosome-associated, or nucleolar. Many have been identified in genome-scale protein localization studies [11,12] as well as studies of narrower focus [13-18]. Such experiments usually represent only snapshots of cells in particular states. Furthermore, native protein localization might have been altered when proteins are expressed with fusion tags or as yeast two-hybrid baits or preys. Therefore, it is likely that many additional nucleolar or ribosome-associated proteins are still undiscovered. In support of this hypothesis, studies on the proteomes of human and mouse-ear cress nucleoli [8,9,19,20] identified hundreds of proteins that were unknown before or have not yet been linked to the nucleolus. The lists of nucleolar proteins from these distantly related eukaryotes were only partially overlapping. Moreover, Andersen and colleagues [9,21] found that a large proportion of human nucleolar proteins localize to the nucleolus only transiently, which might also have rendered their discovery in yeast more difficult.

In this study, we aim to extend the fragmentary knowledge about the protein parts list of yeast nucleoli. We present a computational approach to predict novel nucleolar or ribosome biosynthesis proteins of yeast using data from orthologous nucleolar proteins and data sets on pairwise protein interactions or protein complexes. Using a naïve Bayesian classifier we predict novel proteins associated with nucleolar or ribosomal components at high estimated sensitivity and specificity. We study the evolution of these proteins using phylogenetic profiles across 84 prokaryotic and eukaryotic organisms, thereby complementing and extending earlier computational studies on the function and evolution of the nucleolus [21,22]. Finally, we investigate expression patterns of nucleolar and ribosome-associated genes to characterize the substructure of the nucleolar expression program.

Results and discussion

Prologue

This section is divided into three parts. In the first section, we describe a comprehensive list of yeast proteins that we predict to be associated with nucleolar or ribosomal components. Note that in the following paragraphs such proteins will be termed nucleolar or ribosomal component-associated (NRCA) proteins. NRCA proteins do not necessarily have to be associated with the ribosome or to be localized in nucleoli during their whole life cycle. Instead, it is possible that a predicted NRCA protein localizes to the nucleolus only temporarily or binds to nucleolar components outside the nucleolus. All proteins that associate with ribosomal and nucleolar components are the targets of our predictions. In this way we would like to capture all proteins that have the potential to exert important functions on nucleolar and ribosomal biology. In the second part of the study, the identified proteins are subjected to phylogenetic profiling, thereby providing insights into the evolution of the nucleolus and ribosome assembly. Finally, we characterize the gene expression program for NRCA proteins by comparison of expression patterns of diverse functionally or evolutionarily related sets of genes.

Prediction of novel nucleolar and ribosome-associated proteins

A prerequisite for comprehensive functional and evolutionary characterization of the nucleolus and the ribosomal machinery is a complete parts list of its proteins. We applied naïve Bayesian classification to extend the known list of 439 proteins associated with nucleolar and ribosomal components in yeast towards a complete inventory of such proteins. Before prediction of new factors, we performed an extensive cross-validation of our naïve Bayesian classifier to judge whether we are able to predict NRCA proteins with considerable accuracy (Figure 1). To this end, we built 1,000 training sets, performed a cross-validation and obtained 1,000 receiver operating characteristic (ROC) curves. The average area under the ROC curve (AUC) was approximately 0.98, which generally indicates a classifier of high performance. Based on cross-validation and ROC analysis on the training sets, we chose a conservative threshold of $log(O_{post}) > 4$ for the prediction of new NRCA proteins. During cross validation we predicted nucleolar proteins at a sensitivity of 50.4% and a specificity of 98.6% using this threshold, indicating that our predictions are very conservative.

Out of 6,281 proteins that were not annotated as NRCA proteins before, we predicted a further 62 to be linked to nucleolus/ribosome biology (Table 1, Figure 2). The experimental evidence underlying our predictions can be encoded in a 7-bit binary data string. All data strings that occurred in our analysis are summarized in Table 2 along with the prediction results obtained for them. When sensitivity/specificity estimates of the cross-validation runs hold, we estimate that there is approximately 1 false positive prediction among the

62 proteins and that we missed about another 62 proteins by our approach. We conclude that the complete inventory of nucleolar and ribosome-associated proteins in yeast comprises 439 previously known proteins, 62 predicted in this analysis, and about another 62 proteins that remain to be discovered. Thus, we hypothesize that, in total, approximately 560 genes (more than 8% of the total gene content) encode proteins related to nucleolar or ribosomal biology in yeast.

The majority of newly predicted NRCA proteins belong to four functional classes. The first class consists of proteins that were known as regulators of translation before: the translation initiation factors TIF1, SUI3, SUI2, TIF2, GCD1, TIF4631, the translation elongation factors TEF1, TEF4, EFT1, SPT5, the translational release factor SUP45, and the translocon component KAR2. We identified these proteins not only because of their physical interactions with other translation factors or ribosome components, but also because each factor has orthologs in human and/or mouse-ear cress that have been detected in nucleoli. Although the ribosomal association of these factors was known before, their appearance in the nucleolus is surprising. It lends further support to the hypothesis that ribosomal subunits in the nucleus already have translational competence [23-25]. Alternatively, the nucleolar translation factors could support the assembly or quality control of ribosomes, for example, by ensuring through their physical presence that their future binding sites are assembled and modified correctly.

The second class comprises factors that are linked to transcription. Whereas RNA polymerase I is the natural polymerase for the transcription of rRNA genes in the nucleolus, we additionally predicted the nucleolar association of the RNA polymerase II factors SUA7, RPO21, DST1, TFG2, RPB3, TIF4631, and TAF14, and the RNA polymerase III factors RPO31 and RET1. Several of these factors (RPO21, TIF4631, TAF14, RPB3, RPO31, RET1) have not been identified in nucleolar preparations, but were linked to other nucleolar proteins by shared participation in protein complexes and/or interactions in independent experiments. Therefore, it is possible that they associate with nucleolar/ribosomal proteins only outside the nucleolus. The remaining factors were all identified in at least one nucleolar purification experiment, suggesting that they could play yet undiscovered roles as regulators of ribosomal gene expression by RNA polymerase I.

As a third group, we predicted several components of the splicing apparatus to occur also in the nucleolus [26,27]. Among these are components of the major spliceosomal subcomplexes, namely the U1 small nuclear (sn)RNP protein SMD2, the U4/U6 snRNP factors PRP3 and PRP4, the U2A snRNP protein LEA1, the U2 components PRP9 and HSH49, the U5 snRNP protein PRP8, and the Sm core proteins SMX2 and SMD3. Furthermore, we predict the nucleolar localization of the exon junction complex component SUB2 and the spliceosome disassembly protein PRP43. U3 snRNP proteins are already known to contribute to early steps in ribosome assembly and are components of the 90S processosome. We propose that the identified spliceosomal proteins have as yet unknown functions in the assembly of ribosomes and/or other nucleolar RNPs.

The fourth class is linked to the regulation of genomic DNA structure and chromatin. The nucleolar association of several nucleosome components like histone H2A.2 (HTA2), H4 (HHF2), H2B.2 (HTB2), H2B (HFB1), and an H2A variant of the F/Z family (HTZ1) is not surprising as genomic DNA is an integral part of nucleoli that are formed by fusion of so-called nucleolar organizer regions (NORs), stretches of genomic DNA carrying rRNA genes. DNA topoisomerase I (TOP1) could be required to relax tension in DNA structure in NORs, either during replication or transcription. SPT16 is an essential general chromatin assembly factor that is known to assist in RNA polymerase II transcription. Rvb1p (RVB1) is also essential for yeast viability and known as a component of chromatin remodeling complexes. Our results suggest that both proteins are involved in remodeling the chromatin of NORs.

Putative biochemical functions of several further predicted nucleolar proteins are in accordance with a role in nucleolus or ribosome maturation. The gene *DHH1* encodes an RNA helicase of the DEAD box family that was not found in nucleoli of ear cress or human, but interacted with known nucleolar proteins in four independent data sets (Table 1). Another DEAD box RNA helicase encoded by *DBP2* was found in nucleoli and in nucleolar complexes. In combination with their putative biochemical function, this is strong evidence that both RNA helicases play a role in nucleolar RNP assembly. The *BCP1* gene is largely of unknown function, but its deletion is lethal in yeast. It has been linked to nuclear transport and maturation of ribosomes through interactions with a ribosomal lysine methyltransferase (RKM1), to a RAN-binding protein (KAP123), to a ribosomal protein (RPL23A) and to its essentiality for nuclear export of the Mss4p protein. Although little is known about the cellular function of the heat shock proteins HSP82 and SSA2, their occurrence in nucleoli is not surprising because protein folding is a fundamental process during RNP assembly. Similarly, it seems reasonable to assume a ribosomal function for the karyopherins alpha and beta (KAP95, SRP1). The Uso1p-related myosin-like protein (MLP1) is linked to the interior side of the nuclear envelope and nuclear pore. It is proposed to act in the nuclear retention of unspliced messengers. Its identification in nucleolar preparations suggests that it fulfills a similar role in the control of RNA or RNP processing in the nucleolus.

Furthermore, there were several surprising predictions of novel nucleolar proteins. Two subunits (CKA1 and CKB2) of yeast casein kinase 2 (CK2) were predicted to be nucleolar. CK2 is known as a pleiotropic regulator of the cell cycle and has recently been linked to the regulation of chromatin [28].

