FUNCTIONAL GENOMICS EVIDENCE UNEARTHS NEW MOONLIGHTING ROLES OF OUTER RING COAT NUCLEOPORINS

SREP-13-05332A

Katerina R. Katsani[1], Manuel Irimia[2], Christos Karapiperis[3], Zacharias G. Scouras[3], Benjamin J. Blencowe[2], Vasilis J. Promponas[4], Christos A. Ouzounis[2,4,5*]

[1] Department of Molecular Biology & Genetics, Democritus University of Thrace, GR-68100 Alexandroupolis, Greece

[2] Donnelly Centre for Cellular & Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario M5S 3E1, Canada

[3] Department of Genetics, Development & Molecular Biology, School of Biology, Faculty of Sciences, Aristotle University of Thessaloniki, GR-54124 Thessalonica, Greece

[4] Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, PO Box 20537, CY-1678 Nicosia, Cyprus

[5] Institute of Applied Biosciences, Centre for Research & Technology, PO Box 361, GR-57001 Thessalonica, Greece

[*] current address: Chemical Process Engineering Research Institute, Centre for Research & Technology, PO Box 361, GR-57001 Thessalonica, Greece

[*]corresponding author: CAO • email: ouzounis@certh.gr

SUPPLEMENTARY INFORMATION

SUPPLEMENTARY TEXT

NUP160 [FBpp0079788]

Nup160 – encoded in *D. melanogaster* by the CG4738 gene – is the longest protein in this group with 1411 residues, containing a number of WD-40 repeats, and other elements. Nup160 together with another group member, Nup96, are lethal hybrid incompatibility genes between *D. melanogaster* and *D. simulans*[1,2].

Sequence profile searches with CG4738 converge at iteration 6, with 316 homologs including 3 homologs of known structure (**Supplementary Table 1**). Certain alleged homologs (e.g. hypothetical protein AaeL_AAEL001399, GI:157118605) are not identified by our profile searches and therefore deemed as false positives (containing DUF3639). [Case 01:] Region 5-1183 matches the 1532-residue protein from the fungus *Metarhizium anisopliae* ARSEF 23 (GI:322708659)[3] in positions 370-1532 signifying that the Nup160 domain exists in this protein whose N-terminal domain (positions 25-320) corresponds to the RAD51 signature (DMC1/archaeal radA). Its closest homolog, the 1165-residue protein MAC_04219 (GI:322698013) from *Metarhizium acridum* CQMa 102 contains the core Nup160 domain in positions 28-575; the remainder C-terminal region 576-1165 also matches Nup160 without the RAD51 domain (which is present in another 348-residue protein of *M. acridum*, GI:322698012) – traces of this domain arrangement can be found in the genome of *Chaetomium thermophilum*[4] – negative strand 549120-549323 (frame –2) and 549805-549867 (frame –1) as neighboring genes, possibly a pyro-sequencing artifact. [Case 02:] Nup160 region 168-1187 matches a 1365-residue protein from *Ogataea parapolymorpha* DL-1 (GI:320583072) at positions 102-1016; the C-terminal domain of this protein contains a protoporphyrinogen oxidase signature – plus a fragment sequence in *Dekkera bruxellensis* AWRI1499 (GI:385306005), both annotated automatically. [Case 03:] A 1791-residue protein from *Dictyostelium discoideum* AX4 (GI:166240053) with Nup160 (27-1553) contains the CcmE/CycJ domain (positions 1555-1659). Importantly, in this case, the N-terminal Nup160 domain is detectable in related organisms with exactly the same configuration at its C-terminal end (Nup160/CcmE): an unannotated 1625-residue protein from *D. purpureum* (GI:330796511), an annotated 1535-residue protein from *Polysphondylium pallidum* PN500 (GI:281210825) and yet another partially annotated 1727-residue protein from *D. fasciculatum* (GI:328873820) (**Supplementary Fig. 5**).

At convergence, the search returns more homologs, including remotely related sequences not previously characterized as Nup160s. To the extent of our knowledge, the significantly similar sequences contain one false positive, namely protein CBY40925.1 (GI:313214600) from *Oikopleura dioica*, a rapidly evolving tunicate[5]. [Case 04:] The 590-residue protein OsI_05816 (GI:218190037) from *Oryza sativa* contains domain DUF1668 at positions 4-332, while its C-terminal region (residues 397-589) clearly identifies Nup160 domains in reverse database searches. This co-occurrence has not been observed previously. [Case 05:] A 2247-residue protein from *Tribolium castaneum* (GI:189241332) contains the Nup160 domain at positions 862-2237 (C-terminus) while its N-terminal region contains an aminopeptidase domain (positions 1-652) (Pfam_01433), equivalent to another entry in the database (GI:270013150), probably the correct version – an error possibly attributable to misguided gene prediction in the first case (**Supplementary Table 2**).

