SCIENTIFIC REPORTS

OPEN

Absence of N-terminal acetyltransferase diversification during evolution of eukaryotic organisms

Om Singh Rathore[1,2,3], Alexandra Faustino[1,2], Pedro Prudêncio[1,2,5], Petra Van Damme[6,7], Cymon J. Cox[4] & Rui Gonçalo Martinho[1,2,5]

Protein N-terminal acetylation is an ancient and ubiquitous co-translational modification catalyzed by a highly conserved family of N-terminal acetyltransferases (NATs). Prokaryotes have at least 3 NATs, whereas humans have six distinct but highly conserved NATs, suggesting an increase in regulatory complexity of this modification during eukaryotic evolution. Despite this, and against our initial expectations, we determined that NAT diversification did not occur in the eukaryotes, as all six major human NATs were most likely present in the Last Eukaryotic Common Ancestor (LECA). Furthermore, we also observed that some NATs were actually secondarily lost during evolution of major eukaryotic lineages; therefore, the increased complexity of the higher eukaryotic proteome occurred without a concomitant diversification of NAT complexes.

The genetic code is almost universal and the decoding molecular machine, the ribosome, is highly conserved across all known living organisms[1]. The narrow dimension of the ribosome exit tunnel (the cavity from which the nascent peptide emerges) precludes large domain folding of the nascent protein. This creates a window of opportunity for modification of protein residues that would be otherwise inaccessible due to folding. Indeed, co-translational modifications are widespread in cells throughout all three traditional kingdoms of life. Among others, these modifications include the proteolytic excision of the initial methionine and protein N-terminal acetylation (Nt-acetylation)[2–4], which involves the transfer of an acetyl group from acetyl-CoA to the protein alpha-amino group[4]. Although Nt-acetylation is an ubiquitous modification in eukaryotes, its prevalence varies, having a protein frequency of 50–70% in *Saccharomyces cerevisiae* (budding yeast), 70–80% in *Drosophila melanogaster* (fruit fly), and 80–90% in *Homo sapiens* (humans) and *Arabidopsis thaliana* (flowering plant)[5–9]. In eubacteria typically less than 10% of proteins are (partially) N-terminally acetylated, whereas in archaeal species it varies between 14–29% of all studied proteins[3,10].

Nt-acetylation may influence protein half-life[11–15], localization and export[16,17], protein-protein and protein-lipid interactions[18–22], the correct organization and function of the cellular cytoskeleton[23–25], nuclear chromatin[26,27], and vesicular compartment[28]. Its mis-regulation is frequently associated with tumor development and aggressiveness[29], and distinct human syndromes[30–33].

Nt-acetylation is catalyzed by a highly conserved family of N-terminal acetyltransferases (NATs). Prokaryotes have at least three NATs[3,34–37], whereas *H. sapiens* has six distinct but highly conserved NATs (NatA-F) (Fig. 1A). While some of these NATs are protein complexes requiring different catalytic and auxiliary subunits (e.g. NatA, NatB, and NatC), other NATs are able to Nt-acetylate independently of protein partners (e.g. NatD, NatF, and possibly NatE)[9,25,38–47]. NATs have distinct substrate specificity profiles, where substrate recognition depends on the identity of the first 2–5 amino acids of the elongating polypeptide[6,9,39,48–51].

[1]Department of Biomedical Sciences and Medicine, Faro, Portugal. [2]Center for Biomedical Research (CBMR), Faro, Portugal. [3]ProRegeM-PhD Program in Mechanisms of Disease and Regenerative Medicine, Faro, Portugal. [4]Center of Marine Sciences, University of Algarve, Faro, Portugal. [5]Instituto Gulbenkian de Ciência, Rua da Quinta Grande 6, Oeiras 2781-901, Portugal. [6]Department of Medical Protein Research, VIB, B-9000 Ghent, Belgium. [7]Department of Biochemistry, Ghent University, B-9000 Ghent, Belgium. Correspondence and requests for materials should be addressed to R.G.M. (email: rgmartinho@ualg.pt)
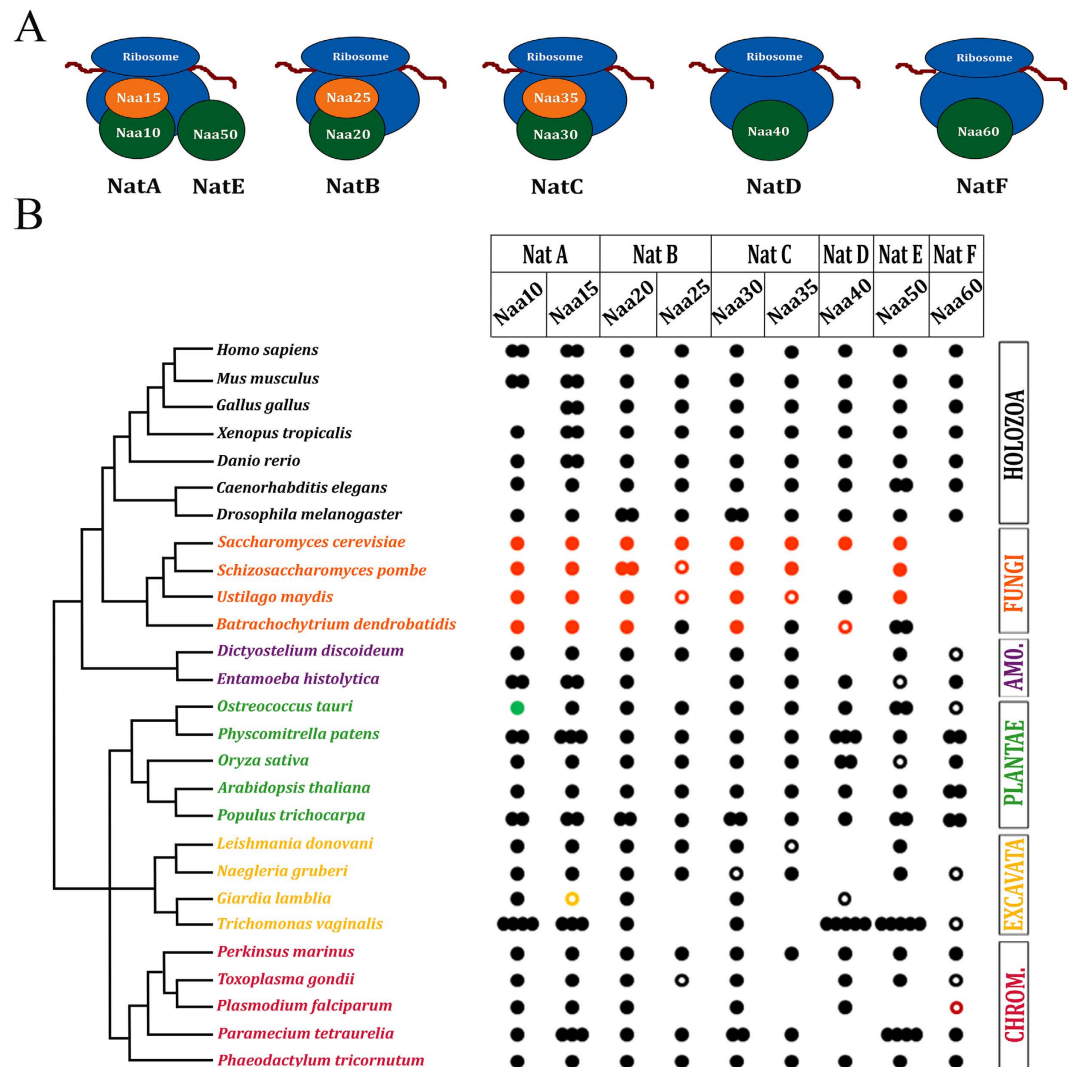
Figure 1. All six major human NAT complexes (NatA-F) were most likely present in the Last Eukaryotic Common Ancestor (LECA). (**A**) Subunits of all six major human NAT complexes (NatA-F). Catalytic subunits are shown in green, whereas regulatory subunits are shown in orange. (**B**) Catalytic and regulatory subunits of all six major human NATs complexes were identified across the eukaryotic tree of life, suggesting they were all present in the LECA. NATs subunit orthologs were identified in 27 species representative of the eukaryotic tree of life[59–63]. Naa60 (NatF) was apparently secondarily lost in fungi. Results are indicated according to reciprocal blastp E-value score ("filled dot" = E-value score lower than e^{-8}; "open dot" = E-value score between e^{-8}-e^{-03}; "no dot" = E-value score higher than e^{-03}). Black dot indicates NAT was identified using *H. sapiens* ortholog; orange dot indicates that NAT was identified using *S. cerevisiae* ortholog; green, yellow and red dots indicate that NATs were identified, respectively, using the phylogenetically closest plant, excavate and chromalveolata species ortholog. In the case of species-specific gene duplication, the number of dots is equivalent to the number of identified NAT paralogs. Phylogenetic distribution shown in this figure was previously reported[59–63]. Details of the original 73 analyzed eukaryotic species are shown in Supplementary Fig. 1 and Supplementary Table 1.

Nascent proteins are synthesized with a N-terminal methionine (also known as the initiator methionine or iMet), but if the second residue is non-bulky the iMet is frequently co-translationally removed by methionine aminopeptidases and the second residue is Nt-acetylated by NatA[6,49,50]. If the iMet is not excised, it can be Nt-acetylated by the other NATs. NatA and NatB are the major NATs in eukaryotic cells, which together Nt-acetylate approximately 60% of all *H. sapiens* proteins, while NatC, NatE, and NatF together Nt-acetylate only 15–20% of the proteome[9,52]. By contrast, the single archaeal NAT, possibly a direct ancestor of the eukaryotic NATs, has the ability to Nt-acetylate both NatA and NatE-type substrates of eukaryotes[34]. Such ancestral relationship implies the evolution of NAT substrate specialization and diversification in the eukaryote lineage.