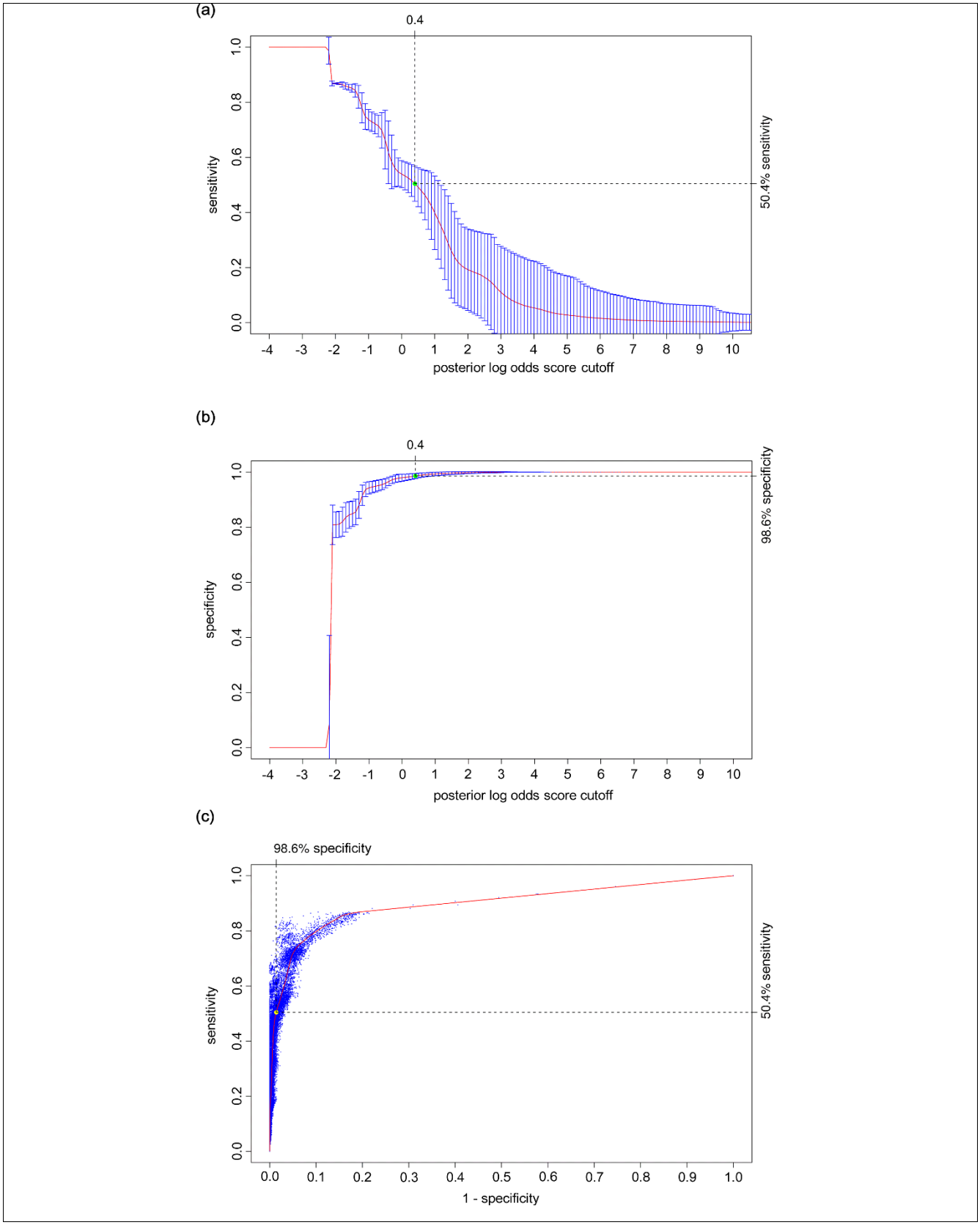

Figure 1 *(see legend on next page)*

Figure 1 *(see previous page)*
Estimation of prediction accuracy. The accuracy of predictions was estimated from 1,000 runs of 10-fold cross-validations using 1,000 alternative training sets (see Materials and methods). The threshold/working point used for the final predictions of new nucleolar proteins is marked in each plot. **(a)** The sensitivity (*SE* = *TP*/(*TP* + *FN*)) of our classifier is plotted over different thresholds of classifier scores (log posterior odds ratios) applied to each cross-validation run. The logarithmic posterior odds ratios indicate how likely it is under the naïve Bayesian model that a protein is an NRCA protein (positive scores) versus that it is not an NRCA protein (negative scores). A single point on the line and its error bar stems from calculations of the average sensitivity and its standard deviation obtained from 1,000 cross-validation runs using a distinct classification score threshold. Confidence intervals are ± 2-fold standard deviation intervals around the mean. Note that at the threshold that was finally used for prediction (0.4) we expect to reach a sensitivity of 50.4%. This means that we have probably still missed as many NRCA proteins as we have predicted (62). **(b)** The specificity (*SP* = *TN*/(*TN* + *FP*)) of our classifier is plotted over different thresholds of classifier thresholds (log posterior odds ratios) that were applied on results of each of 1,000 cross-validation runs. Confidence intervals are ± 2-fold standard deviation intervals around the mean. Note that at the finally used threshold of 0.4 the specificity reaches 0.986, meaning that we expect only 1.4% of false positives among our predictions. **(c)** The ROC curve of our classifier is plotted as sensitivity versus (1-specificity). Each individual data point reflects predictions at a single cross-validation run when a single prediction threshold is applied. The central line is based on averaged SE/SP values for each threshold applied. The ROC curve gives an impression of the quality of a classifier. It is a general indicator of classification performance. The bigger the AUC, the better the classifier. We obtained an AUC value of 0.98, which generally indicates a classification of high quality. The ROC curve was also the basis for the selection of our final classifier threshold, as it illustrates the trade-off between sensitivity and specificity. We chose to be very conservative (high specificity) for the sake of missing true NRCA proteins (lower sensitivity).

Therefore, we hypothesize that CK2 regulates chromatin accessibility in nucleolar organizer regions. Casein kinase 1 is known for its function in intracellular vesicle transport and secretion [29]. A nucleolar role of casein kinase 1 (HHR) was not known during preparation of this manuscript, but was published during the revision stage (see Note added in proof). An F1 beta subunit component of the F1F0-ATPase complex (ATP2) has been detected in nucleolus purifications of both ear cress and human. This strongly suggests a dual function for this protein in respiration and the nucleolus. The nucleolar localization of a mitochondrial ADP/ATP carrier protein (AAC3) was also detected in both model organisms and is supported by protein interactions to nucleolar proteins.

We note that, in total, only 11 of 62 proteins have been identified solely on the basis of protein interactions; the remaining 51 proteins have nucleolar orthologs in model species. We expect that the latter perform yet undiscovered functions in the nucleolus, although they have been linked to extra-nucleolar or even cytosolic processes like splicing, nuclear ribosome import/export, or translation before. The former are candidates for yeast-specific nucleolar localization or for extra-nucleolar ribosome maturation. Further functional characterization is hardly possible using only presently available data and would, therefore, require additional experiments.

Note added in proof: validation of our predictions in the current literature

During revision of this manuscript we became aware of several old and new articles that add experimental evidence to some predictions of nucleolar or ribosome-associated proteins made in this manuscript. We were not of aware of the ribosomal or nucleolar roles of these proteins before, because such annotations were missing in the *Saccharomyces* Genome Database (SGD) database at the time of analysis. In the following we shortly describe these findings of others. Lebaron *et al.* [30] and Leeds *et al.* [31] found that the Prp43 protein, a putative DEAH helicase, is a component of multiple pre-ribosomal particles and localizes to the nucleolus. We predicted a nucleolar role of Prp43 via evidence from nucleolar preparations in model organisms and from protein-protein interactions. Schafer *et al.* [32] have shown recently that the protein kinase HRR25 (casein kinase I) binds pre-40S particles, phosphorylates Rps3 and the maturation factor Enp1, and is required for maturation of the 40S subunit *in vivo*. We predicted a ribosomal/nucleolar role for HRR25 based on the occurrence of the human HRR25 ortholog in nucleolar preparations and on the co-occurrence of HRR25 with other nucleolar proteins in affinity-purified protein complexes (Table 1). In 2001, Bond *et al.* [33] had already shown that DBP2 is not only involved in nonsense-mediated mRNA decay, but is also a ribosome biogenesis factor as DBP2 mutant cells are deficient in free 60S subunits and 25S rRNA is significantly reduced. This link has apparently escaped the attention of SGD database curators for years. We rediscovered the link of DBP2 with ribosomal biology through a prediction based on nucleolar localization of the human DBP2 ortholog and through interactions with nucleolar proteins in protein complex data of two independent studies (see table 1). In 2000, Edwards *et al.* [34] found that yeast topoisomerase TOP1 localizes to the nucleolus dependent on its interaction with nucleolin. We rediscovered this link because of the co-occurrence of yeast TOP1 in protein interactions and complexes with nucleolar components and the nucleolar localization of human TOP1. These four cases are independent experimental validations of our predictions.

Phylogenetic profiling of nucleolar and ribosome-associated proteins

We established presence-absence patterns of genes across multiple organisms, so called phylogenetic profiles, for all 501 NRCA proteins (Figures 2, 3, 4) to investigate their ancestry in the three domains of life. We identified a large cluster of 83 yeast proteins by hierarchical clustering with orthologs in the majority of archaeal species under investigation, but only single orthologs in bacteria (Figure 4). Among the archaeal proteins were many maturation factors and components of the ribosome. From a biochemical viewpoint, together with a few proteins that are ubiquitous in all domains of life, these

Table 1

Classification results and annotation for 62 novel predicted nucleolar/ribosome-associated proteins

Gene	ORF	Hs	At	Ue	It	Kr	Ga	Ho	log(O)	Description
SUA7	YPR086W	1	0	1	0	1	0	o	0.665	TFIIB subunit (transcription initiation factor) factor E
HTA1	YDR225W	1	0	1	0	0	1	0	0.612	Histone H2A
HSC82	YMR186W	1	1	0	0	1	0	0	0.697	Heat shock protein
TIF1	YKR059W	1	1	0	0	1	0	0	0.699	Translation initiation factor 4A
PRP4	YPR178W	1	0	0	0	1	0	1	0.703	U4/U6 snRNP 52 kDa protein
KAR2	YJL034W	1	1	0	0	0	0	0	0.684	Component of ER translocon
HTA2	YBL003C	1	0	0	0	1	1	0	0.724	Histone H2A.2
AAC3	YBR085W	1	1	0	0	0	1	0	0.686	Mitochondrial ADP/ATP carrier - member of the mitochondrial carrier (MCF) family
RFC2	YJR068W	1	1	0	0	0	0	0	0.686	DNA replication factor C 41 kDa subunit
TEF1	YPR080W	1	1	0	0	0	0	0	0.686	Translation elongation factor eEF1 alpha-A chain cytosolic
SMX2	YFL017W-A	1	1	0	0	1	0	0	0.696	snRNP G protein (the homologue of the human Sm-G)
BCP1	YDR361C	1	1	0	0	0	0	0	0.686	Similarity to hypothetical protein *S. pombe*
LEA1	YPL213W	1	1	0	0	1	0	0	0.704	U2 A snRNP protein
HSP82	YPL240C	1	1	0	0	0	0	0	0.686	Heat shock protein
SMD3	YLR147C	1	1	0	0	1	0	0	0.699	Spliceosomal snRNA-associated Sm core protein required for pre-mRNA splicing
TIF2	YJL138C	1	1	0	0	0	1	0	0.686	Translation initiation factor eIF4A
None	YBR025C	1	0	1	0	0	1	0	0.610	Strong similarity to Ylf1p
SPT16	YGL207W	1	1	0	0	1	1	0	0.705	General chromatin factor
SUI2	YJR007W	1	0	0	0	1	1	0	0.720	Translation initiation factor eIF2 alpha chain
HSH49	YOR319W	0	1	0	1	0	1	0	0.702	Essential yeast splicing factor
DED1	YOR204W	0	1	0	0	0	0	1	0.716	ATP-dependent RNA helicase
HTB1	YDR224C	1	1	1	0	0	1	0	0.709	Histone H2B
HRR25*	YPL204W	1	0	0	0	1	1	0	0.718	Casein kinase I Ser/Thr/Tyr protein kinase
SSA2	YLL024C	1	1	0	0	0	0	0	0.686	Heat shock protein of HSP70 family cytosolic
SRP1	YNL189W	0	1	1	0	1	1	1	0.696	Karyopherin-alpha or importin
SUB2	YDL084W	1	1	0	0	0	0	0	0.686	Probably involved in pre-mRNA splicing
CKA1	YIL035C	1	0	0	0	1	1	1	0.698	Casein kinase II catalytic alpha chain
PRP43*	YGL120C	1	1	1	0	1	1	0	0.695	Involved in spliceosome disassembly
SUI3	YPL237W	1	0	0	0	1	1	0	0.721	Translation initiation factor eIF2 beta subunit
DST1	YGL043W	1	0	0	0	0	0	1	0.692	TFIIS (transcription elongation factor)
PRP8	YHR165C	1	0	0	0	1	1	0	0.721	U5 snRNP protein pre-mRNA splicing factor
PRP9	YDL030W	1	0	1	0	1	0	0	0.667	Pre-mRNA splicing factor (snRNA-associated protein)
SUP45	YBR143C	1	0	1	0	1	1	0	0.704	Translational release factor
ASC1	YMR116C	1	1	0	0	1	0	0	0.698	40S small subunit ribosomal protein
DBP2*	YNL112W	1	0	0	0	1	1	0	0.719	ATP-dependent RNA helicase of DEAD box family
CKB2	YOR039W	1	0	0	1	1	1	0	0.710	Casein kinase II beta chain
YRA1	YDR381W	1	0	0	0	1	1	0	0.720	RNA annealing protein
GCD11	YER025W	1	0	1	0	0	1	0	0.609	Translation initiation factor eIF2 gamma chain
TFG2	YGR005C	1	0	0	0	1	1	1	0.695	TFIIF subunit (transcription initiation factor) 54 kDa
TOP1*	YOL006C	1	0	1	1	1	0	0	0.693	DNA topoisomerase I
BRR2	YER172C	1	1	0	0	0	0	1	0.708	RNA helicase-related protein
RVB1	YDR190C	1	1	1	0	0	1	0	0.709	RUVB-like protein
MLP1	YKR095W	1	1	0	0	0	0	0	0.686	Myosin-like protein related to Uso1p
HTZ1	YOL012C	1	1	0	0	0	0	0	0.685	Evolutionarily conserved member of the histone H2A F/Z family of histone variants
ATP2	YJR121W	1	1	0	0	0	0	0	0.685	F1F0-ATPase complex F1 beta subunit
SMD2	YLR275W	1	1	0	1	1	0	0	0.688	U1 snRNP protein of the Sm class