NUP133 [FBpp0083695]

Nup133 – encoded in *D. melanogaster* by the CG6958 gene – is a 1200 residue subunit, built from an N-terminal β-propeller motif, followed by a C-terminal α-helical domain, structurally related to the ScNup157/Nup170 (vertebrate Nup155) of the Nic96 complex[6]. Moreover, Nup133 is anchored to the NPC via Nup107 in tail-to-tail fashion[7].

The Nup133 domain identifies homologs across a wide range of taxa with a certain degree of variation (N-terminal, Pfam_08801). Following iteration 3 by excluding a single false positive, an unknown low-complexity protein from *Ectocarpus siliculosus* (GI:299472445)[8], the search yields 320 Nup133-containing sequences, including a falsely annotated DNA topoisomerase 2-alpha from *Trichinella spiralis* (GI:339241837)[9] not detected in Pfam, while recognizing Nup133 homologs in reverse database searches (**Supplementary Table 1**). Exceptionally, in three cases, this family is found in a multi-domain arrangement with other domains, namely (i) [Case 06:] in the 1970-residue *Rhodototula glutinis* ATCC 204091 (GI:342321222) in region 1124-1866 (Nup133-like), a Nup170-like region at positions 600-1118 and an arginase/agmatinase domain (residues 187-540), (ii) [Case 07:] in the 1613-residue hypothetical protein from *Pyrenophora tritici-repentis* Pt-1C-BFP (GI:189196830) (residues 360-1498) and a STF2-like N-terminus with a fungal hypothetical protein, and (iii) [Case 08:] in the 2555-residue *Dictyostelium fasiculatum* DFA_09912 protein (GI:328866705) at residues 1585-2555C is Nup133, while residues 990–1500 correspond to an amylase catalytic domain (**Supplementary Table 2**).

NUP107 [FBpp0079710]

Nup107 – encoded in *D. melanogaster* by the essential CG6743 gene – is the prototype member of the group[10] required for the assembly of a subset of Nups in mature NPCs[11,12]. Interestingly, Nup107 is folded in an α-helical element present similarly to other subunits of the complex (Nup85, ScNup145C/Nup96), as well as in ScNic96/Nup93 and, curiously, in Sec31 and Sec16 of the COPII vesicle coating system[6,13]. The 845-residue sequence from *D. melanogaster* identifies homologs that have been previously characterized as well as newly identified entries (Nup84/107; Pfam_04121).

The search converges at iteration 7, recovering 340 homologs from a wide phylogenetic range, while putative hits of low sequence similarity such as Nic96/Nup93 family members are not admitted (**Supplementary Table 1**). Five proteins of known structure match Nup107: 3IKO(CFI) match positions 64-535, 3JRO(C) matches positions 64-499 while 3I4R/3CQC3/3CQG(A) match positions 569-844 thus covering its entire length (**Supplementary Table 1**).

Four longer proteins present a complex domain structure: (i) [Case 09:] a 1234-residue protein from horse (GI:338726199) (contains domain 341/425-end; N-terminal: tetratricopeptide repeat domain - TPR); (ii) [Case 10:] a 1042-residue protein from *Acromyrmex echinatior* (GI:332016695) contains the Nup107 domain (residues 215-1041); this protein also contains a 200-residue long N-terminal PBP/GOBP (pheromone binding/general-odorant binding) domain, found also in the same species within the Signal recognition particle 72 kDa protein (GI:332023052) at positions 6-108; this Srp72 protein contains an APC3 (anaphase-promoting complex subunit 3; Pfam_12895) at positions 154-204 (unpublished) – this domain pattern is present in *Solenopsis invicta* (GI:322788827); (iii) [Case 11:] a 2763-residue protein from *Rhodotorula glutinis* ATCC 204091 (GI:342319308), which beyond residue 800 (Nup107, residues 1-802) also contains two ARF-binding sites, a Sec7 domain (positions 1900-2000) and a pleckstrin domain (positions 2200-2346); (iv) [Case 12:] a 1490-residue protein from *Clonorchis sinensis* (GI:358337287) matches Nup107 (residues 1-926) and then at positions 935-1490C DUF1767 (positions 937-1017) and a UBA (Ubiquitin Associated) domain (positions 1224-1260) (**Supplementary Table 2**).