Although absent in *S. cerevisiae*, NatF has been identified in *H. sapiens*, *D. melanogaster*, and more recently in *A. thaliana*[9,53]. NatF enzymatic activity is responsible for a significant increase in proteome Nt-acetylation when comparing *S. cerevisiae* to *H. sapiens*[9], and specifically targets transmembrane proteins[28]. Indeed, since the number and degree of Nt-acetylated proteins and NAT diversity is higher in *H. sapiens* and *D. melanogaster*

compared to *S. cerevisiae* and prokaryotes, it has been proposed that an increase in the regulatory complexity of this co-translational modification has occurred during evolution of higher eukaryotes[9,28]. Yet, the precise nature of these changes and their functional consequences remains poorly understood, as genome-wide studies across the eukaryotic tree of life are lacking.

In this work, we investigated the diversification of NATs during evolution of eukaryotic organisms. We concluded that most diversification of NATs happened before the evolution of eukaryotes, as our data strongly suggest that all six major human NATs were most likely present in the Last Eukaryotic Common Ancestor (LECA). Furthermore, we also observed that some NATs were secondarily lost during evolution of major eukaryotic lineages. Therefore, and although some clade-specific NAT duplications exist across the eukaryotic tree of life[54–56], the increased complexity of the higher eukaryotic proteome occurred without a concomitant diversification of the major NAT complexes.

Results

All six major human NATs (NatA-F) were most likely present in the Last Eukaryotic Common Ancestor (LECA).

Since the number of distinct NAT complexes and N-terminally acetylated proteins increased in *H. sapiens* and *D. melanogaster* when compared to *S. cerevisiae* and prokaryotes[3,9], we hypothesized that NATs had diversified during evolution of eukaryotic organisms. To test this hypothesis we selected 27 species representative of the major clades of eukaryotic tree of life: seven holozoa, four fungi, two amoebozoans, five plants, four excavates, and five chromalveolata. Bidirectional protein BLAST[57] searches and HMMER[58] gene model analyses were used to assess the presence or absence of the putative catalytic and regulatory subunits of the six major *H. sapiens* NAT complexes, or when necessary, NATs orthologs from more closely related species (e.g. *S. cerevisiae* for analysis of most fungi species). We considered NAT subunits orthologs on the basis of E-value of reciprocal best-hit in blastp (Fig. 1B; Supplementary Table 1). Detailed results for the originally analyzed 73 eukaryotic species are shown in Supplementary Fig. 1 and Supplementary Table 1.

Our analysis identified orthologs for the catalytic and regulatory subunits of all six NatA-F complexes (respectively, Naa10, Naa15, Naa20, Naa25, Naa30, Naa35, Naa40, Naa50, and Naa60) (Fig. 1A) in holozoa, amoebozoa, plantae, excavata, and chromalveolata (Fig. 1B). Consistent with our previous observation in *S. cerevisiae*[9], we failed to detect Naa60 in the four fungi species (Fig. 1B). When considering the wide phylogenetic distribution of the analyzed eukaryotic species (Fig. 1B)[59–63], we concluded that all major known *H. sapiens* NAT complexes were most likely present in the Last Eukaryotic Common Ancestor (LECA). Our results also suggested that Naa60 was secondarily lost in fungi (see below).

Identified NATs are most likely catalytically active.

If the identified NATs are catalytically active, the acetyl-CoA binding domain and catalytically active residues should be conserved among these proteins. To test this hypothesis we performed multiple sequence alignment of the protein sequences encoding the catalytic subunits of NATs, guided by the recently identified substrate-binding and catalytically active residues of *H. sapiens* Naa10 and Naa50[64,65]. Full protein alignment of distinct NAT catalytic subunit identified multiple highly conserved domains (data not shown). The acetyl coenzyme A binding motif, RxxGxG/A, is a sequence feature that is highly conserved among enzymes of the N-acyltransferase superfamily[66,67] and is highly conserved in most NATs (Fig. 2C). The catalytically active residues of Naa10 ($\alpha1$–$\alpha2$ loop 'E'; $\beta5$ helix 'R'; $\beta6$-7 helix 'Y') and Naa50 ($\beta4$ helix 'Y'; $\beta5$ helix 'H') ("inverted triangles" indicated on Fig. 2A,B)[64,65], and the substrate binding residues for Naa10 ($\alpha1$–$\alpha2$ loop 'L', 'E', and 'Y'; $\beta6$-7 helix 'Y') and Naa50 ($\alpha1$–$\alpha2$ loop 'Y'; $\beta4$ helix 'M'; $\beta5$ helix 'H'; $\beta6$-7 helix 'Y') ("plus signs" indicated on Fig. 2A,B)[64,65], were similarly highly conserved among most identified Naa10 and Naa50 orthologs.

In addition, all major functional domains of Naa10 and Naa50 ($\alpha1$–$\alpha2$ loop, $\beta4$, $\beta5$ and $\beta6$-7) were also highly conserved among Naa20, Naa30, and Naa60 (Fig. 2A,B). The only exception was Naa30, where the catalytically active arginine (R) in the $\beta5$ helix of Naa10 was replaced by a highly conserved glutamic acid (E) residue (Fig. 2A,B). Interestingly, the catalytically active glutamic acid (E) in $\alpha1$–$\alpha2$ loop and arginine (R) in the $\beta5$ helix were flanked by residues that varied specifically between Naa10, Naa20, and Naa30 (Fig. 2A,B). This confirmed that most identified NATs were correctly assigned to Naa10, Naa20, and Naa30, and suggested that this motif variation might be important for their distinct substrate specificities. Our results strongly suggest that most identified NATs are catalytically active N-terminal acetyltransferases. NATs whose canonical catalytically active residues vary from the *H. sapiens* orthologs are indicated in Supplementary Fig. 2.

Clade-specific NAT loss during evolution of eukaryotic organisms.

Although all major NAT complexes were most likely present in the LECA, we previously observed that Naa60 (the catalytic subunit of NatF) was absent in *S. cerevisiae*[9]. To test whether NatF is likely absent in all fungi, we analyzed 13 species representative of the fungal kingdom, and confirmed that Naa60 orthologs are absent (Fig. 3). The wide phylogenetic distribution of the analyzed fungal species suggests that Naa60 was most likely absent in the progenitor of all fungi.

We also investigated the genomes of species in the phylum microspora, as they are eukaryotic unicellular organisms and obligatory intracellular parasites that exhibit extreme genome reduction and gene loss[68]. We detected orthologs for Naa10, Naa20, and Naa50 in most microsporidia, yet we failed to identify orthologs for Naa30, Naa40, and Naa60 (Fig. 3). Loss of Naa40 and Naa60 was also observed within some but not all excavata (another large group of unicellular organisms) (Fig. 3). It is therefore evident that although all six NATs present in *H. sapiens* were most likely present in the LECA, some NATs have been lost during evolution of major eukaryotic lineages.