Table 1 *(Continued)*

Classification results and annotation for 62 novel predicted nucleolar/ribosome-associated proteins

PRP3	YDR473C	1	0	0	0	1	0	1	0.704	Essential splicing factor
EFT1	YOR133W	1	1	0	0	0	0	0	0.682	Translation elongation factor eEF2
HTB2	YBL002W	1	1	0	0	0	1	0	0.690	Histone H2B.2
TEF4	YKL081W	1	0	0	0	1	1	0	0.718	Translation elongation factor eEF1 gamma chain
HHF2	YNL030W	1	1	0	0	1	0	0	0.695	Histone H4
Predictions based solely on protein interactions										
RPO21	YDL140C	0	0	0	0	1	1	1	0.728	DNA-directed RNA polymerase II 215 kDa subunit
DHH1	YDL160C	0	0	1	1	1	1	0	0.714	Putative RNA helicase of the DEAD box family
CFT1	YDR301W	0	0	0	0	1	1	1	0.731	Pre-mRNA 3-end processing factor CF II
KAP95	YLR347C	0	0	1	0	1	1	1	0.689	Karyopherin-beta
SPT5	YML010W	0	0	0	0	1	1	1	0.732	Transcription elongation protein
TAF14	YPL129W	0	0	0	0	1	1	1	0.733	TFIIF subunit (transcription initiation factor) 30 kDa
RPB3	YIL021W	0	0	0	0	1	1	1	0.728	DNA-directed RNA-polymerase II 45 kDa
RPO31	YOR116C	0	0	0	0	1	1	1	0.729	DNA-directed RNA polymerase III 160 kDa subunit
TIF4631	YGR162W	0	0	0	0	1	1	1	0.734	mRNA cap-binding protein (eIF4F) 150K subunit
PRP24	YMR268C	0	0	0	0	1	1	1	0.734	Pre-mRNA splicing factor
RET1	YOR207C	0	0	0	0	1	1	1	0.731	DNA-directed RNA polymerase III 130 kDa subunit

The data used for classification and the detailed prediction results are listed for all 62 proteins that passed our threshold of $O_{post} > 0.4$. These proteins had not been annotated as associated with nucleolar or ribosomal components before, but were classified as such in our analysis. A literature survey for the predicted proteins revealed that for four proteins a role in the nucleolus and ribosome biogenesis had already been established (see Note added in proof). The lower part of the table lists 11 proteins that were predicted as NRCA proteins solely on the basis of shared participation in complexes or interactions. For these proteins, we do not necessarily predict a nucleolar localization, but direct interaction with nucleolar/ribosomal components at least under one specific cellular condition at an unspecified locus within the cell. *Four proteins for which recent articles have confirmed a role in ribosome biogenesis or the nucleolus. The results are supplemented by a concise annotation for each protein from the Comprehensive Yeast Genome Database (CYGD) [72]. The header line contains abbreviations describing the column content: Gene, gene symbol of yeast gene; ORF, yeast open reading frame ID; Hs, orthology to human nucleolar protein; At, orthology to mouse-ear cress nucleolar protein; It, link to nucleolar protein via Y2H interaction in Ito dataset; Ue, link to nucleolar protein via Y2H interaction in Uetz dataset; Ga, link to nucleolar protein via participation in a complex in Gavin data set; Ho, link to nucleolar protein via participation in a complex in Ho data set; Kr, link to nucleolar protein via participation in a complex in Krogan data set; log(O), average posterior odds ratio from all prediction runs in which the protein was not used for training; Description, concise description of protein function.

archaeal-type proteins seem to represent the functional core of the nucleolus and of ribosome maturation.

There is a considerable, but much smaller, fraction of nucleolar proteins that have orthologs in bacteria, but not in archaea (Figure 3). Among these are RRP5, which is essential for the processing of 18S and 5.8S rRNA [35], and the 3'-5' exonuclease DIS3, which is required for the processing of 5.8S rRNA and is a component of the exosome [36]. Eukaryotes have employed these bacterial-type proteins for the processing of archaeal-type ribosomes. More detailed phylogenetic studies will have to show whether these bacterial-type proteins are even of alpha-proteobacterial (that is, mitochondrial/hydrogenosomal) origin. Interestingly, several proteins of mitochondrial ribosomes seem to localize to the nucleolus (MRPS28, MRPL9, MRPL23, YML6). Unlike most other mitochondrial ribosomal proteins, YML6 is essential for yeast viability, indicating that it does not exclusively function in mitochondria. The dual nucleolar and mitochondrial localization of these proteins means that they have taken over important functions in nuclear ribosome maturation in addition to their roles in mitochondrial ribosomes. RNAase III encoded by the *RTS1* gene is involved in the processing of U2 snRNA, highlighting also the chimeric evolutionary origin of the machinery for RNA splicing. The tRNA-isopentenyltranferase MOD5 is known as one of the few proteins that occur in three subcellular compartments: cytosol, mitochondria, and the nucleus [37]. Its phylogenetic profile shows that MOD5 shares a common sequence ancestor with bacteria. The finding that eukaryotes employed bacterial-type, possibly mitochondrial, proteins to supplement the archaeal-type ribosome maturation machinery is congruent with earlier observations on the level of protein domains [22].

The largest fraction of yeast NRCA proteins has multiple orthologs in eukaryotes, but none in prokaryotes. Many of these proteins can be regarded as eukaryotic inventions. This group spans the whole range of nucleolar and ribosome-related functions. Explicitly, we investigated the profiles of components of the 90S processosome, a large complex attached to freshly transcribed rRNA that performs early maturation steps before ribosomal proteins and rRNA are assembled into subunits. The 90S processosome proteins do not show strong similarity to prokaryotic proteins, although they are strongly conserved in eukaryotes (Figure 4). As ribosome assembly in eukaryotes is much more complex than in prokaryotes, the finding that the 90S processosomal machinery has no prokaryotic counterpart is not surprising.

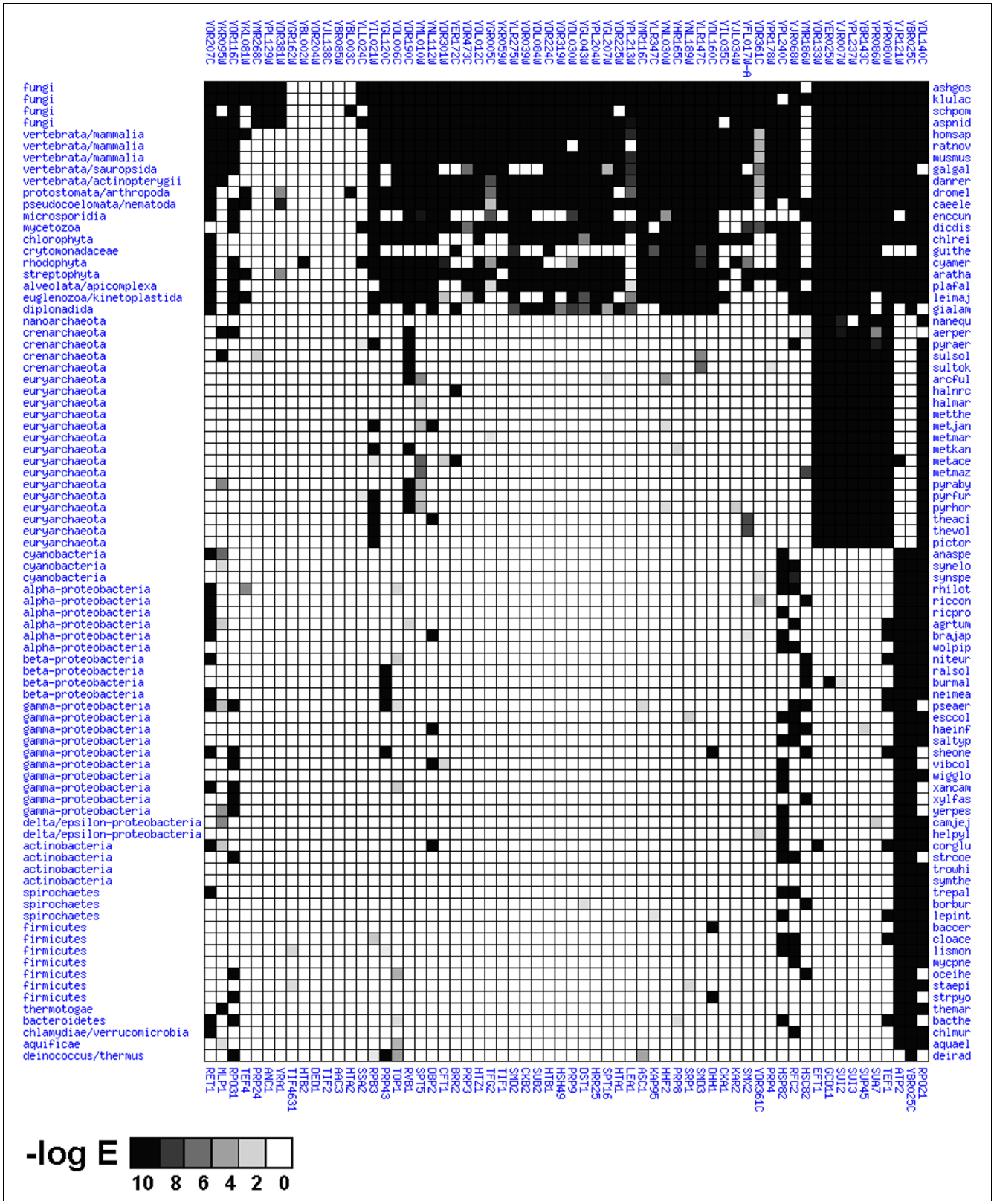

Figure 2 *(see legend on next page)*

Figure 2 *(see previous page)*
Phylogenetic profiling of novel nucleolar/ribosome-associated proteins. Phylogenetic profiles of 62 previously unrecovered nucleolar/ribosome-associated proteins of yeast across 84 organisms. The profiles were generated using the best reciprocal hit method with yeast as a reference organism (see Materials and methods). Abbreviations given on the top of the plot represent organism names (first three letters for genus and first three letters of species names; see Materials and methods for a translation of abbreviations into organism names). Further taxonomic annotation is given on the bottom of the plot. Yeast open reading frame identifiers are given on the left side, and gene names and descriptions are given on the right side of the plot. The significance of sequence similarity is visualized by different shades of gray that reflect the logarithmic expectation (E) value from reciprocal BLAST searches (shown at the bottom of the figure). Here, the E values of BLAST searches using target proteome sequences as queries versus the yeast proteome reference database are shown. The genes are ordered according to hierarchical clustering (see Materials and methods).