NUP98 [FBpp0083851]

Nup98 (Nup98-96) – encoded in *D. melanogaster* by the CG10198 gene – is involved in lethal hybrid incompatibility, like Nup160, between *D. melanogaster* and *D. simulans*[2]. The largest Nup98 transcript (Nup98-RA) encodes a single, large precursor polyprotein of 1960 residues that, in other eukaryotes, and thus presumably in flies, autoproteolytically cleaves itself yielding separate Nup98 and Nup96 proteins (Nup145N and Nup145C respectively in yeast), with only Nup96/Nup145-C being a stable member of the Y-complex[14].

The full-length 1960-residue sequence identifies succinctly its shortest isoform of the auto-proteolytic product as well as multiple variants of longer length, primarily from lower organisms including fungi (autoproteolytic autopeptidase Pfam_04096). Despite its wide distribution across eukaryotes, Nup98 (Nup98-96) is a conserved, well-defined domain represented by either the long form (around 2000 residues) or the short form (around 1000 residues). The N-terminus contains multiple FG repeats and regions common with other nucleoporins, evidently including orthologs SONB/Nup189 (**Data Supplement DS09**). The strongly conserved sequence features of Nup98 characterize this well-defined protein family that can be readily detected: with a limit set at 500 hits, the lowest sequence similarity level observed is 24% identity.

[Case 13:] An exception to this pattern is the presence of a 2177-residue protein in *Fusarium oxysporum* Fo5176, which further to the N-terminal 2000-residue Nup98 domain also contains a SET domain (positions 2037-2147) (unpublished, GI:342873147), also found in *Metarhizium acridum* CQMa 102 (GI:322698664) and *M. anisopliae* ARSEF 23 (GI:322711125) [3]; it is interesting to note that a similar pattern of gene fusion between the N-terminal domain of Nup98 and other SET-containing proteins have been observed in human leukemia[15]. [Case 14:] Another domain association is found in a 2823-residue protein from *Monosiga brevicollis* MX1 (GI:167537610), which, following the Nup98 region in positions 800-2823C, also contains a beta galactosidase-associated domain (DUF1680) (positions ~100-700) and a galactose binding lectin domain (Pfam_02140) (positions 695-788) (**Supplementary Table 2**). The region 1314-1769 matches proteins of known structure (e.g. 3IKO), which include the complex with SEC13[16]. In total, there are >500 sequences detected at iteration 10, without false positives and a very deep phylogenetic distribution across the eukaryotes (**Supplementary Table 1**).

NUP75 [FBpp0085954]

Nup75 – encoded in *D. melanogaster* by the CG5733 gene – corresponds to a 668 residue long protein. Nup75 (Nup85 in yeast) shares the same α-helical fold with another two members of the group (Nup107/ScNup84, and Nup96/ScNup145C) also found in Nic96 and interestingly the COPII coatomer element Sec31, suggesting an evolutionary link between NPCs and COP/clathrin-like membrane coats[13,17,18]. Structural homologs are co-crystallized with *S. cerevisiae* Seh-1[13,18].

The length variation of the archetypal 668-residue protein is due to a 300-residue long N-terminal extension in a significant majority of fungal proteins. Iteration 4 yields 314 homologs (Pfam_07575) (**Supplementary Table 1**). The family exhibits wide distribution across eukaryotes as previously demonstrated, including amoebozoa and Bacillariophyta (stramenopiles) e.g. *Thallassiosira pseudonana* CCMP1335 (GI:223997390)[19]. Curiously, one puzzling hit is from the Nup98 family in *Harpegnathos saltator* (GI:307191801)[20] (**Data Supplement DS09**). However, there are five cases of length variation not detected previously, which correspond to possible domain involvement of Nup75 into other associations (**Supplementary Table 2**).

[Case 15:] First, the Nup75 domain is found in the 1192-residue Nup153 from *Naegleria gruberi* (GI:290983204)[21] with the N-terminus (residues 1-580) having FG-repeats and the subsequent region (residues 601-1192) the Nup75 domain. The existence of a gene annotated as Nup98 in *N. gruberi* (GI: 290978415) indicates that this is a genuine instance of Nup75 in this species. [Case 16:] Second, the 1419-residue enzyme acetyl-CoA carboxylase biotin carboxylase (carbamoyl-phosphate synthase subunit L, involved in pyrimidine and arginine biosynthesis) from *Rhodotorula glutinis* ATCC 204091 (GI:342319109) contains an N-terminal Nup75 domain (residues 27-628) followed by the enzyme region (residues 690-1409). [Case 17:] Third, the N-terminal domain (residues 353-896) of a 1899-residue aconitase from *Metarhizium acridum* CQMa 102 (GI:322702102)[3] identifies the Nup75 domain from *M. anisopliae* and other pathogenic fungi (residues 244-898), while the C-terminal domain (900-1899) detects its aconitase 'orthologs'. [Case 18:] Fourth, the 2055-residue *Chlorella variabilis* protein 51187 (GI:307108886) N-terminus (residues 175-847) is also homologous to Nup75[22]; the C-terminus of this protein, more than 1000-residues long, contains two instances of the SMC domain

(structural maintenance of chromosomes/condensins). [Case 19:] Fifth, the 1828-residue long Vitamin h transporter 1 from *Grosmannia clavigera* N-terminus (residues 400-1027, up to 1260) (GI:320590333)[23] also exhibits the presence of Nup75; the remainder of this protein contains a MFS (major facilitator superfamily) domain (residues ~1581-1678) involved in secondary transport. Note that none of these domain associations of Nup75 do not appear to be supported by nucleotide sequence searches (see **Methods** and **Supplementary Table 2**).