Surprisingly, in birds, we failed to identify Naa10 in galliformes (e.g. *Gallus gallus* (chicken) and *Meleagris gallopavo* (turkey)), passeriformes (e.g. *Taenopygia guttata* and *Ficedula albicollis*), and psittaciformes (e.g. *Melopsittacus undulatus* (budgerigar)) (Supplementary Fig. 4). However, given a) the extent of
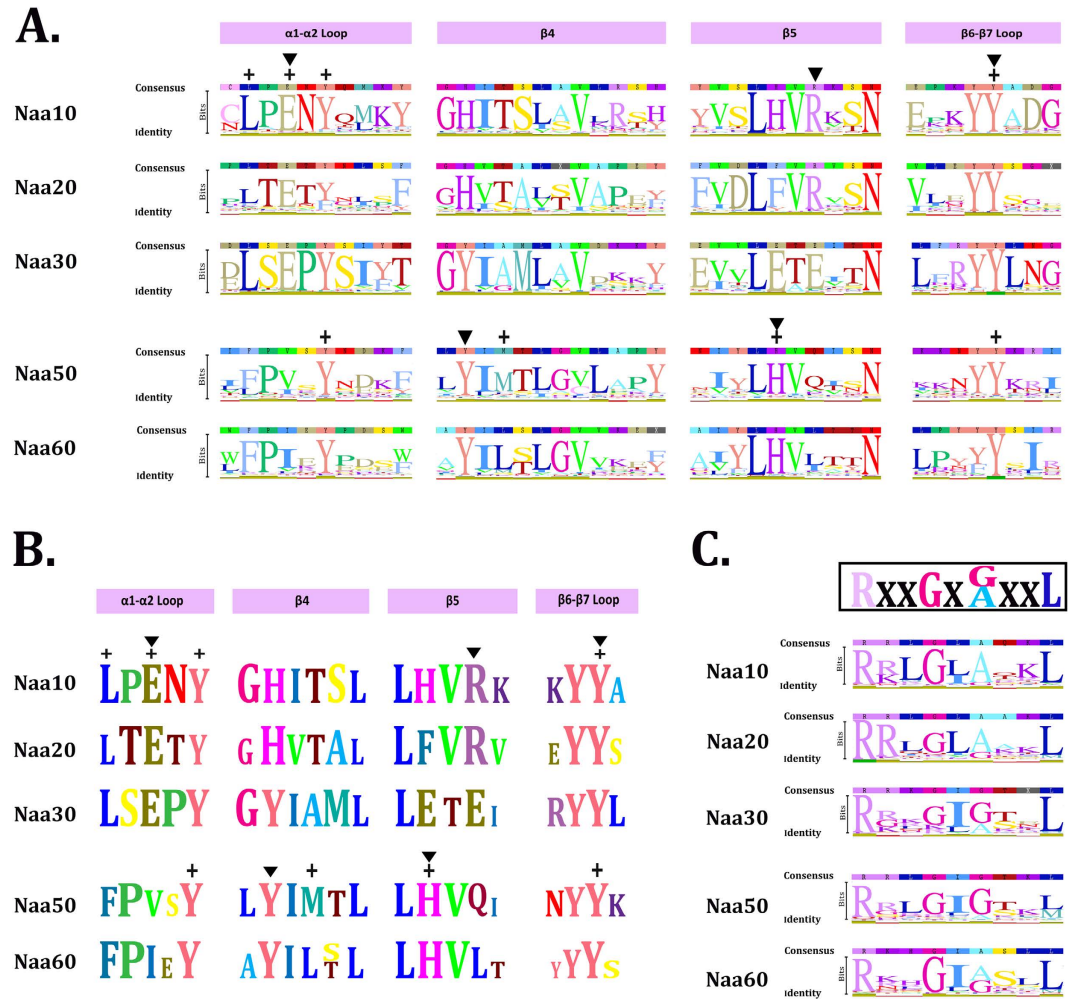
Figure 2. Identified NATs are most likely catalytically active. Identified NATs are most likely catalytically active as the acetyl-CoA binding domain[66,67] and the catalytically active residues from *H. sapiens* NAT´s[64,65] are highly conserved among these proteins. Protein sequences were retrieved from the genomes of 73 species representative of the eukaryotic tree of life and aligned. Letters height represents the degree of conservation of each amino acid residue for that position. The catalytically active residues (▼) and substrate binding residues (+) for Naa10 and Naa50 are indicated above each sequence alignment[64,65]. (**A** and **B**) The catalytically active residues of Naa10 (α1–α2 loop 'E'; β5 helix 'R'; β6-7 helix 'Y') and Naa50 (β4 helix 'Y'; β5 helix 'H') are highly conserved among most identified NATs. The substrate binding residues for Naa10 (α1–α2 loop 'L', 'E', and 'Y'; β6-7 helix 'Y') and Naa50 (α1–α2 loop 'Y'; β4 helix 'M'; β5 helix 'H'; β6-7 helix 'Y') are also highly conserved among most identified NATs. The substrate binding and catalytically active residues of Naa10 are similarly highly conserved in Naa20 and Naa30. The only exception is the catalytically active arginine (R) in the β5 helix, which was replaced in Naa30 by a highly conserved glutamic acid (E) residue. The catalytically active glutamic acid (E) in α1–α2 loop and arginine (R) in the β5 helix are flanked by residues that vary specifically within Naa10, Naa20, and Naa30. (**C**) The acetyl coenzyme A binding motif, RxxGxG/A, which is highly conserved among enzymes of the N-acyltransferase superfamily[66,67], is similarly conserved among most identified NATs.

conservation of Naa10 across the eukaryotic tree of life (Fig. 1B; Supplementary Fig. 1), b) NatA's broad role in protein Nt-acetylation, c) the identification of the Naa10-interacting subunit Naa15 (Supplementary Fig. 1; Supplementary Fig. 4), d) the identification of methionine aminopeptidases (MetAP) in all tested bird species (data not shown), and e) that Naa10 is present in some other birds (e.g. *Falco cherrug* (falcon) and *Anas platyrhynchos* (duck)) (Supplementary Fig. 1; Supplementary Fig. 4), we concluded that the absence of Naa10 in some birds is probably due to an unknown gap in the publicly available genome data. Undetectable Naa10 homologs due to rapid sequence divergence is not likely an explanation, as the protein sequences of the two identified birds' Naa10 were highly conserved (Supplementary Table 1).

NatE interacts and influences *in vivo* NatA catalytic activity. Naa10 and Naa20 were never or rarely lost during eukaryotic evolution (Fig. 1; Supplementary Fig. 1), even in organisms like microsporidia with extensive gene loss (Fig. 3)[68], possibly because these two NATs Nt-acetylate the majority of the proteome[9,52], likely
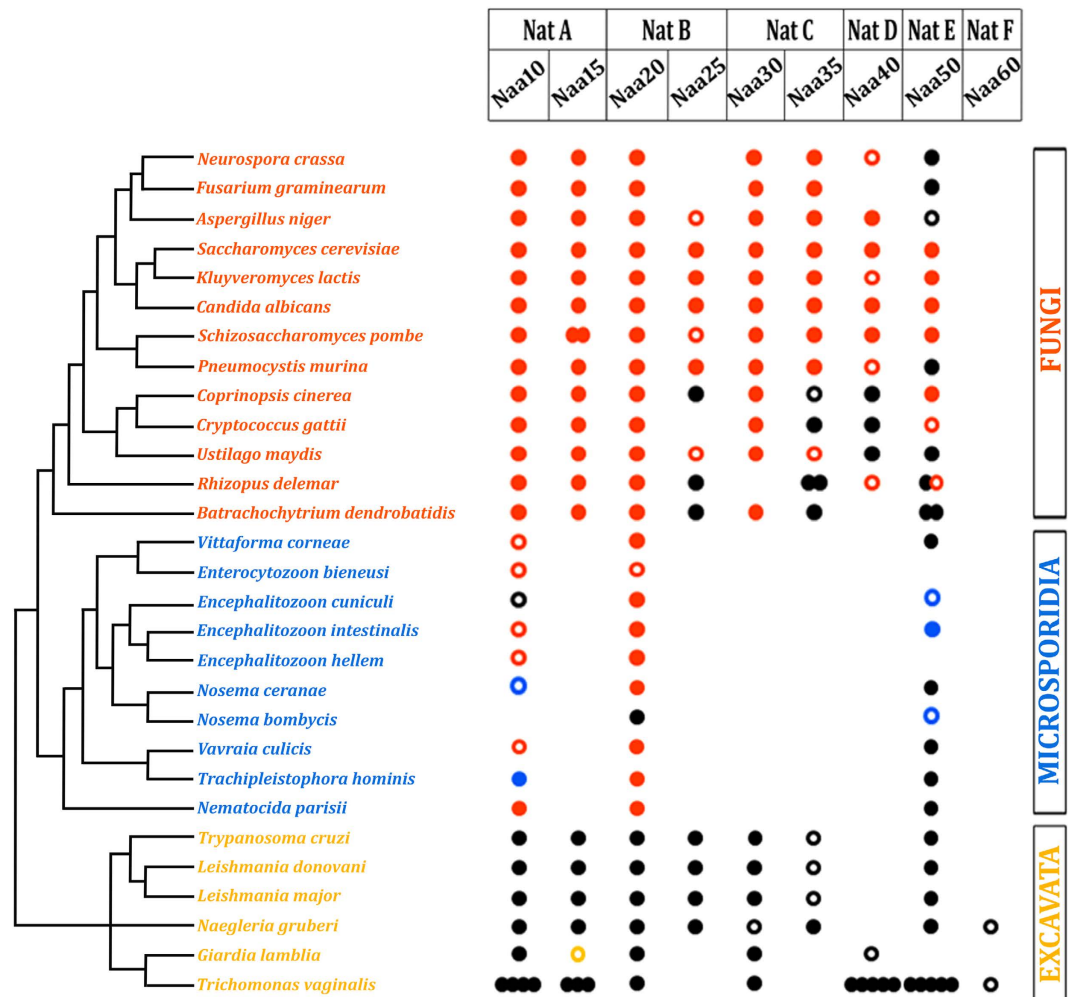
Figure 3. Clade-specific NAT loss during evolution of eukaryotic organisms. NAT complexes have been secondarily lost during evolution of major eukaryotic lineages. Naa10 and Naa20 were never or seldom lost during eukaryotic evolution (Fig. 1) (Supplementary Fig. 1), even in organisms like microsporidia with extensive gene loss[68]. Naa50 was comparatively more resilient to loss in distinct eukaryotic lineages (e.g. microspora and excavata) than Naa30, Naa40, and Naa60. Analysis of 13 species representative of the fungal kingdom shows that Naa60 is absent. The wide phylogenetic distribution of the fungal taxa suggests that Naa60 was most likely absent in the progenitor of all fungi. Results are indicated according to reciprocal blastp E-value score ("filled dot" = E-value score lower than e^{-8}; "open dot" = E-value score between e^{-8}-e^{-03}; "no dot" = E-value score higher than e^{-03}. Black dot indicates NAT was identified using *H. sapiens* ortholog; orange dot indicates that NAT was identified using *S. cerevisiae* ortholog; blue and yellow dots indicate that NATs were identified, respectively, using the phylogenetically closest microsporidia and excavate species ortholog. In the case of species-specific gene duplication, the number of dots is equivalent to the number of identified NAT paralogs. Phylogenetic distribution shown in this figure was previously reported[59–63].

making them indispensable. Surprisingly, Naa50 was comparatively more resilient to loss in distinct eukaryotic lineages than Naa30, Naa40, and Naa60 (Fig. 3; Supplementary Fig. 1).

Drosophila Naa50 is encoded by the gene separation anxiety (san) and is required for sister chromatid cohesion and chromosome segregation during mitosis[69,70]. Consistently with previous reports[69,71,72], Naa50 (~20kDa) physically interacted with NatA (Naa10 and Naa15; respectively ~22kDa and ~103kDa) (Fig. 4A,B). Naa10 and Naa50-containing complexes of approximately 150-350kDa were identified in wild-type *Drosophila* embryos (Fig. 4D), next to the detection of significant levels of monomeric Naa50. There was however no notable change in the apparent size of the Naa10-containing complexes in protein extracts from san[3] mutant embryos (Fig. 4C versus 4D), which suggested that NatA complex integrity was not affected by loss of Naa50.