It shows that a large machinery of proteins acting in concert at an early step during ribosome maturation has been invented exclusively for the eukaryotic branch of life.

Implications for hypotheses on the origin of eukaryotes

What do all these results mean with respect to hypotheses about the origin of eukaryotes? Although a phylogenetic profile can reveal a prokaryotic ancestry, it can not prove a prokaryotic origin of a nucleolar protein. This question has to be studied for all proteins by single phylogenetic analyses that are beyond the scope of this study. When the first genomes were available in the late 1990s, sequence comparisons led to the postulates that 'informational' proteins in eukaryotes stem from archaea and 'operational' proteins stem from bacteria and several authors have put forward hypotheses on the origin of eukaryotes based on 'genome fusion' [38-42]. Kurland *et al.* [43] have recently called these interpretations into question and argued that whole-genome sequence comparisons, many phylogenetic analyses (in which eukaryotic proteins do not branch within archaeal or bacterial orthologs), and so called eukaryote-specific cellular signature structures (CSSs) rather show that eukaryotes represent a primordial lineage and are not just an amalgamation of prokaryotic genomes. According to another recent hypothesis, eukaryotes, archaea and bacteria each evolved by independent transitions from the RNA world to the DNA world through viral transduction [44]. The latter two hypotheses postulate that eukaryotes comprise a lineage as equally old as bacteria and archaea and are, hereafter, referred to as 'primordial eukaryote' hypotheses.

According to 'genome fusion' hypotheses, the existence of nucleolar proteins of archaeal and bacterial type would mean that the nucleolus is chimeric, with building blocks acquired from both archaea and bacteria. In contrast, 'primordial eukaryote' hypotheses would either explain prokaryotic-type proteins by gene uptake (either by horizontal gene transfer, viral transfer or endosymbiosis) or by common ancestry with genes in the last universal common ancestor (LUCA) with subsequent loss in either the bacterial or archaeal lineage.

The fact that the largest fraction of nucleolar proteins lacks counterparts in prokaryotes suggests that the nucleolus is primarily a eukaryotic invention. According to 'genome fusion' hypotheses, the many eukaryote-specific nucleolar proteins would have evolved after the genome fusion that led to the first eukaryote, thus at a relatively late time point in evolution. According to the 'primordial eukaryote' view, eukaryote-specific nucleolar proteins would be as equally old as the prokaryote-type proteins and should also be witnesses of early eukaryote (and even earliest cellular) evolution.

So far, considerations based on phylogenetic profiling do not rule out either type of hypothesis. However, our study also shows that proteins of the functional core of nucleoli are not distributed evenly across the three evolutionary groups (archaeal like, bacterial like, eukaryote specific). It is the archaeal-like set of proteins in combination with the ubiquitous proteins that represent the functional core of nucleoli and ribosome maturation. This leads us to the postulate that bacterial-type and eukaryote-specific proteins were assembled around an archaeal-type functional core, and, therefore, emerged later in the ribosome maturation machinery. How does this fit into the different types of hypotheses?

The timely order of cellular transitions outlined above would fit the 'genome fusion' hypotheses in which nucleoli evolved as a compensatory mechanism to prevent dilution of ribosome assembly factors in an early eukaryotic lineage [22]. This would have been necessary to maintain the efficiency of ribosome assembly in eukaryotes. At some time point the eukaryotic lineage must have evolved towards larger cell sizes, a development made possible by more efficient catabolism via mitochondria or hydrogenosomes [22]. In this scenario, nucleoli have emerged after the mitochondrial precursor symbiont entered its host cell, probably as a result of special pressure exerted by larger cell volumes.

Under such a hypothesis of nucleolar evolution based on 'genome fusion' it is possible that eukaryotes with mitochondria (or mitochondrial/hydrogenosomal remnants) exist that have never evolved nucleoli. In contrast, eukaryotes with nucleoli and without mitochondria would not be compatible with the hypothesis. Today, the existence of a eukaryote that lacks either mitochondria or nucleoli (or remnants of them) has not been proven [45]. Recently, Xin *et al.* [46] described a typical nucleolar protein in *Giardia lamblia* and concluded that Giardia once had nucleoli. We conclude that, so far, 'genome fusion' hypotheses are compatible with current data on nucleolar evolution.

Table 2

Summary of effective prediction rules obtained by Bayesian classification

Hs	At	Ue	It	Kr	Ga	Ho	Prediction: associated with nucleolar or ribosomal component?
0	0	0	0	0	0	0	No
0	0	0	0	0	0	I	No
0	0	0	0	0	I	0	No
0	0	0	0	0	I	I	No
0	0	0	0	I	0	0	No
0	0	0	0	I	0	I	No
0	0	0	0	I	I	0	No
0	0	0	I	0	0	0	No
0	0	0	I	0	I	0	No
0	0	0	I	I	0	0	No
0	0	0	I	I	I	0	No
0	0	I	0	0	0	0	No
0	0	I	0	0	0	I	No
0	0	I	0	0	I	0	No
0	0	I	0	I	0	0	No
0	0	I	0	I	I	0	No
0	0	I	I	0	0	0	No
0	0	I	I	0	I	0	No
0	I	0	0	0	0	0	No
0	I	0	0	0	I	0	No
0	I	0	0	I	0	0	No
I	0	0	0	0	0	0	No
I	0	0	0	0	I	0	No
I	0	0	0	I	0	0	No
I	0	0	I	0	0	0	No
I	0	I	0	0	0	0	No
0	0	0	0	I	I	I	Yes
0	0	I	0	I	I	I	Yes
0	0	I	I	I	I	0	Yes
0	I	0	0	0	0	I	Yes
0	I	0	I	0	I	0	Yes
0	I	I	0	I	I	I	Yes
I	0	0	0	0	0	I	Yes
I	0	0	0	I	0	I	Yes
I	0	0	0	I	I	0	Yes
I	0	0	0	I	I	I	Yes
I	0	0	I	I	I	0	Yes
I	0	I	0	0	I	0	Yes
I	0	I	0	I	0	0	Yes
I	0	I	0	I	I	0	Yes
I	0	I	I	I	0	0	Yes

Table 2 *(Continued)*

Summary of effective prediction rules obtained by Bayesian classification

I	I	0	0	0	0	0	Yes
I	I	0	0	0	0	I	Yes
I	I	0	0	0	I	0	Yes
I	I	0	0	I	0	0	Yes
I	I	0	0	I	I	0	Yes
I	I	0	I	I	0	0	Yes
I	I	I	0	0	I	0	Yes
I	I	I	0	I	I	0	Yes

Our Bayesian classification approach assigns a distinct prediction to each possible binary pattern that could be associated with a protein. With the seven data sources used here, only a limited number of 128 different combinations of binary evidences is possible. Here, all binary patterns that occur in our data set are enumerated. They are supplemented by the prediction result to illustrate which input data generates which prediction. Note that neither a single protein interaction nor a single occurrence in nucleoli of model organisms is sufficient for a positive prediction, and that evidence from three protein interaction experiments is necessary for a positive prediction in the absence of evidence based on orthologs in nucleolar preparations of model organisms. Column headers denote the data source: Hs, orthology to human nucleolar protein; At, orthology to mouse-ear cress nucleolar protein; Ue, link to nucleolar protein via Y2H interaction in Uetz dataset; It, link to nucleolar protein via Y2H interaction in Ito dataset; Kr, link to nucleolar protein via participation in a complex in Krogan data set; Ga, link to nucleolar protein via participation in a complex in Gavin data set; Ho, link to nucleolar protein via participation in a complex in Ho data set.

'Primordial eukaryote' hypotheses presented so far have been less specific about the timely order of events that generated eukaryotic signature structures. Also, the driving forces that led to major eukaryotic signature structures have not been proposed. Hypotheses that postulate a eukaryotic 'raptor' as the host cell that acquired mitochondria imply that a nucleolus and nucleolar structures like 90S processosomes (eukaryotic signature structures) preceded mitochondria. This means that all eukaryotes with mitochondria should also have nucleoli or nucleolar remnants. It seems as if all known eukaryotes fulfill this criterion. However, unlike for 'genome fusion' hypotheses, one might argue that eukaryotes with nuclei and nucleoli that never had mitochondria/hydrogenosomes should have survived until today. But, as many recent studies have shown, the existence of such eukaryotes has not so far been proven [45].

In summary, our phylogenetic profiles are not sufficient to rule out either 'primordial eukaryote' or 'genome fusion' hypotheses. However, each future hypothesis about eukaryotic origins would also have to explain the hallmarks of nucleolar evolution highlighted above, that is, the archaeal nature of the functional core of nucleoli to which bacterial-type additions and many eukaryote-specific proteins were recruited.

Distinct nucleolar gene expression programs: the ribosome and the 90S processosome

The expression compendium of Hughes *et al.* [47] reflects a considerable part of the global yeast expression program. We studied this data set to identify particular groups of yeast genes that are expressed similarly across the 300 experiments of this global genetic perturbation study (Figure 5).

First, we compared the correlation of expression between genes that encode nucleolar and ribosome-associated proteins with the correlation within all other yeast proteins. There is considerably higher correlation of expression patterns among NRCA protein-encoding genes, suggesting that there is a special nucleolar expression program. One might suspect that the ancient archaeal core of nucleoli, which includes many ribosomal proteins and maturation factors, constitutes a nucleolar subcomponent that exhibits an especially high degree of expression co-regulation. Therefore, we divided our set of nucleolar/ribosome-associated proteins into an archaeal set and a non-archaeal set and compared the correlation of expression within these groups. The distributions of correlation coefficients look rather similar, suggesting that evolutionary age or sequence conservation is not paralleled by tight expression co-regulation.

In contrast, the protein components of the ribosome show a marked co-regulation that is much stronger than the co-regulation observed for all nucleolar proteins. The 90S processosome is a large particle formed around unprocessed rRNA (see previous section). Surprisingly, we found that the degree of co-regulation among 90S processosomal genes is comparable with, if not higher than, that among ribosomal protein genes. We next asked whether co-regulation of ribosomal and 90S-processosomal genes is coupled, that is, whether they are under the control of the same expression program. The cross-comparison of expression vectors of genes from both particles suggests that their expression is different (Figure 5).

We examined this difference in more detail by unsupervised clustering of expression data for a fused list of genes from both large complexes (Figure 6). Hierarchical clustering revealed that the majority of genes were distributed among two large clusters. One cluster was composed nearly entirely of ribosomal proteins, and the other cluster nearly entirely of 90S processosome proteins. We concluded that the 90S-processosomal expression program is highly co-regulated, but different from the ribosomal program. Thus, the 90S processosome proteins not only differ from their ribosomal functional associates with respect to evolution (see above),

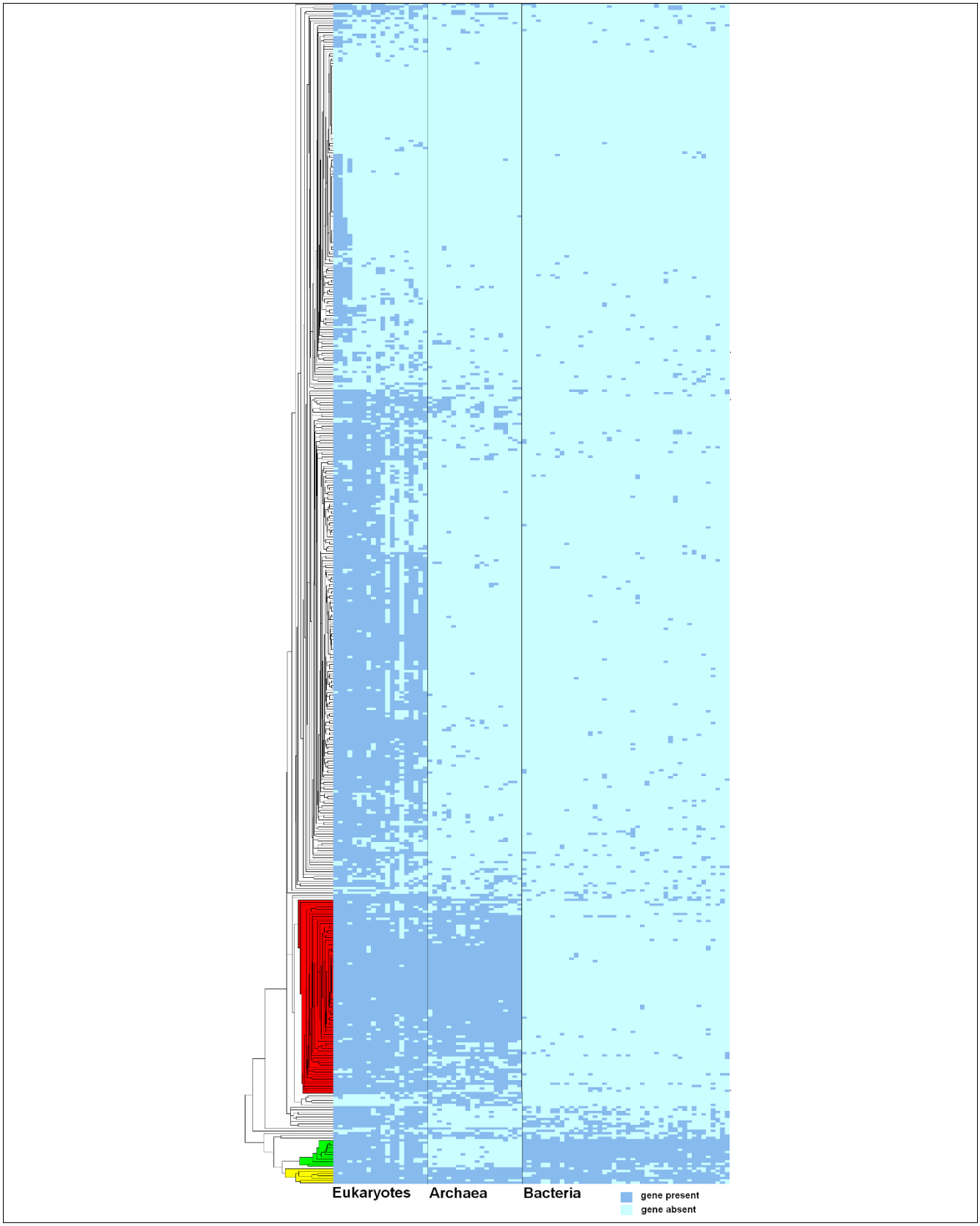

Figure 3 *(see legend on next page)*

Figure 3 *(see previous page)*
Hierarchical clustering of phylogenetic profiles of nucleolar proteins. Phylogenetic profiles of all 501 nucleolar or ribosome-associated proteins. Organisms vary along the horizontal axis, proteins along the vertical axis. Presence of a gene is indicated by dark blue, absence by light blue. Organisms from the three domains of life are separated by black bars. The dendrogram resulting from protein-wise hierarchical clustering is given on the left. Several evolutionarily meaningful clusters emerged, which are colored in the dendrogram: red, proteins of archaeal origin; yellow, ubiquitous proteins; green, proteins of (eu-)bacterial origin. Note that the eukaryote-only genes constitute the largest group, followed by the archaea/eukaryote group. There is a considerable number of genes with orthologs only in bacteria and eukaryota, but not in archaea.

but also with respect to gene expression. We note that there is a large overlap between the sets of proteins of the 90S processosome and the so-called Ribi (ribosome biosynthesis) regulon [48-50]. Of 52 proteins of the 90S processosome, 46 are also components of the Ribi regulon. We propose that the 90S processosomal genes constitute a functionally defined module of the Ribi regulon.

Furthermore, phylogenetic profiles suggest that most 90S processosome components are not just remnants of prokaryotic precursor proteins that could stem from an amalgamation of archaeal and bacterial contributions during the origin of eukaryotes. The eukaryote-specific conservation of many 90S processosome proteins rather suggests that the 90S processosome emerged solely during eukaryotic evolution. Thus, the 90S processosome can be regarded as an ancient eukaryote-specific functional module.

Conclusion

Baker's yeast is the major model organism for the study of eukaryotic nucleolar processes, in particular the assembly of ribosomes. However, recent studies in other eukaryotic model organisms suggest that only a fraction of nucleolar and ribosome biogenesis proteins of *S. cerevisiae* is known today. Using large-scale data sets of nucleolar proteins in *Homo sapiens* and *Arabidopsis thaliana* and protein interactions and complexes in *S. cerevisiae*, we predicted with high confidence that 62 further proteins are associated with nucleolar or ribosomal components, thereby extending the list of nucleolar/ribosomal component-associated proteins to 501. A survey of their presence-absence patterns across 84 organisms from all domains of life confirmed a shared ancestry of the nucleolar functional core with archaea. It also revealed several additions of bacterial character, and that the majority of nucleolus- and ribosome-associated proteins in yeast are eukaryote-specific. Proteins of the 90S processosome tend to be conserved across eukaryotes, but not in prokaryotes. In summary, this suggests an exclusive emergence of many nucleolar ribosome maturation factors in the eukaryote lineage. These findings represent novel insights into transitions leading to eukaryote-specific structures and represent cornerstones that have to be addressed by future hypotheses on the origin of eukaryotes. Furthermore, the analysis of a public gene expression compendium revealed that genes encoding the 90S processosome are nearly as tightly regulated as genes encoding ribosomal proteins, but that the gene expression programs of the ribosome and 90S processosome are distinct.

Materials and methods

Compiling data for classification

Let each yeast protein have an associated data vector $\vec{x}_k = \{x_1, x_2, ..., x_K\}$ with x_k denoting an individual experimental observation for experiment k. This binary data vector carries information that will be used to predict whether a single protein is nucleolar or not. The total number of different sources of evidence is K. Annotations of nucleolar localization for 439 yeast proteins were retrieved from the SGD. We used seven sources of evidence ($K = 7$), hereafter also termed data columns, to judge whether a yeast protein is likely to be nucleolar or not. By default, all observations were set to $x_k = 0$.

Data columns $k = 1$, $k = 2$ contain information on whether a yeast protein has an ortholog in human/*Arabidopsis* that has been detected in purified nucleoli of these organisms by mass spectrometry. We determined orthology relationships between yeast proteins and human/*Arabidopsis* proteins using INPARANOID [51]. If a yeast protein has an ortholog in a model organism that was found in nucleoli, we set the observation in the associated data vector to $x_k = 1$. For data columns $k = 3$, $k = 4$ we used the yeast two-hybrid data sets for protein-protein interactions of Uetz *et al.* [52] and Ito *et al.* [53]. Whenever a yeast protein was involved in a pairwise protein interaction with another protein that is among the 219 known nucleolar proteins, we set the associated observation to $x_k = 1$. For data columns $k = 5$, $k = 6$, $k = 7$ we used the yeast protein complex data sets of Gavin *et al.* [54], Ho *et al.* [55] and Krogan *et al.* [56]. Whenever a yeast protein interacted with a nucleolar protein through shared participation in a complex, we set the associated observation to $x_k = 1$. The resulting data matrix with K observations for each yeast protein was used for the prediction of new nucleolar proteins.

The naïve Bayesian classifier

We use a Bayesian formalism to contrast the hypothesis that a given protein is nucleolar ($H = nuc$) with the hypothesis that it is not nucleolar ($H = \overline{nuc}$). According to Bayes rule, the conditional probability that a protein is nucleolar given its associated data $\vec{x}$ is:

Figure 4 *(see legend on next page)*

Figure 4 *(see previous page)*
Phylogenetic profiling of the 90S processosome. Phylogenetic profiles of known yeast 90S processosome proteins across 84 organisms. Abbreviations given on the top of the plot represent organism names (first three letters for genus and first three letters of species names; see Materials and methods for a translation of abbreviations into organism names). Further taxonomic annotation is given on the bottom of the plot. Yeast open reading frame identifiers are given on the left side, and gene names and descriptions are given on the right side of the plot. The significance of sequence similarity is visualized by different shades of gray that reflect the logarithmic expectation (E) value from reciprocal BLAST searches (shown at the bottom of the figure). Here, the E values of BLAST searches using target proteome sequences as queries versus the yeast proteome reference database are shown. The genes are ordered according to hierarchical clustering (see Materials and methods). Note that there are only a few proteins with many prokaryotic orthologs when compared to Figure 3.

$$P(nuc \mid \vec{x}) = \frac{P(\vec{x} \mid nuc) \times P(nuc)}{P(\vec{x})} \qquad (1)$$

where $P(\vec{x} \mid nuc)$ is the likelihood $L(nuc)$ of the data $\vec{x}$ under the hypothesis $H = nuc$. The conditional probability that a protein is not nucleolar is assigned accordingly.

The posterior odds ratio O_{post} reflects how much more likely it is that a particular protein m is nucleolar than that it is not:

$$O_{post} = \frac{P(nuc \mid \vec{x})}{P(\overline{nuc} \mid \vec{x})} = \frac{P(\vec{x} \mid nuc)}{P(\vec{x} \mid \overline{nuc})} \times \frac{P(nuc)}{P(\overline{nuc})} = LR \times O_{prior} \qquad (2)$$

The prior odds ratio O_{prior} expresses the prior belief that an unknown protein is nucleolar before seeing its associated data. A lower bound on this prior is estimated from current knowledge about the number of nucleolar proteins (439) and the number of total proteins (6,720) in yeast. We set O_{prior} = 439/(6,720 - 439). The first term in the last equation is the likelihood ratio $LR = L(nuc)/L(\overline{nuc})$ of the data given a pair of hypotheses (here $H = nuc$ or $H = \overline{nuc}$). The likelihood ratio contains all information on how we should update our prior belief that a particular protein is nucleolar in the light of its associated data. Thus, the posterior odds ratio can be thought of as an updated version of the prior odds ratio after the data have been seen.

How is the likelihood ratio *LR* calculated? In naïve Bayesian classification one 'naïvely' assumes that the observations from different data sources are independent. Then, the likelihood ratio of a complete set of observed data points is just the product of the likelihood ratios for individual observations:

$$LR = \prod_{k=1}^{K} LR_k = \prod_{k=1}^{K} \frac{L_k(nuc)}{L_k(\overline{nuc})} \qquad (3)$$

The individual likelihoods $L_k(H)$, $H \in \{nuc, \overline{nuc}\}$ can easily be calculated from the positive and negative training data, that is, sets of proteins that are nucleolar and that are not nucleolar and their associated data. Individual data points x_k are binary. Let all $n(H)$, $H \in \{nuc, \overline{nuc}\}$ be the number of proteins in the training data that fulfill hypothesis H. Let all $n(x_k, H)$, $H \in \{nuc, \overline{nuc}\}$ be the number of proteins in the training data that fulfill hypothesis H and have associated data x_k (either 0 or 1). Thus, for a new protein the likelihoods $L_k(H)$, $H \in \{nuc, \overline{nuc}\}$ are calculated as follows:

$$L_k(nuc) = P(x_k \mid nuc) = \frac{n(x_k, nuc)}{n(nuc)} \qquad (4)$$

$$L_k(\overline{nuc}) = P(x_k \mid \overline{nuc}) = \frac{n(x_k, \overline{nuc})}{n(\overline{nuc})} \qquad (5)$$

Training the classifier and estimation of classification performance

To train our classifier for the prediction of new nucleolar proteins we needed positive and negative training data, that is, sets of known nucleolar and non-nucleolar proteins. We retrieved overlapping lists of 219 known nucleolar proteins, 239 proteins acting in ribosome biosynthesis, and 159 proteins associated with cytosolic ribosomes from the SGD. This resulted in a non-redundant set of 439 nucleolar proteins, which we used as positive training cases.