The NatA-interacting chaperone-like protein HYPK[73] was also identified as a subunit of the immunoprecipitated Naa50-containing complex (Fig. 4B), suggesting that Naa50 interacts (directly or indirectly) not only with Naa10-Naa15 but also with a subset of other NatA-interacting proteins. This supports the hypothesis that the relative levels of ribosome associated Naa50/NatA and MetAPs (iMet-aminopeptidases) are important for the correct processing of the nascent polypeptide N-termini[74] and further confirms the functional crosstalk between these two NAT complexes.
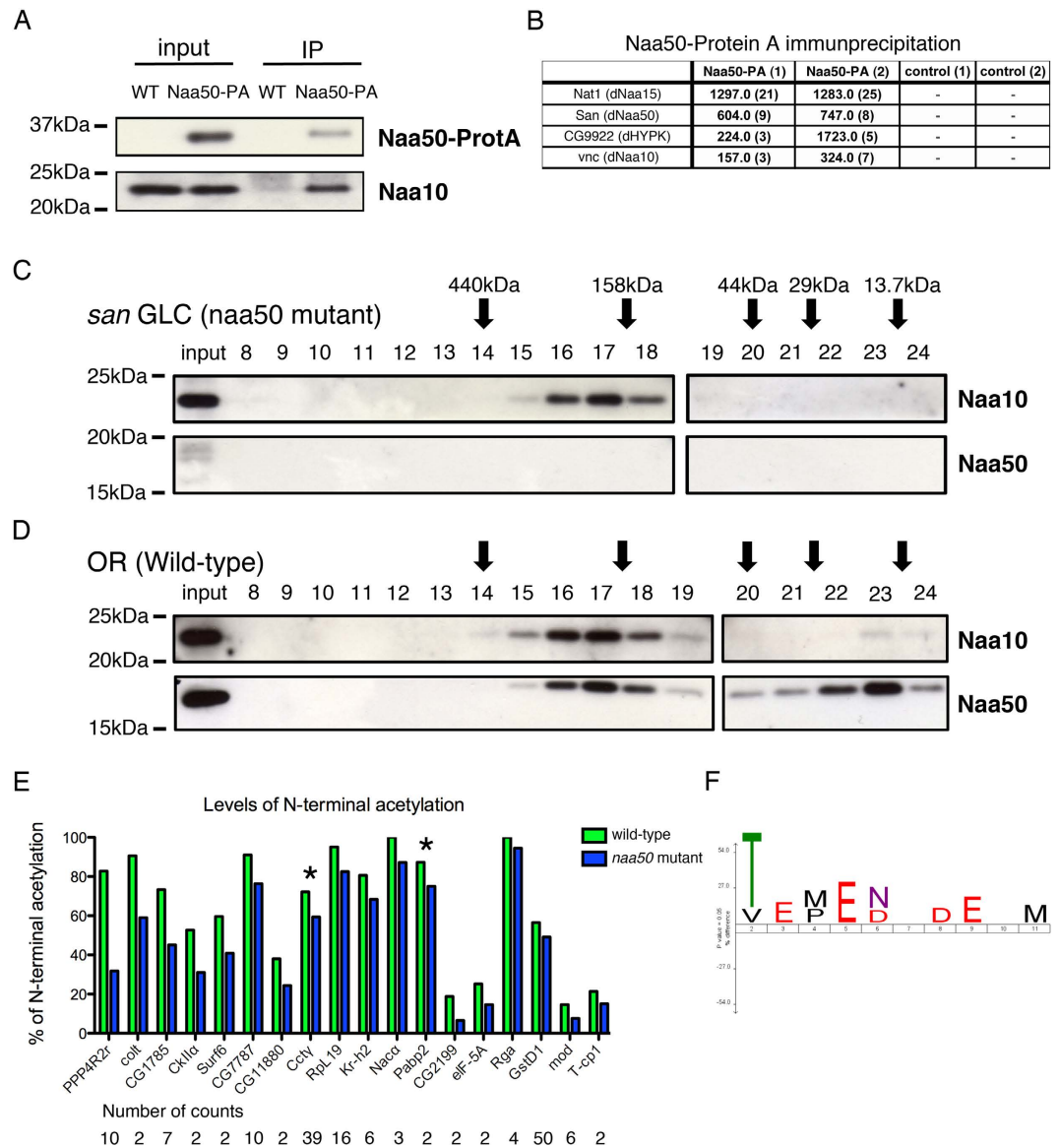
Figure 4. NatE interacts and influences *in vivo* NatA catalytic activity. *Drosophila* Naa50 is encoded by the gene separation anxiety (san). Naa50 physically interacts with NatA, and its loss influences the *in vivo* activity of NatA. (**A** and **B**). Naa50 (~20 kDa) physically interacts with NatA complex (Naa10 and Naa15) and HYPK/CG9922; respectively ~22 kDa, ~103 kDa, and ~14 kDa). Immunoprecipitated proteins were identified by western-blot (**A**) or by liquid chromatography coupled to tandem mass spectrometry (**B**). (**C** and **D**). Integrity of Naa10-containing complexes was apparently not affected after loss of Naa50. Size-exclusion chromatography of *san*[3] mutant (**C**) and control (**D**) protein extracts from 0–2 h embryo collections. After separation, each fraction was analyzed by western blot. Naa10-containing complexes with approximately 150–350 kDa were similarly identified in wild type and *san*[3] mutant embryos (**D** versus **C**). There were significant levels of monomeric Naa50 (~20 kDa) in wild-type embryos (**D**). (**E** and **F**) Loss of Naa50 negatively influences the *in vivo* activity of NatA. The proteome of control and *san*[3] maternal mutant embryos (0–2h collection) was analyzed by N-terminal COFRADIC[75,89]. (**E**) The N-terminal acetylation status of 265 unique N-termini was unequivocally determined, and the affected N-termini are shown for control (green bars) and *san*[3] mutant embryos (blue bars). (**F**) Consensus sequence of affected N-termini suggested that most displayed a NatA-like substrate specificity (i.e. predominantly Thr-, Val-, Ser-, and Ala-starting N-termini). Only two of the affected N-termini started with a methionine (MF- and MQ-) (see asterisks indicated in panel E), which is consistent with the *H. sapiens* Naa50 substrate specificity[39,74].

Since the physical interaction between NatA and NatE (Naa50) is conserved between *D. melanogaster* and *H. sapiens*, and in line with the recent observation that six canonical NatA substrates were less Nt-acetylated in *S. cerevisiae* lacking Naa50[74], we hypothesized that Naa50 was infrequently lost during eukaryotic evolution because its absence partially inhibits NatA enzymatic activity. To further support this hypothesis, we performed N-terminal
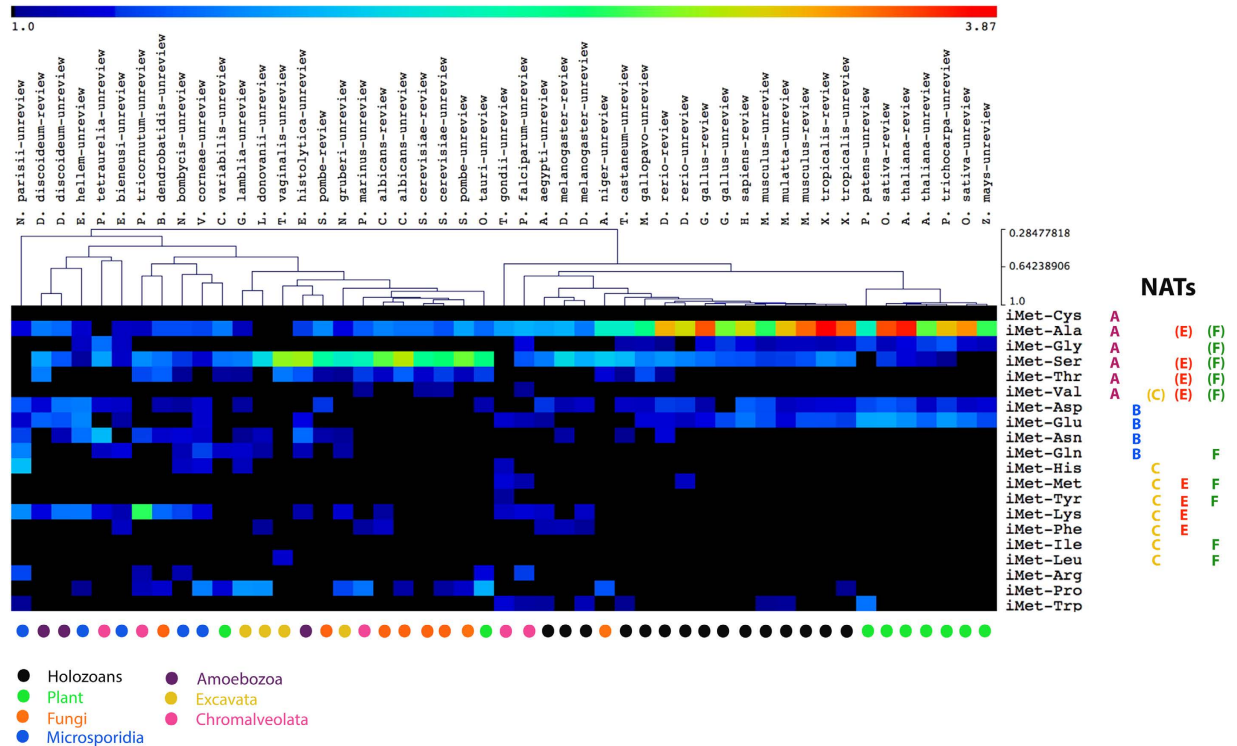
Figure 5. No detectable proteome adaptation to NAT loss. There are major biases in the frequency of amino acid usage at the N-terminal second position when compared to total proteome. However there is no detectable proteome adaptation to NAT loss. Loss of NatC and NatF in microsporidia (blue dots) and in some excavata (yellow dots) (Fig. 3) does not correlate with an increase usage frequency of NatA and NatB substrates. Comparatively to holozoans (black dots) and fungi (orange dots), microsporidia show an under-representation of alanine and serine residues (NatA substrates), no obvious over-representation of NatB substrates, and comparatively to higher eukaryotes, a moderate over-representation of lysine (NatC and NatE substrates). Amino acid usage frequency bias for each N-terminal position was analyzed by calculating the amino acid usage frequency for each position divided by its frequency in the total proteome (for more experimental detail see material and methods). The fold enrichment heat map shows the over-representation range (1 to 3.87 fold enrichment) of each amino acid at the N-terminal second position when compared to the total proteome. For each species and for each amino acid, it was attributed a black color (≤ 1.0) when the amino acid is under-represented in the N-terminal second position compared to its total proteome usage frequency. A detailed breakdown of the values used in this heat map is shown in Supplementary Table 2.