The acquisition of negative training cases was not as straightforward, because we suspect that, among the remaining 6,720 - 439 = 6,281 yeast proteins, a considerable number are nucleolar ones. We consciously decided not to chose a biologically motivated approach to acquire negative training examples to avoid introducing an unknown biological bias into the negative training set (for example, by taking only extracellular or organelle proteins as negative training data). Instead, we obtained 1,000 random samples of 439 proteins (the same size as the positive set) from all but the 439 nucleolar proteins.

Each sample of negative training cases was combined with the unique positive training set to yield a complete training data set. For each of these 1,000 training data sets we performed 10-fold cross-validation. We applied a range of thresholds to determine the sensitivity ($SE = TP/(TP + FN)$) and specificity ($SP = TN/(TN + FP)$) and the ROC curve for each of the 1,000 cross-validation runs. We determined the average AUC from 1,000 cross-validation runs to judge the quality of our classifier.

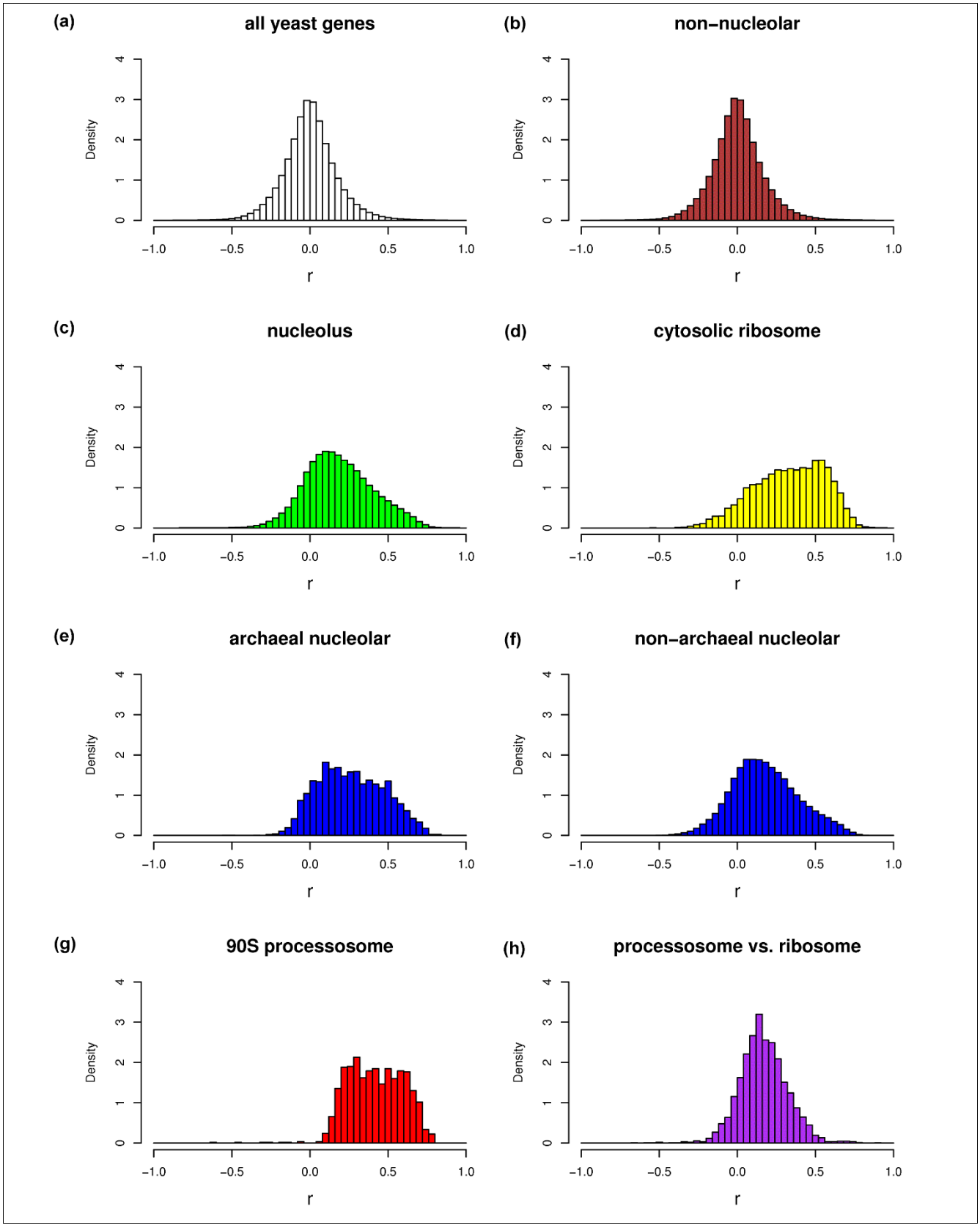

Figure 5 *(see legend on next page)*

Figure 5 *(see previous page)*
Survey of nucleolar/ribosomal gene expression. Histograms of sets of pairwise Pearson correlation coefficients computed from vectors of gene expression ratios for gene pairs. The distributions of Pearson correlation coefficients (each obtained from the pairwise comparison of expression profiles of two genes) gives an impression of the global similarity of expression patterns in a group of genes. Random data would give a Pearson correlation coefficient distribution centered around 0 (no correlation). The more a distribution deviates towards +1 compared to a 0-centered bell shape, the more similar a group of genes is expressed across the whole expression compendium. Gene pairs were formed within or between the functional/evolutionarily-defined groups of genes that are under investigation here. **(a)** Correlation within all yeast genes. **(b)** Correlation within genes that do not encode nucleolar proteins. **(c)** Correlation within genes for nucleolar proteins. **(d)** Correlation within genes for ribosomal or ribosome-associated proteins. **(e)** Correlation within nucleolar genes that stem from archaea. **(f)** Correlation within nucleolar genes that do not stem from archaea. **(g)** Correlation within genes that encode 90S processosome components. **(h)** Correlation between genes for ribosome proteins and 90S processosome proteins. Note that the distributions for the ribosomal protein genes and the 90S processosome strongly deviate from the rather 0-centered distribution of 'all genes-versus-all gene' comparisons. However, the distribution for gene pairs in which one partner is a 90S processosome component and the other partner is a ribosomal component deviate much less from the random shape and, thus, indicate distinct expression programs.

Prediction of novel nucleolus or ribosome-associated proteins

After encouraging cross-validation results, we tried to predict new nucleolar proteins from the set of 6,281 proteins not previously assigned as nucleolar using a similar strategy. We randomly sampled 1,000 sets of 439 negative training cases from the 6,281 proteins and combined each with the 439 positive training cases, thus yielding 1,000 training data sets. We built 1,000 classifiers from these training data sets. With each classifier we made predictions for all proteins not used for training. For a single prediction we used a threshold of $log(O_{post}) = 0.4$. Application of this threshold led to a sensitivity of 50.5% and a specificity of 98.6% during cross-validation. In total, we obtained approximately 900 predictions for each protein (less than 1,000 because we only considered predictions in which the protein was not used for training). The actual classifier decision as to whether a single protein is nucleolar or not was a majority vote based on all approximately 900 predictions.

Based on the prediction results (see Results and discussion; Figure 1), we estimate that our set of 6,281 non-NRCA proteins - for which we assumed that the majority of cases are not NRCA proteins - does probably contain approximately 124 positives (1.97%). Thus, in retrospect, we estimate that, on average, 9 of the presumed non-NRCA proteins are actually positive when we sample 439 proteins at random to compile a negative training set. This probably led to a slight reduction in sensitivity and specificity of the classifier. However, as we can not compile a better set of negative training cases without introducing a systematic bias, we argue that repeated random sampling of negative training cases is an adequate procedure for this classification problem.

Phylogenetic profiling of yeast nucleolar proteins

We obtained sequences from 84 complete proteomes from the ftp server of the European Bioinformatics Institute or the genome download sites at the Wellcome Trust Sanger Institute, The Institute for Genomics Research (TIGR), and the Marine Biological Laboratory. Protein sequences of *S. cerevisiae* were retrieved from the Munich information center for protein sequences [57-60]. The proteomes of the following organisms were used as target proteomes to derive phylogenetic profiles for yeast proteins that served us as reference proteins. Eukaryota: ashgos, *Ashbya gossypii*; klulac, *Kluyveromyces lactis*; schpom, *Schizosaccharomyces pombe*; aspnid, *Aspergillus nidulans*; homsap, *Homo sapiens*; ratnov, *Rattus norvegicus*; musmus, *Mus musculus*; galgal, *Gallus gallus*; danrer, *Danio rerio*; dromel, *Drosophila melanogaster*; caeele, *Caenorhabditis elegans*; enccun, *Encephalitozoon cuniculi*; dicdis, *Dictyostelium discoideum*; chlrei, *Chlamydomonas reinhardtii*; guithe, *Guillardia theta*; cyamer, *Cyanidioschyzon merolae*; aratha, *Arabidopsis thaliana*; plafal, *Plasmodium falciparum*; leimaj, *Leishmania major*; gialam, *Giadia lamblia*. Archaea: nanequ, *Nanoarchaeum equitans*; aerper, *Aeropyrum pernix*; pyraer, *Pyrobaculum aerophilum*; sulsol, *Sulfolobus solfataricus*; sultok, *Sulfolobus tokodaii*; arcful, *Archaeoglobus fulgidus*; halnrc, *Halobacterium* sp.; halmar, *Haloarcula marismortui*; metthe, *Methanobacterium thermoautotrophicum*; metjan, *Methanococcus jannaschii*; metmar, *Methanococcus maripaludis*; metkan, *Methanopyrus kandleri*; metace, *Methanosarcina acetivorans*; metmaz, *Methanosarcina mazei*; pyraby, *Pyrococcus abyssi*; pyrfur, *Pyrococcus furiosus*; pyrhor, *Pyrococcus horikoshii*; theaci, *Thermoplasma acidophilum*; thevol, *Thermoplasma volcanium*; pictor, *Picrophilus torridus*. Bacteria: anaspe, *Anabaena* sp.; synelo, *Synechococcus elongates*; synspe, *Synechocystis* sp.; rhilot, *Rhizobium loti*; riccon, *Rickettsia conorii*; ricpro, *Rickettsia prowazekii*; agrtum, *Agrobacterium tumefaciens*; brajap, *Bradyrhizobium japonicum*; wolpip, *Wolbachia pipientis*; niteur, *Nitrosomonas europaea*; ralsol, *Ralstonia solanacearum*; burmal, *Burkholderia mallei*; neimea, *Neisseria meningitis*; pseaer, *Pseudomonas aeruginosa*; esccol, *Escherichia coli*; haeinf, *Haemophilus influenzae*; saltyp, *Salmonella typhimurium*; sheone, *Shewanella oneidensis*; vibcol, *Vibrio cholerae*; wigglo, *Wigglesworthia glossinidia brevipalpis*; xancam, *Xanthomonas campestris*; xylfas, *Xylella fastidiosa*; yerpes, *Yersinia pestis*; camjej, *Campylobacter jejuni*; helpyl, *Helicobacter pylori*; corglu, *Corynebacterium glutamicum*; strcoe, *Streptomyces coelicolor*; trowhi, *Tropheryma whipplei*; symthe, *Symbiobacterium thermophilum*; trepal, *Treponema pallidum*; borbur, *Borrelia burgdorferi*; lepint, *Leptospira interrogans*; backer, *Bacillus cereus*; cloace, *Clostridium acetobutylicum*; lismon, *Listeria monocytogenes*; mycpne, *Mycoplasma pneumoniae*; oceihe,