COmbined FRActional DIagonal Chromatography (COFRADIC) analysis[75] on the proteomes isolated from control and *san*[3] maternal mutant *Drosophila* embryos to assay their Nt-acetylomes, and to look for differences in N-terminal acetylation states. The *in vivo* N-terminal acetylation levels of 265 proteins were unequivocally determined in the proteomes of both control and *san*[3] mutant embryos (Fig. 4E,F). 18 protein N-termini were identified as more Nt-acetylated in the control proteome expressing Naa50 as compared to the *san* mutant proteome (absence of Naa50). Consistent with our hypothesis, 16 out of 18 of the *san* mutant affected N-termini displayed NatA-like substrate specificity (Fig. 4E,F). Only two affected N-termini started with a methionine (MF- and MQ- starting N-termini) (see asterisks on Fig. 4E), the presumed *H. sapiens* Naa50 substrate specificity[39,74]. This suggests that loss of Naa50 in *D. melanogaster* embryos was associated *in vivo* with a reduction of NatA activity, which fully supports the hypothesis that Naa50 was infrequently lost during eukaryotic evolution because its absence partially impairs NatA enzymatic activity.

No detectable proteome adaptation to NAT loss.

NAT substrate specificity is largely defined by the identity of the substrate N-terminal second amino acid residue[6,9,39,48–51,76]. Since different NATs have distinct substrate specificities, and since these enzymes Nt-acetylate a large number of proteins, we tested whether the loss of NatC, NatD, and NatF in microsporidia was associated with a detectable decrease of their substrate N-termini when compared to NatA and NatB substrate N-termini. To investigate if there was such proteome-wide adaptation to NAT loss, we analyzed the amino acid frequency in N-terminal second, third, fourth, fifth, and sixth residue positions for 36 species representative of the eukaryotic tree of life, including 5 microsporidia species.

Previously, it was observed that the amino acid preference for the N-terminal second position varies, being lysine the most common amino acid residue in prokaryotes, serine the most common for lower eukaryotes, and alanine the most common for animals[9,52,77,78]. Consistently, we identified major biases in the frequency of amino acid usage at the N-terminal second position within distinct eukaryotic clades, including vertebrates and plants (Fig. 5; Supplementary Fig. 5; Supplementary Table 2). In vertebrates and plants, alanine residues (NatA

substrate[6,7,50]) were significantly over-represented in the N-terminal second position when compared to the total frequency of usage of this amino acid in the proteome (Fig. 5; Supplementary Table 2), whereas serine residues (NatA substrate[6,7,50]) were significantly over-represented in the N-terminal second position of fungi and excavates (Fig. 5; Supplementary Table 2). Invertebrates showed instead an over-representation of both alanine and serine residues in the N-terminal second position (Fig. 5; Supplementary Table 2)[9]. Other amino acid residues like leucine, although commonly used in the entire proteome[52], were nevertheless significantly under-represented at this position (data not shown). The N-terminal residue usage frequency biases were significantly less pronounced for the third, fourth, fifth and sixth N-terminal amino acid residues (Supplementary Fig. 5; Supplementary Table 2).

The N-terminal second residue position frequency bias observed in microsporidia does not support the hypothesis of a proteome-wide adaptation to loss of NatC, NatD, and NatF, with an increase frequency of NatA and NatB substrate N-termini. Compared to higher eukaryotes, microsporidia showed an under-representation of alanine and serine (NatA substrates), no obvious over-representation of NatB substrates, and a moderate over-representation of lysine (NatC and NatE substrates[9,23,25,39,76]). The moderate over-representation of lysine within these microorganisms is most likely not related to NAT loss, as the diatom *Phaeodactylum tricornutum* (chromalveolata), for example, showed an unusual over-representation of lysine within the N-terminal second position (Fig. 5) without NAT loss (Fig. 1). Some microsporidia, amoebozoa, chromalveolata, and excavata microorganisms also showed a moderate increase in the usage frequency of asparagine and glutamine residues (NatB substrates) in the N-terminal second position (Fig. 5). This over-representation is most likely also not related with NAT loss as *Paramecium tetraurelia* (chromalveolata), for example, showed a significant increase of these two residues (Fig. 5) without extensive NAT loss (Fig. 1). Consistently, differences in *S. cerevisiae* N-terminal sequences could not explain the absence of NatF in this unicellular organism[9].

We concluded that although the N-terminal second residue position is under significant clade-specific constraints when compared to other nearby residues, there is no evidence supporting the hypothesis of a proteome-wide adaptation of proteins N-termini to NAT loss. The extensive loss of NATs might be accommodated by an increased substrate redundancy of the remaining NATs in microsporidia. Alternatively, it is also possible that Nt-acetylation is not functionally rate limiting for most proteins.

Discussion

The origin of eukaryotes is still poorly understood. The organizational complexity of a eukaryotic cell is significantly higher than any known prokaryote. Eukaryotic cell compartmentalization is supported by an elaborate endomembrane system and by an actin/tubulin-based cytoskeleton. Conservation of the major features of eukaryotic cell organization and of a large set of genes demonstrates that they were already present in the Last Eukaryotic Common Ancestor (LECA)[79]. Protein Nt-acetylation is an ancient and highly conserved protein modification. Although the number of NAT complexes significantly increased in eukaryotic cells when compared to prokaryotic organisms, our work demonstrates that all major human NAT complexes diversified before the evolution of eukaryotic organisms, as they were already present in LECA.

Clade-specific NAT duplications exist across the eukaryotic tree of life[54–56]. In mammals, gene duplications of both Naa10 and Naa15 resulted in Naa11 and Naa16 paralogs, respectively[54,55]. Both gene paralogs encode proteins that are likely to participate in functional NatA complexes. Although human Naa11 and Naa16 share 81% and 70% sequence identity with Naa10 and Naa15, respectively[54,55], Naa16 expression is particularly high on certain tissues like adrenal gland, mammary gland, heart, thymus, and testis[55]. Expression of mouse Naa11 is similarly restricted, being particularly high during spermatogenesis[80]. Since mouse Naa10 locates on chromosome X, Naa11 expression is likely to have a compensatory function after the loss of Naa10 expression (approximately in 50% of the sperm) during meiosis. Therefore, and although the increased complexity of the higher eukaryotic proteome clearly occurred without a concomitant diversification of the major known NAT complexes, clade-specific gene duplications likely happened to address specific functional requirements.

It is possible to argue that the NAT substrate repertoire has not been significantly increased during eukaryotic evolution, as these enzymes mostly rely on the identity of the substrate N-terminal second residue for specificity. Nevertheless, and next to purely housekeeping functions for Nt-acetylation, the regulatory complexity of this co-translational modification is likely to have increased in higher eukaryotes as their proteome and cellular behavior became more complex. Consistently, detectable co-evolution between NATs substrate specificities and the eukaryotic proteome has recently been reported in humans for NatA and NatD[47,81]. In the absence of NAT diversification, we speculate that transcriptional and post-transcriptional regulation of NAT activity are likely to play increasingly important roles in the regulation of Nt-acetylation during development of multicellular organisms[82].

Material and Methods

Identification of NAT orthologs. Reference protein sequences of NAT complexes subunits (Naa10, Naa15, Naa20, Naa25, Naa30, Naa35, Naa40, Naa50 and Naa60) in *H. sapiens* and *S. cerevisiae* were obtained through literature mining[9,83]. These sequences, or sequences from closely related species, were used to identify putative orthologs from 73 species (Supplementary Fig. 1) representative of the eukaryotic tree of life[59–63]. Publicly available genome databases (e.g. NCBI and Ensembl) were used to retrieve these sequences. Two steps were used to verify the identity of these putative orthologs: we used reciprocal bidirectional protein BLAST (blastp)[57]. The retrieved proteins were only considered *bona fide* orthologs if they corresponded to the blastp best hit in both directions. The results were subjectively divided in three classes: "filled dot": reciprocal blastp E-value score lower than e^{-8}; "empty dot": reciprocal blastp E-value score between e^{-8}- e^{-03}; "no dot": reciprocal blastp E-value scores higher than e^{-03} (Fig. 1B, Fig. 3, Supplementary Fig. 1–4, Supplementary Table 1). To identify apparently missing ortholog sequences, additional searches were performed using protein sequence queries from phylogenetically closely related organisms. For example, *Populus trichocarpa* (seed plant) protein sequences were used for ortholog

identification in the closely related (in the context of eukaryotes) *Osteococcus tauri* (green algae). To confirm the bidirectional blastp results, we built a separate hidden markov model for each NAT protein using HMMER 3.0[58]. Multiple sequence alignments of ortholog protein sequences positively identified by reciprocal BLAST searches were used to calculate hidden markov models from GenBank non-redundant (nr) protein database (downloaded August 2013). These HMM profiles allowed us to review the fit of the sequence to our orthologue classification.

Sequence alignment and functional domain arrangement. For the identification of conserved domains within NAT catalytic subunits all identified orthologs of Naa10, Naa20, Naa30, Naa50, and Naa60 were retrieved from 73 eukaryotic species and aligned using ClustalW[84] program in the Geneious software (version 6.1.8) with default values. Sequence alignments were manually edited by removing ambiguously aligned sequences and gaps in Gblocks[85]. Regions in the alignment with more than 50% gaps were removed. Four domains α1–α2 loop, β4 helix, β5 helix and β6-7 helix were previously shown to be catalytically important in NATs[64,65]. These highly conserved domains were used to validate the alignment. The three catalytically active residues of Naa10 (α1–α2 loop 'E', β5 helix 'R' and β6-7 helix 'Y)' were identified and marked as (▼) (Fig. 2A,B). Substitution of the canonical catalytically active residues for alternative residues is shown in Supplementary Figs 2 and 3. The two catalytically active residues of Naa50 (β4 helix 'Y' and β5 helix 'H') were marked as (▼) (Fig. 2A,B). The acetyl coenzyme A (AcCoA) binding motif[66,67], RxxGxG/A in NATs was manually assigned in the aligned sequences (Fig. 2C).

Fly work and genetics. *Drosophila melanogaster* flies were raised using standard techniques[86]. san[3] mutant allele was previously reported as a loss-of-function allele of the gene separation anxiety (san)[70], which encodes *Drosophila* Naa50. Maternal mutant embryos of *san*[3] were generated using the FLP/FRT ovo[D] system[87].

The Naa50 open-reading frame (from cDNA clone AT27602) was subcloned into a vector containing the UASp promoter and C-terminal Protein A-tag (Gateway, Life Technologies). All constructs were then used to generate transgenic fly stocks (BestGene).

Protein extracts. Collections of 0–2 hr (after egg-laying) *Drosophila* embryo were washed in PBS + 0.1% tween 20, dechorionated with 50% of bleach (commercial solution) for 5 minutes, and thoroughly washed in water. Protein extraction was performed through homogenization of embryos in NB lysis buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 2 mM EDTA, 0.1% NP-40), 1 mM DTT, 10 mM NaF, and EDTA-free protease inhibitor cocktail (Roche, Germany). Extracts were centrifuged at 4 °C and 21,100 g for 3 minutes. Supernatant was collected (being the upper lipid-rich layer avoided as much as possible) and centrifuged twice more. Total protein concentration was determined using Bio-Rad Bradford protein assay (BioRad, Hercules, CA, USA).

Antibodies. Primary antibodies used were rabbit anti-San/Naa50 (1:1000)[69] and rabbit anti-Ard1/Naa10 (1:2000) (Santa Cruz; FL-235)[88]. Secondary detection in western-blot was performed with rabbit HRP-conjugated antibodies (Jackson ImmunoResearch) used at a final concentration of 1:5000.

Co-Immunoprecipitation. Total proteins extracts of *Drosophila* embryos expressing Naa50/San fused to a C-terminal Protein A-tag (Naa50/San-PA) were obtained. Dynabeads M-270 Epoxy (Invitrogen, Grand Island, NY, USA) pre-incubated with rabbit IgG immunoglobulins (5 μg IgGs/ 1 mg of beads) (Dynabeads antibody coupling kit, 14311D, Invitrogen), were added to 1.5 mg of protein extract and incubated for 1 hour at 4 °C. After washing the beads 3 times with NB buffer, protein elution was performed with 100 μl of 100 mM Glycine pH 3.0 during 1 minute and stopped with 10 μl of 1 M Tris Base pH 10.8. Eluted proteins were precipitated at −20 °C using 5 volumes of acetone. Precipitated samples were analyzed by liquid chromatography coupled to tandem mass spectrometry (Mass Spectrometry Laboratory, Institute of Biochemistry and Biophysics, Poland). For western-blot analysis eluted proteins were boiled in SDS-sample buffer for 5 minutes.

Size exclusion chromatography. Size-exclusion chromatography was performed by applying protein extracts of 0–2 h collections from FRT42B *san*[3] mutants or control (Oregon R) *Drosophila* embryos. Extracts were prepared in NB2 buffer (150 mM NaCl, 50 mM Tris-HCl pH 7.5, 2 mM EDTA, 0.01% NP-40), 1 mM DTT and EDTA-free protease inhibitor cocktail (Roche). Subsequently, 2 mg of protein extract were fractionated using Superose 6 10/300 GL column (GE Healthcare) in NB2 buffer and analyzed by western blot using antibodies against Naa10/Ard1[88] and Naa50/San.

Mass spectrometry. Liquid chromatography coupled to tandem mass spectrometry was performed at the Mass Spectrometry Laboratory (Institute of Biochemistry and Biophysics, Poland).

Briefly, peptides mixtures were analyzed by LC-MS-MS/MS (liquid chromatography coupled to tandem mass spectrometry) using Nano-Acquity (Waters, Milford, MA, USA) LC system and Orbitrap Velos mass spectrometer (Thermo Electron Corp., San Jose, CA, USA). Prior to analysis, proteins were subjected to standard 'in-solution digestion' procedure, during which proteins were reduced with 100 mM DTT (for 30 minutes at 56 °C), alkylated with 0.5 M iodoacetamide (45 minutes in darkroom at room temperature), and digested overnight with trypsin (Sequencing Grade Modified Trypsin-Promega V5111). The peptide mixture was applied to an RP-18 precolumn (nanoACQUITY Symmetry C18-Waters 186003514) using water containing 0.1% TFA as mobile phase, then transferred to nano-HPLC RP-18 column (nanoACQUITY BEH C18-Waters 186003545) using an acetonitrile gradient (0%–35% AcN in 180 min) in the presence of 0.05% formic acid with a flow rate of 250 nl/min. The column outlet was directly coupled to the ion source of the spectrometer, operating in the regime of data dependent MS to MS/MS switch. A blank run ensuring no cross contamination from previous samples preceded each analysis.

Raw data were processed by Mascot Distiller followed by Mascot Search (Matrix Science, London, UK, on-site license) against Flybase database. Search parameters for precursor and product ions mass tolerance were 15 ppm and 0.4 Da, respectively, enzyme specificity: trypsin, missed cleavage sites allowed: 0, fixed modification of cysteine by carbamidomethylation, and variable modification of methionine oxidation. Peptides with Mascot Score exceeding the threshold value corresponding to <5% false positive rate, calculated by Mascot procedure, and with the Mascot score above 30 were considered to be positively identified.

N-terminal COFRADIC analysis. The proteome of control (Oregon R) and *san*[3] maternal mutant *Drosophila* embryos (0–2 h collection) were analyzed by N-terminal COFRADIC analysis as described previously[75,89]. Overall, 432 unique N-termini derived from 399 unique SwissProt *Drosophila* protein accessions were identified. N-termini that partially retain their initiator-Met and/or have most probably alternative or miss-annotated translation products led to the higher number of unique N-termini as compared to their corresponding identified accession. Quantification of the degree of Nt-acetylation was performed as described[9]. A significant variation in the degree of Nt-acetylation was set to 10% or more (p ≤ 0.01). Further, N-termini were considered to be affected by *san* deletion that were fully acetylated in the control strain and displayed a reduction of more than 5% in the *san* deletion strain were also considered.

As such, the N-terminal acetylation states of the N-termini identified in both setups (292 unique N-termini) were comparatively analyzed. In this study we distinguished between *in vivo* N-terminal acetylation and free N-termini by making use of *in vitro* AcD3C13-acetylation (i.e. 5Da heavier form of NHS-acetate which introduces an 13C2 and D3 labeled acetyl moiety on all free amines).

An N-terminus is defined as 1) a peptide that is *in vivo* N-terminal acetylated or *in vitro* AcD3C13-acetylated peptide (i.e. an *in vivo* free N-terminus) and starts in the Swiss-Prot database at position 1 or 2 of the annotated protein sequence, or 2) an *in vivo* N-terminal acetylated peptide, with a protein sequence starting position beyond position two. Only peptides displaying an internal start position and of which at least one N-terminal acetylated peptide was identified in one of the two setups analyzed were considered (i.e. internal AcD3C13-acetylated peptides were not considered in this study).

Overall the N-terminal acetylation levels of 265 out of the 292 unique N-termini identified in both setups could be unequivocally determined in control as well as the *san* mutant setup, and for the affected N-termini, the difference (Δ) in the N-terminal acetylation status of the control versus *san* mutant setup was analyzed and is shown (Fig. 4E). In total, 10 unique N-termini (start 1/2) and 4 N-termini (internal start position) displayed in san mutant embryos a decrease of at least 10% in the overall N-terminal acetylation. 12 out of 14 (or 87%) of the *san* mutant affected N-termini displayed NatA-like substrate specificity (i.e. predominantly Thr (8/14 or 57%), Ser, Val and Ala-starting N-termini (Fig. 4F). Two affected N-termini started with a methionine (MF- and MQ- starting N-termini) (see "stars" in Fig. 4E). 4 unique N-termini displayed a decrease of more than 5% (but less than 10%) in the overall N-terminal acetylation. All four N-termini displayed NatA-like substrate specificity. LC-MS/MS analysis, data processing and storage were performed as described[9].

Whole proteome sequences and frequency analysis. We selected 38 species (ten holozoa, seven plant, five fungi, five microsporidia, four excavata, five chromalveolata, and two amoeboza) to cover the full range of eukaryotic proteome complexity. Whole datasets of proteins from these organisms with complete proteome annotations available in UniProt were downloaded as FASTA formats in April 2015 (http://www.uniprot.org). We used reviewed (UniProtKB/Swiss-Prot) and unreviewed (UniProtKB/TrEMBL) protein entries for each analyzed species. Yet, if the reviewed proteome had less than thousand proteins than only the more extensive unreviewed proteome was used.