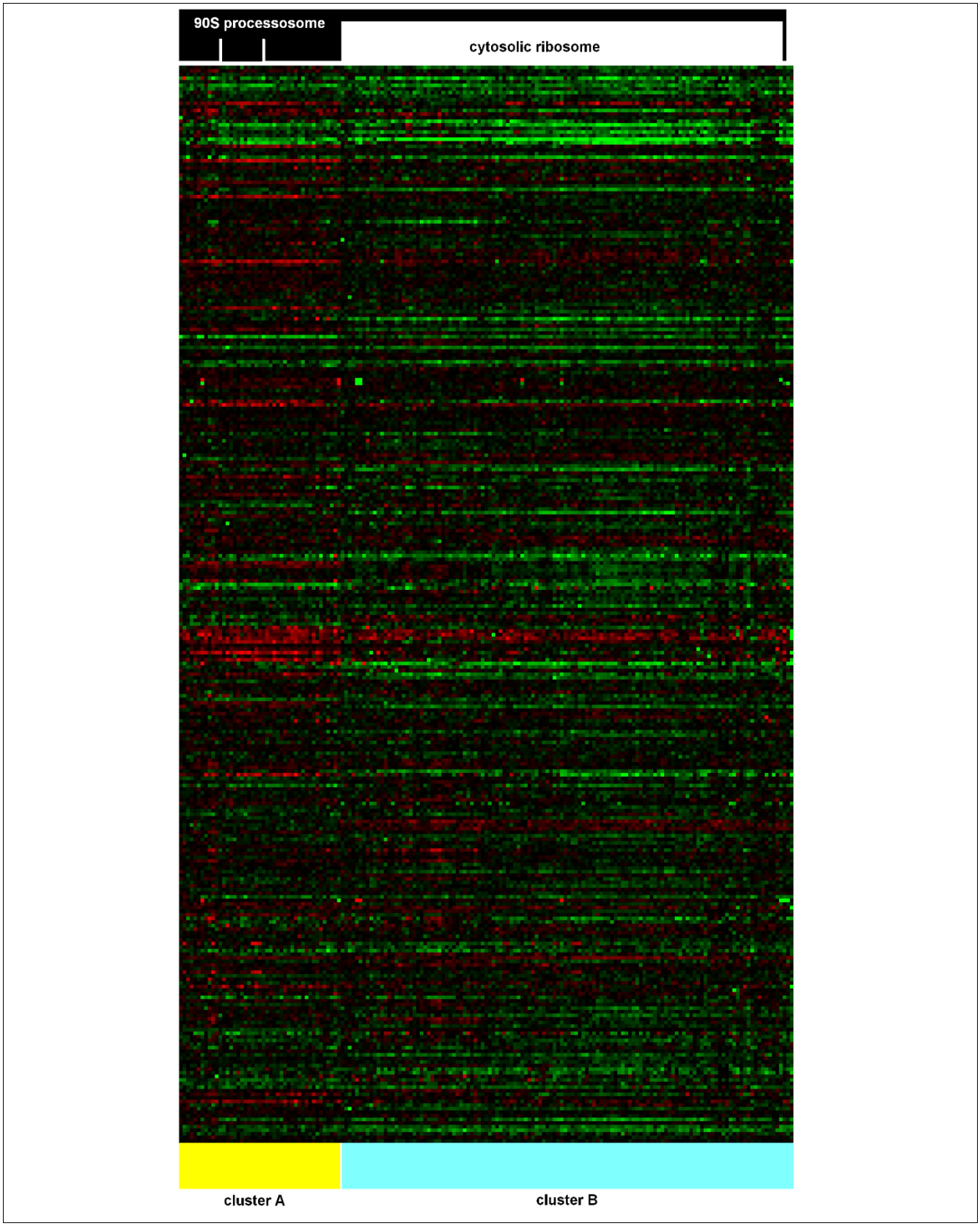

Figure 6 *(see legend on next page)*

Figure 6 *(see previous page)*
Hierarchical clustering of gene expression patterns of ribosomal and processosomal protein genes. The central plot shows color-coded expression ratios as supplied in the ROSETTA expression compendium [47] for genes encoding ribosomal and 90S-processosomal proteins. Genes vary along the horizontal axis, expression experiments vary along the vertical axis. Top: 90S-processosomal genes are marked in black, ribosomal protein genes are marked in white. Bottom: hierarchical clustering yields two large clusters, here marked in cyan and in yellow, that comprise approximately 80% of all ribosomal/processosomal genes (171 of 211). Only genes of these clusters are shown here. Note that only three genes are not clustered according to their membership to either the ribosome or the 90S processosome. The separation of the 90S processosomal and ribosomal protein genes by hierarchical clustering (an unsupervised approach) confirms that the ribosomal and 90S processosomal expression programs are distinct from each other (Figure 5).

Oceanobacillus iheyensis; staepi, *Staphylococcus epidermidis*; strpyo, *Streptococcus pyogenes*; themar, *Thermotoga maritime*; bacthe, *Bacteroides thetaiotaomicron*; chlmur, *Chlamydia muridarum*; aquael, *Aquifex aeolicus*; deirad, *Deinococcus radiodurans*.

We applied the 'best reciprocal hit' (BRH) method to find orthologous protein pairs between the reference proteome and a target proteome. The BRH method is an approximative method for ortholog identification that has been applied by many other groups before to identify orthologs for pairs of organisms with considerable accuracy (see, for example, [61-65]. The BRH method performs worse than more sophisticated phylogeny-based techniques (like reconciliation of phylogenetic and species tree), but, because of its simplicity, it is especially suited for phylogenetic profiling of proteomes of dozens of organisms. More advanced schemes based on pairwise sequence matching (for example, INPARANOID) are also able to find so called 'in-paralogs' (paralogs in one organism that have to be called orthologous to a protein in a second organism with equal right, because they emerged from a duplication after the evolutionary split of these organisms). For ortholog phylogenetic profiling, detection of such paralogs is not so important because the profile scores come from the best hit anyway. Compared to clustering approaches for ortholog identification (COG, orthoMCL), the BRH method relies only on pairwise comparisons of a reference proteome with target genomes and, therefore, more closely adheres to the original definition of orthology, which is always defined between two species. For a more detailed discussion of the BRH method for phylogenetic profiling we refer to our recent study [66].

In our implementation of BRH-based phylogenetic profiling we used yeast proteins as queries to carry out BLASTP searches [67] against each of the 84 proteomes. The best hits for each yeast protein in the individual proteomes were recorded. Then, we used those protein sequences that were identified as best hits as queries in reciprocal BLAST searches of the complete yeast proteome. For BLAST searches we used default parameters (BLOSUM62 matrix, SEG filter on, gap open penalty: 11, gap extension penalty: 1) and an expectation (E) value threshold of $E < 0.1$. Only reciprocal best hits were considered for the construction and visualization of phylogenetic profiles. For visualization, the E values of prokaryote-versus-yeast-proteome searches were color-coded in a yeast-protein-versus-prokaryotic-species matrix, our phylogenetic profile, using white and various levels of gray.

For clustering of phylogenetic profiles, we obtained binary (presence-absence/1 or 0) phylogenetic profiles for all nucleolar proteins of yeast identified here. These profiles were subjected to hierarchical clustering using the centroid method and city-block distances using the software CLUSTER 3.0 [68]. The result helped us to identify sets of nucleolar proteins that stem from archaea, bacteria, or emerged late in eukaryotes. We visualized the results using the Java Treeview software [69] or our own phylogenetic profile viewer [70].

Expression analysis of nucleolar protein components

We performed an analysis of nucleolar expression patterns across 300 experimental conditions of the ROSETTA yeast expression compendium [47]. We aimed to compare co-expression of genes within or between several functionally or evolutionary related groups of genes. Therefore, we determined Pearson correlation coefficients for pairs of genes within or between groups that were calculated from their paired vectors of logarithmic expression ratios across different experimental conditions. We investigated groups that were either identified during the preceding analysis (proteins of the nucleolus, archaeal proteins of the nucleolus, non-archaeal proteins of the nucleolus) or obtained from external resources (the cytosolic ribosome components as recorded by SGD and Gene Ontology [71], 90S processosome proteins listed by Fromont-Racine *et al.* [3]). Histograms of correlation coefficients were determined for each group or group cross-comparison to visualize the degree of co-regulation. Additionally, the expression of genes encoding the cytosolic ribosome and the 90S processosome were investigated using the CLUSTER software (version 3.0). We performed a hierarchical clustering of the original logarithmic expression ratios of these genes using the centroid method and the un-centered correlation option. We visualized the results using the Java Treeview software [69].

Additional data files

The following additional data are available with the online version of this manuscript. Additional data file 1 contains information about known nucleolar proteins and their results

during classifier cross-validation. Additional data file 2 contains all classification results for proteins not predicted as nucleolar.

Acknowledgements

We would like to thank Yun Wan Lam, Angus Lamond, David Marshall, and John Brown for providing nucleolar protein sequences of human and mouse-ear cress. We thank Hannes Luz for valuable comments at various stages of manuscript preparation and Knud H Nierhaus for valuable input on ribosomal biology.