We analyzed the amino acid residue frequency bias (over or under-representation) for each N-terminal position by calculating the residue frequency for each N-terminal position divided by the total proteome frequency of this residue (program script for N-termini analysis is available upon request). MultiExperiment Viewer (MeV), which is part of the TM4 Microarray Software Suite (v4.9), was used to perform hierarchical clustering and to generate heat maps with frequency of amino acid at second, third, fourth fifth, and sixth position[90]. The parameters used for the hierarchical clustering were the sample tree, Pearson correlation and the average linkage method.

References

1. Marintchev, A. & Wagner, G. Translation initiation: structures, mechanisms and evolution. *Q Rev Biophys* **37,** 197–284 (2004).
2. Jones, J. D. & O'Connor, C. D. Protein acetylation in prokaryotes. *Proteomics* **11,** 3012–3022 (2011).
3. Soppa, J. Protein acetylation in archaea, bacteria, and eukaryotes. *Archaea* **2010,** (2010); doi: 10.1155/2010/820681.
4. Starheim, K. K., Gevaert, K. & Arnesen, T. Protein N-terminal acetyltransferases: when the start matters. *Trends Biochem Sci* **37,** 152–161 (2012).
5. Brown, J. L. & Roberts, W. K. Evidence that approximately eighty per cent of the soluble proteins from Ehrlich ascites cells are Nalpha-acetylated. *J Biol Chem* **251,** 1009–1014 (1976).
6. Arnesen, T. *et al.* Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans. *Proc Natl Acad Sci USA* **106,** 8157–8162 (2009).
7. Bienvenut, W. V. *et al.* Comparative large scale characterization of plant versus mammal proteins reveals similar and idiosyncratic N-alpha-acetylation features. *Mol Cell Proteomics* **11,** M111 015131 (2012).
8. Goetze, S. *et al.* Identification and functional characterization of N-terminally acetylated proteins in Drosophila melanogaster. *PLoS Biol* **7,** e1000236 (2009).
9. Van Damme, P. *et al.* NatF contributes to an evolutionary shift in protein N-terminal acetylation and is important for normal chromosome segregation. *PLoS Genet* **7,** e1002169 (2011).
10. Bienvenut, W. V., Giglione, C. & Meinnel, T. Proteome-wide analysis of the amino terminal status of Escherichia coli proteins at the steady-state and upon deformylation inhibition. *Proteomics* **15,** 2503–2518 (2015).
11. Hershko, A., Heller, H., Eytan, E., Kaklij, G. & Rose, I. A. Role of the alpha-amino group of protein in ubiquitin-mediated protein breakdown. *Proc Natl Acad Sci USA* **81,** 7021–7025 (1984).

12. Hoshiyasu, S. *et al.* Potential involvement of N-terminal acetylation in the quantitative regulation of the epsilon subunit of chloroplast ATP synthase under drought stress. *Biosci Biotechnol Biochem* **77,** 998–1007 (2013).

13. Zattas, D., Adle, D. J., Rubenstein, E. M. & Hochstrasser, M. N-terminal acetylation of the yeast Derlin Der1 is essential for Hrd1 ubiquitin-ligase activity toward luminal ER substrates. *Mol Biol Cell* **24,** 890–900 (2013).

14. Hwang, C. S., Shemorry, A. & Varshavsky, A. N-terminal acetylation of cellular proteins creates specific degradation signals. *Science* **327,** 973–977 (2010).

15. Shemorry, A., Hwang, C. S. & Varshavsky, A. Control of protein quality and stoichiometries by N-terminal acetylation and the N-end rule pathway. *Mol Cell* **50,** 540–551 (2013).

16. Forte, G. M., Pool, M. R. & Stirling, C. J. N-terminal acetylation inhibits protein targeting to the endoplasmic reticulum. *PLoS Biol* **9,** e1001073 (2011).

17. Murthi, A. & Hopper, A. K. Genome-wide screen for inner nuclear membrane protein targeting in Saccharomyces cerevisiae: roles for N-acetylation and an integral membrane protein. *Genetics* **170,** 1553–1560 (2005).

18. Scott, D. C., Monda, J. K., Bennett, E. J., Harper, J. W. & Schulman, B. A. N-terminal acetylation acts as an avidity enhancer within an interconnected multiprotein complex. *Science* **334,** 674–678 (2011).

19. Behnia, R., Panic, B., Whyte, J. R. & Munro, S. Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p. *Nat Cell Biol* **6,** 405–413 (2004).

20. Setty, S. R., Strochlic, T. I., Tong, A. H., Boone, C. & Burd, C. G. Golgi targeting of ARF-like GTPase Arl3p requires its Nalpha-acetylation and the integral membrane protein Sys1p. *Nat Cell Biol* **6,** 414–419 (2004).

21. Hofmann, I. & Munro, S. An N-terminally acetylated Arf-like GTPase is localised to lysosomes and affects their motility. *J Cell Sci* **119,** 1494–1503 (2006).

22. Dikiy, I. & Eliezer, D. N-terminal acetylation stabilizes N-terminal helicity in lipid- and micelle-bound alpha-synuclein and increases its affinity for physiological membranes. *J Biol Chem* **289,** 3652–3665 (2014).

23. Polevoda, B., Cardillo, T. S., Doyle, T. C., Bedi, G. S. & Sherman, F. Nat3p and Mdm20p are required for function of yeast NatB Nalpha-terminal acetyltransferase and of actin and tropomyosin. *J Biol Chem* **278,** 30686–30697 (2003).

24. Singer, J. M. & Shaw, J. M. Mdm20 protein functions with Nat3 protein to acetylate Tpm1 protein and regulate tropomyosin-actin interactions in budding yeast. *Proc Natl Acad Sci USA* **100,** 7644–7649 (2003).

25. Van Damme, P. *et al.* N-terminal acetylome analyses and functional insights of the N-terminal acetyltransferase NatB. *Proc Natl Acad Sci USA* **109,** 12449–12454 (2012).

26. Arnaudo, N. *et al.* The N-terminal acetylation of Sir3 stabilizes its binding to the nucleosome core particle. *Nat Struct Mol Biol* **20,** 1119–1121 (2013).

27. van Welsem, T. *et al.* Synthetic lethal screens identify gene silencing processes in yeast and implicate the acetylated amino terminus of Sir3 in recognition of the nucleosome core. *Mol Cell Biol* **28,** 3861–3872 (2008).

28. Aksnes, H. *et al.* An organellar nalpha-acetyltransferase, naa60, acetylates cytosolic N termini of transmembrane proteins and maintains Golgi integrity. *Cell Rep* **10,** 1362–1374 (2015).

29. Kalvik, T. V. & Arnesen, T. Protein N-terminal acetyltransferases in cancer. *Oncogene* **32,** 269–276 (2013).

30. Rope, A. F. *et al.* Using VAAST to identify an X-linked disorder resulting in lethality in male infants due to N-terminal acetyltransferase deficiency. *Am J Hum Genet* **89,** 28–43 (2011).

31. Esmailpour, T. *et al.* A splice donor mutation in NAA10 results in the dysregulation of the retinoic acid signalling pathway and causes Lenz microphthalmia syndrome. *J Med Genet* **51,** 185–196 (2014).

32. Myklebust, L. M. *et al.* Biochemical and cellular analysis of Ogden syndrome reveals downstream Nt-acetylation defects. *Hum Mol Genet* **24,** 1956–1976 (2015).

33. Casey, J. P. *et al.* NAA10 mutation causing a novel intellectual disability syndrome with Long QT due to N-terminal acetyltransferase impairment. *Sci Rep* **5,** 16022 (2015).

34. Liszczak, G. & Marmorstein, R. Implications for the evolution of eukaryotic amino-terminal acetyltransferase (NAT) enzymes from the structure of an archaeal ortholog. *Proc Natl Acad Sci USA* **110,** 14652–14657 (2013).

35. Chang, Y. Y. & Hsu, C. H. Structural basis for substrate-specific acetylation of Nalpha-acetyltransferase Ard1 from Sulfolobus solfataricus. *Sci Rep* **5,** 8673 (2015).

36. Tanaka, S., Matsushita, Y., Yoshikawa, A. & Isono, K. Cloning and molecular characterization of the gene rimL which encodes an enzyme acetylating ribosomal protein L12 of Escherichia coli K12. *Mol Gen Genet* **217,** 289–293 (1989).

37. Yoshikawa, A., Isono, S., Sheback, A. & Isono, K. Cloning and nucleotide sequencing of the genes rimI and rimJ which encode enzymes acetylating ribosomal proteins S18 and S5 of Escherichia coli K12. *Mol Gen Genet* **209,** 481–488 (1987).

38. Ametzazurra, A., Larrea, E., Civeira, M. P., Prieto, J. & Aldabe, R. Implication of human N-alpha-acetyltransferase 5 in cellular proliferation and carcinogenesis. *Oncogene* **27,** 7296–7306 (2008).

39. Evjenth, R. *et al.* Human Naa50p (Nat5/San) displays both protein N{alpha} and N{epsilon} acetyltransferase activity. *J Biol Chem* **284,** 31122–31129 (2009).

40. Mullen, J. R. *et al.* Identification and characterization of genes and mutants for an N-terminal acetyltransferase from yeast. *EMBO J* **8,** 2067–2075 (1989).

41. Park, E. C. & Szostak, J. W. ARD1 and NAT1 proteins form a complex that has N-terminal acetyltransferase activity. *EMBO J* **11,** 2087–2093 (1992).