References

1. Nierhaus KH: **The assembly of prokaryotic ribosomes.** *Biochimie* 1991, **73:**739-755.
2. Dez C, Tollervey D: **Ribosome synthesis meets the cell cycle.** *Curr Opin Microbiol* 2004, **7:**631-637.
3. Fromont-Racine M, Senger B, Saveanu C, Fasiolo F: **Ribosome assembly in eukaryotes.** *Gene* 2003, **313:**17-42.
4. Olson MOJ, Hingorani K, Szebeni A: **Conventional and nonconventional roles of the nucleolus.** *Int Rev Cytol* 2002, **219:**199-266.
5. Rudra D, Warner JR: **What better measure than ribosome synthesis?** *Genes Dev* 2004, **18:**2431-2436.
6. Ideue T, Azad AK, ichi Yoshida J, Matsusaka T, Yanagida M, Ohshima Y, Tani T: **The nucleolus is involved in mRNA export from the nucleus in fission yeast.** *J Cell Sci* 2004, **117:**2887-2895.
7. Warner JR: **The economics of ribosome biosynthesis in yeast.** *Trends Biochem Sci* 1999, **24:**437-440.
8. Scherl A, Couté Y, Déon C, Callé A, Kindbeiter K, Sanchez JC, Greco A, Hochstrasser D, Diaz JJ: **Functional proteomic analysis of human nucleolus.** *Mol Biol Cell* 2002, **13:**4100-4109.
9. Andersen JS, Lyon CE, Fox AH, Leung AKL, Lam YW, Steen H, Mann M, Lamond AI: **Directed proteomic analysis of the human nucleolus.** *Curr Biol* 2002, **12:**1-11.
10. Hinsby AM, Kiemer L, Karlberg EO, Lage K, Fausbøll A, Juncker AS, Andersen JS, Mann M, Brunak S: **A wiring of the human nucleolus.** *Mol Cell* 2006, **22:**285-295.
11. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast.** *Nature* 2003, **425:**686-691.
12. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, *et al.*: **Subcellular localization of the yeast proteome.** *Genes Dev* 2002, **16:**707-719.
13. Bassler J, Grandi P, Gadal O, Lessmann T, Petfalski E, Tollervey D, Lechner J, Hurt E: **Identification of a 60S preribosomal particle that is closely linked to nuclear export.** *Mol Cell* 2001, **8:**517-529.
14. Bernstein KA, Baserga SJ: **The small subunit processome is required for cell cycle progression at G1.** *Mol Biol Cell* 2004, **15:**5038-5046.
15. Dragon F, Gallagher JEG, Compagnone-Post PA, Mitchell BM, Porwancher KA, Wehner KA, Wormsley S, Settlage RE, Shabanowitz J, Osheim Y, *et al.*: **A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis.** *Nature* 2002, **417:**967-970.
16. Grandi P, Rybin V, Bassler J, Petfalski E, Strauss D, Marzioch M, Schäfer T, Kuster B, Tschochner H, Tollervey D, *et al.*: **90S pre-ribosomes include the 35S pre-rRNA, the U3 snoRNP, and 40S subunit processing factors but predominantly lack 60S synthesis factors.** *Mol Cell* 2002, **10:**105-115.
17. Harnpicharnchai P, Jakovljevic J, Horsey E, Miles T, Roman J, Rout M, Meagher D, Imai B, Guo Y, Brame CJ, *et al.*: **Composition and functional characterization of yeast 66S ribosome assembly intermediates.** *Mol Cell* 2001, **8:**505-515.
18. Nissan TA, Bassler J, Petfalski E, Tollervey D, Hurt E: **60S pre-ribosome formation viewed from assembly in the nucleolus until export to the cytoplasm.** *EMBO J* 2002, **21:**5539-5547.
19. Andersen JS, Lam YW, Leung AKL, Ong SE, Lyon CE, Lamond AI, Mann M: **Nucleolar proteome dynamics.** *Nature* 2005, **433:**77-83.
20. Pendle AF, Clark GP, Boon R, Lewandowska D, Lam YW, Andersen J, Mann M, Lamond AI, Brown JWS, Shaw PJ: **Proteomic analysis of the *Arabidopsis* nucleolus suggests novel nucleolar functions.** *Mol Biol Cell* 2005, **16:**260-269.
21. Leung AKL, Andersen JS, Mann M, Lamond AI: **Bioinformatic analysis of the nucleolus.** *Biochem J* 2003, **376:**553-569.
22. Staub E, Fiziev P, Rosenthal A, Hinzmann B: **Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire.** *Bioessays* 2004, **26:**567-581.
23. Brogna S, Sato TA, Rosbash M: **Ribosome components are associated with sites of transcription.** *Mol Cell* 2002, **10((1)):**93-104.
24. Iborra FJ, Jackson DA, Cook PR: **The case for nuclear translation.** *J Cell Sci* 2004, **117:**5713-5720.
25. Ishigaki Y, Li X, Serin G, Maquat LE: **Evidence for a pioneer round of mRNA translation: mRNAs subject to nonsense-mediated decay in mammalian cells are bound by CBP80 and CBP20.** *Cell* 2001, **106:**607-617.
26. Zhou Z, Licklider LJ, Gygi SP, Reed R: **Comprehensive proteomic analysis of the human spliceosome.** *Nature* 2002, **419:**182-185.
27. Jurica MS, Moore MJ: **Pre-mRNA splicing: awash in a sea of proteins.** *Mol Cell* 2003, **12:**5-14.
28. Barz T, Ackermann K, Dubois G, Eils R, Pyerin W: **Genome-wide expression screens indicate a global role for protein kinase CK2 in chromatin remodeling.** *J Cell Sci* 2003, **116:**1563-1577.
29. Knippschild U, Gocht A, Wolff S, Huber N, Löhler J, Stöter M: **The casein kinase 1 family: participation in multiple cellular processes in eukaryotes.** *Cell Signal* 2005, **17:**675-689.
30. Lebaron S, Froment C, Fromont-Racine M, Rain JC, Monsarrat B, Caizergues-Ferrer M, Henry Y: **The splicing ATPase prp43p is a component of multiple preribosomal particles.** *Mol Cell Biol* 2005, **25:**9269-9282.
31. Leeds N, Small E, Hiley S, Hughes T, Staley J: **The splicing factor Prp43p, a DEAH box ATPase, functions in ribosome biogenesis.** *Mol Cell Biol* 2006, **26:**513-522.
32. Schafer T, Maco B, Petfalski E, Tollervey D, Bottcher B, Aebi U, Hurt E: **Hrr25-dependent phosphorylation state regulates organization of the pre-40S subunit.** *Nature* 2006, **441:**651-655.
33. Bond A, Mangus D, He F, Jacobson A: **Absence of Dbp2p alters both nonsense-mediated mRNA decay and rRNA processing.** *Mol Cell Biol* 2001, **21:**7366-7379.
34. Edwards T, Saleem A, Shaman J, Dennis T, Gerigk C, Oliveros E, Gartenberg M, Rubin E: **Role for nucleolin/Nsr1 in the cellular localization of topoisomerase I.** *J Biol Chem* 2000, **275:**36181-36188.
35. Eppens NA, Rensen S, Granneman S, Raué HA, Venema J: **The roles of Rrp5p in the synthesis of yeast 18S and 5.8S rRNA can be functionally and physically separated.** *RNA* 1999, **5:**779-793.
36. Suzuki N, Noguchi E, Nakashima N, Oki M, Ohba T, Tartakoff A, Ohishi M, Nishimoto T: **The *Saccharomyces cerevisiae* small GTPase, Gsp1p/Ran, is involved in 3' processing of 7S-to-5.8S rRNA and in degradation of the excised 5'-A0 fragment of 35S pre-rRNA, both of which are carried out by the exosome.** *Genetics* 2001, **158:**613-625.
37. Boguta M, Hunter LA, Shen WC, Gillman EC, Martin NC, Hopper AK: **Subcellular locations of MOD5 proteins: mapping of sequences sufficient for targeting to mitochondria and demonstration that mitochondrial and nuclear isoforms commingle in the cytosol.** *Mol Cell Biol* 1994, **14:**2298-2306.
38. Gogarten JP, Olendzenski L, Hilario E, Simon C, Holsinger KE: **Dating the cenancester of organisms.** *Science* 1996, **274:**1750-1751. author reply 1751-1753
39. Rivera MC, Jain R, Moore JE, Lake JA: **Genomic evidence for two functionally distinct gene classes.** *Proc Natl Acad Sci USA* 1998, **95:**6239-6244.
40. Martin W, Müller M: **The hydrogen hypothesis for the first eukaryote.** *Nature* 1998, **392:**37-41.
41. Moreira , Lopez-Garcia : **Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis.** *J Mol Evol* 1998, **47:**517-530.
42. López-Garcia P, Moreira D: **Metabolic symbiosis at the origin of eukaryotes.** *Trends Biochem Sci* 1999, **24:**88-93.
43. Kurland CG, Collins LJ, Penny D: **Genomics and the irreducible nature of eukaryote cells.** *Science* 2006, **312:**1011-1014.
44. Forterre P: **Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain.** *Proc Natl Acad Sci USA* 2006, **103:**3669-3674.
45. Embley TM, vander Giezen M, Horner DS, Dyal PL, Foster P: **Mitochondria and hydrogenosomes are two forms of the same fundamental organelle.** *Philos Trans R Soc Lond B Biol Sci* 2003, **358:**191-201. discussion 201-202

46. Xin DD, Wen JF, He D, Lu SQ: **Identification of a Giardia krr1 homolog gene and the secondarily anucleolate condition of *Giaridia lamblia*.** *Mol Biol Evol* 2005, **22:**391-394.
47. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
48. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
49. Jorgensen P, Nishikawa JL, Breitkreutz BJ, Tyers M: **Systematic identification of pathways that couple cell growth and division in yeast.** *Science* 2002, **297:**395-400.
50. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters.** *Nat Genet* 2002, **31:**255-265.
51. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314:**1041-1052.
52. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, *et al.*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae.** *Nature* 2000, **403:**623-627.
53. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98:**4569-4574.
54. Gavin AC, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, *et al.*: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415:**141-147.
55. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415:**180-183.
56. Krogan NJ, Peng WT, Cagney G, Robinson MD, Haw R, Zhong G, Guo X, Zhang X, Canadien V, Richards DP, *et al.*: **High-definition macromolecular composition of yeast RNA-processing complexes.** *Mol Cell* 2004, **13:**225-239.
57. Kersey P, Bower L, Morris L, Horne A, Petryszak R, Kanz C, Kanapin A, Das U, Michoud K, Phan I, *et al.*: **Integr8 and Genome Reviews: integrated views of complete genomes and proteomes.** *Nucleic Acids Res* 2005:D297-302.
58. **FTP server of the Wellcome Trust Sanger Institute** [ftp://ftp.sanger.ac.uk/pub/databases/]
59. **GiardiaDB** [http://www.mbl.edu/Giardia]
60. **FTP server of The Institute for Genomics Research (TIGR)** [ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/]
61. Blair J, Ikeo K, Gojobori T, Hedges S: **The evolutionary position of nematodes.** *BMC Evol Biol* 2002, **2:**7.
62. Hutter H, Vogel B, Plenefisch J, Norris C, Proenca R, Spieth J, Guo C, Mastwal S, Zhu X, Scheel J, Hedgecock E: **Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes.** *Science* 2000, **287:**989-994.
63. Snel B, Bork P, Huynen M: **The identification of functional modules from the genomic association of genes.** *Proc Natl Acad Sci USA* 2002, **99:**5890-5895.
64. Wolf Y, Rogozin I, Grishin N, Tatusov R, Koonin E: **Genome trees constructed using five different approaches suggest new major bacterial clades.** *BMC Evol Biol* 2001, **1:**8.
65. Zdobnov E, von Mering C, Letunic I, Torrents D, Suyama M, Copley R, Christophides G, Thomasova D, Holt R, Subramanian G, *et al.*: **Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*.** *Science* 2002, **298:**149-159.
66. Dohm J, Vingron M, Staub E: **Horizontal gene transfer in aminoacyl-tRNA synthetases including leucine-specific subtypes.** *J Mol Evol* 2006, **63:**437-447.
67. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
68. de Hoon MJL, Imoto S, Nolan J, Miyano S: **Open source clustering software.** *Bioinformatics* 2004, **20:**1453-1454.
69. Saldanha AJ: **Java Treeview-extensible visualization of microarray data.** *Bioinformatics* 2004, **20:**3246-3248.
70. **YEAST PHYLPROF** [http://phylprof.molgen.mpg.de/cgi-bin/yeast_phylprof/yeast_phylprof.pl]
71. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, *et al.*: ***Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO).** *Nucleic Acids Res* 2002, **30:**69-72.
72. Güldener U, Münsterkötter M, Kastenmüller G, Strack N, van Helden J, Lemer C, Richelles J, Wodak SJ, Garcia-Martinez J, Perez-Ortin JE, *et al.*: **CYGD: the Comprehensive Yeast Genome Database.** *Nucleic Acids Res* 2005:D364-368.