42. Polevoda, B. & Sherman, F. NatC Nalpha-terminal acetyltransferase of yeast contains three subunits, Mak3p, Mak10p, and Mak31p. *J Biol Chem* **276,** 20154–20159 (2001).

43. Song, O. K., Wang, X., Waterborg, J. H. & Sternglanz, R. An Nalpha-acetyltransferase responsible for acetylation of the N-terminal residues of histones H4 and H2A. *J Biol Chem* **278,** 38109–38112 (2003).

44. Starheim, K. K. *et al.* Identification of the human N(alpha)-acetyltransferase complex B (hNatB): a complex important for cell-cycle progression. *Biochem J* **415,** 325–331 (2008).

45. Arnesen, T. *et al.* Identification and characterization of the human ARD1-NATH protein acetyltransferase complex. *Biochem J* **386,** 433–443 (2005).

46. Starheim, K. K. *et al.* Knockdown of human N alpha-terminal acetyltransferase complex C leads to p53-dependent apoptosis and aberrant human Arl8b localization. *Mol Cell Biol* **29,** 3569–3581 (2009).

47. Hole, K. *et al.* The human N-alpha-acetyltransferase 40 (hNaa40p/hNatD) is conserved from yeast and N-terminally acetylates histones H2A and H4. *PLoS One* **6,** e24713 (2011).

48. Ben-Bassat, A. *et al.* Processing of the initiation methionine from proteins: properties of the Escherichia coli methionine aminopeptidase and its gene structure. *J Bacteriol* **169,** 751–757 (1987).

49. Moerschell, R. P., Hosokawa, Y., Tsunasawa, S. & Sherman, F. The specificities of yeast methionine aminopeptidase and acetylation of amino-terminal methionine *in vivo.* Processing of altered iso-1-cytochromes c created by oligonucleotide transformation. *J Biol Chem* **265,** 19638–19643 (1990).

50. Polevoda, B., Norbeck, J., Takakura, H., Blomberg, A. & Sherman, F. Identification and specificities of N-terminal acetyltransferases from Saccharomyces cerevisiae. *EMBO J* **18,** 6155–6168 (1999).

51. Van Damme, P. *et al.* Proteome-derived peptide libraries allow detailed analysis of the substrate specificities of N(alpha)-acetyltransferases and point to hNaa10p as the post-translational actin N(alpha)-acetyltransferase. *Mol Cell Proteomics* **10,** M110 004580 (2011).

52. Helbig, A. O. *et al.* Profiling of N-acetylated protein termini provides in-depth insights into the N-terminal nature of the proteome. *Mol Cell Proteomics* **9,** 928–939 (2010).
53. Zhu, H. Y. *et al.* *In silico* identification and characterization of N-Terminal acetyltransferase genes of poplar (Populus trichocarpa). *Int J Mol Sci* **15,** 1852–1864 (2014).
54. Arnesen, T. *et al.* Characterization of hARD2, a processed hARD1 gene duplicate, encoding a human protein N-alpha-acetyltransferase. *BMC Biochem* **7,** 13 (2006).
55. Arnesen, T. *et al.* A novel human NatA Nalpha-terminal acetyltransferase complex: hNaa16p-hNaa10p (hNat2-hArd1). *BMC Biochem* **10,** 15 (2009).
56. Dinh, T. V. *et al.* Molecular identification and functional characterization of the first Nalpha-acetyltransferase in plastids by global acetylome profiling. *Proteomics* **15,** 2426–2435 (2015).
57. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215,** 403–410 (1990).
58. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39,** W29–37 (2011).
59. James, T. Y. *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443,** 818–822 (2006).
60. Yoon, H. S. *et al.* Broadly sampled multigene trees of eukaryotes. *BMC Evol Biol* **8,** 14 (2008).
61. Lang, D. *et al.* Genome-wide phylogenetic comparative analysis of plant transcriptional regulation: a timeline of loss, gain, expansion, and correlation with complexity. *Genome Biol Evol* **2,** 488–503 (2010).
62. Capella-Gutierrez, S., Marcet-Houben, M. & Gabaldon, T. Phylogenomics supports microsporidia as the earliest diverging clade of sequenced fungi. *BMC Biol* **10,** 47 (2012).
63. Huerta-Cepas, J., Marcet-Houben, M. & Gabaldón, T. A nested phylogenetic reconstruction approach provides scalable resolution in the eukaryotic Tree Of Life. *PeerJ PrePrints* 2:e223v1 (2014).
64. Liszczak, G., Arnesen, T. & Marmorstein, R. Structure of a Ternary Naa50p (NAT5/SAN) N-terminal Acetyltransferase Complex Reveals the Molecular Basis for Substrate-specific Acetylation. *J Biol Chem* **286,** 37002–37010 (2011).
65. Liszczak, G. *et al.* Molecular basis for N-terminal acetylation by the heterodimeric NatA complex. *Nat Struct Mol Biol* **20,** 1098–1105 (2013).
66. Zybailov, B. *et al.* Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* **3,** e1994 (2008).
67. Angus-Hill, M. L., Dutnall, R. N., Tafrov, S. T., Sternglanz, R. & Ramakrishnan, V. Crystal structure of the histone acetyltransferase Hpa2: A tetrameric member of the Gcn5-related N-acetyltransferase superfamily. *J Mol Biol* **294,** 1311–1325 (1999).
68. Katinka, M. D. *et al.* Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. *Nature* **414,** 450–453 (2001).
69. Williams, B. C. *et al.* Two putative acetyltransferases, san and deco, are required for establishing sister chromatid cohesion in Drosophila. *Curr Biol* **13,** 2025–2036 (2003).
70. Pimenta-Marques, A. *et al.* Differential requirements of a mitotic acetyltransferase in somatic and germ line cells. *Dev Biol* **323,** 197–206 (2008).
71. Gautschi, M. *et al.* The yeast N(alpha)-acetyltransferase NatA is quantitatively anchored to the ribosome and interacts with nascent polypeptides. *Mol Cell Biol* **23,** 7403–7414 (2003).
72. Arnesen, T. *et al.* Cloning and characterization of hNAT5/hSAN: an evolutionarily conserved component of the NatA protein N-alpha-acetyltransferase complex. *Gene* **371,** 291–295 (2006).
73. Arnesen, T. *et al.* The chaperone-like protein HYPK acts together with NatA in cotranslational N-terminal acetylation and prevention of Huntingtin aggregation. *Mol Cell Biol* **30,** 1898–1909 (2010).
74. Van Damme, P., Hole, K., Gevaert, K. & Arnesen, T. N-terminal acetylome analysis reveals the specificity of Naa50 (Nat5) and suggests a kinetic competition between N-terminal acetyltransferases and methionine aminopeptidases. *Proteomics* **15,** 2436–2446 (2015).
75. Van Damme, P. *et al.* A review of COFRADIC techniques targeting protein N-terminal acetylation. *BMC Proc* **3** Suppl 6, S6 (2009).
76. Polevoda, B. & Sherman, F. N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins. *J Mol Biol* **325,** 595–622 (2003).
77. Sanchez, J. Alanine is the main second amino acid in vertebrate proteins and its coding entails increased use of the rare codon GCG. *Biochem Biophys Res Commun* **373,** 589–592 (2008).
78. Shemesh, R., Novik, A. & Cohen, Y. Follow the leader: preference for specific amino acids directly following the initial methionine in proteins of different organisms. *Genomics Proteomics Bioinformatics* **8,** 180–189 (2010).
79. Koonin, E. V. The origin and early evolution of eukaryotes in the light of phylogenomics. *Genome Biol* **11,** 209 (2010).
80. Pang, A. L. *et al.* Cloning, characterization, and expression analysis of the novel acetyltransferase retrogene Ard1b in the mouse. *Biol Reprod* **81,** 302–309 (2009).
81. Van Damme, P., Stove, S. I., Glomnes, N., Gevaert, K. & Arnesen, T. A Saccharomyces cerevisiae model reveals *in vivo* functional impairment of the Ogden syndrome N-terminal acetyltransferase Naa10S37P mutant. *Mol Cell Proteomics* **13,** 2031–2041 (2014).
82. Silva, R. D. & Martinho, R. G. Developmental roles of Protein N-terminal acetylation. *Proteomics* **15,** 2402–2409 (2015).
83. Polevoda, B., Arnesen, T. & Sherman, F. A synopsis of eukaryotic Nalpha-terminal acetyltransferases: nomenclature, subunits and substrates. *BMC Proc* **3** Suppl 6, S2 (2009).
84. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23,** 2947–2948 (2007).
85. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56,** 564–577 (2007).
86. Guilgur, L. G. *et al.* Requirement for highly efficient pre-mRNA splicing during Drosophila early embryonic development. *Elife* **3,** e02181 (2014).
87. Chou, T. B. & Perrimon, N. Use of a yeast site-specific recombinase to produce female germline chimeras in Drosophila. *Genetics* **131,** 643–653 (1992).
88. Wang, Y. *et al.* Drosophila variable nurse cells encodes arrest defective 1 (ARD1), the catalytic subunit of the major N-terminal acetyltransferase complex. *Dev Dyn* **239,** 2813–2827 (2010).
89. Staes, A. *et al.* Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat Protoc* **6,** 1130–1141 (2011).
90. Saeed, A. I. *et al.* TM4 microarray software suite. *Methods Enzymol* **411,** 134–193 (2006).
91. Zhang, G. *et al.* Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346,** 1311–1320 (2014).

Acknowledgements

Author Contributions

O.S.R. bioinformatics analysis and interpretation of data, drafting and revising the article; A.F. and P.P. biochemical analysis of Naa10 and Naa50-containing complexes; P.V.D. COFRADIC analysis, drafting and revising the article; C.J.C. bioinformatics analysis and interpretation of data, drafting and revising the article; R.G.M. conception and design, interpretation of data, drafting and revising the article.

Additional Information