NUP43 [FBpp0082998]

Nup43 and Nup37 were initially identified only in vertebrates and are absent from the *S. cerevisiae* complex[24,25]. The 358-residue Nup43 – encoded in *D. melanogaster* by the CG7671 gene – contains WD40 repeats in positions 88-358. Recently, the three-dimensional structure of human Nup43 was determined (PDB identifier: 4I79, unpublished). The N-terminal domain specifically recognizes 102 homologs by iteration 4 (**Supplementary Table 1**) including *Drosophila* and other insects, lower and then higher vertebrates, and echinozoa (GI:115623733, hypothetical protein is the sea urchin Nup43 ortholog).

It is interesting to note that Nup43 is detected in fungi for the first time[26], in the species *Batrachochytrium dendrobatidis* JAM81 BATDEDRAFT_37327 (GI:328767770), 379 residues long. No other 'orthologs' in fungi can be detected. The significance of this taxonomic anomaly is not understood at present; the closest relatives of this protein are *Drosophila* Nup43s, possibly due to sampling/over-representation at this taxonomic range. Despite claims to the contrary[27], we were unable to establish homology between the masked Nup43 query sequence against a number of plant proteins previously characterized as Nup43 orthologs e.g. in *Medicago truncatula* (GI:355511317); instead, all these proteins contain WD40 repeats and in some cases are annotated as such. Curiously, the hamster sequence is only 161-residue long (GI:344241050); the reported missing region is also absent from certain mammalian isoforms. The human sequence (GI:55663479) contains a 61-residue N-terminal domain of unknown nature, also found in other related species. A sequence from sea urchin *Strongylocentrotus purpuratus* (GI:115623733) included a 79-residue direct repeat (118-206) 100% identical with an adjacent sequence (227-305) – later updated to a corrected version (GI:390334509).

[Case 20:] The search also returns an *Ascaris suum* (nematode) MIF 448-residue protein (GI:324503956)[28], appearing to contain a macrophage migration inhibitory factor (MIF) domain[29] at its N-terminus (positions 2-94) and the Nup43 region at the C-terminus (positions 115-174). As this finding is not supported in related species or across databases – in fact, in the same species there is a 115-residue protein (GI:49257069) without the Nup43 domain – while the 2976-bp sequence was generated by a shotgun/assembly project, it is doubtful that this association of MIF with Nup43 reflects reality and is likely to be a short-read/assembly artifact, not supported by nucleotide searches (**Supplementary Table 2**).

[Case 21:] Surprisingly, Nup43 is found at the N-terminal region of 1000-residue long proteins of six Apocrita insects with DHX15 helicase at the remaining sequence (**Supplementary Table 2**): *Nasonia vitripennis* (GI:345482402), *Camponotus floridanus* (GI:307170456), *Apis mellifera* (GI:328780322), *Solenopsis invicta* (GI:322796692), *Bombus impatiens* (GI:350397130), *Bombus terrestris* (GI:340725762). This region has not been recognized as unique to the Nup43 family and was generally annotated as a WD40 repeat region. We unequivocally demonstrate that the DHX domain is present in these insect Nup43s, while missing from the *Drosophila* Nup43s. These organisms do not contain another copy of Nup43; thus the N-terminus of these genes can be considered as the Nup43 ortholog in these species. Gene CG11107 in *D. melanogaster* and its orthologs in *Drosophila* and other insect species is reciprocally the DHX15 ortholog of the above six proteins, thus implying that it might be functionally associated with Nup43[30]. The recent immunoisolation of a Werner Helicase interacting protein (WHIP) with the human Y-complex also points towards a potential functional link between the Y-complex and helicases[31].

NUP37 [FBpp0084187]

The smallest member of the group, the 320-residue Nup37 protein – encoded in *D. melanogaster* by the CG11875 gene – contains WD40 repeats in positions 72-320C. Recently, the three-dimensional structure of *Schizosaccharomyces pombe* Nup37 was determined (4FHL, 4FHM, 4FHN, 4GQ1, 4GQ2)[32,33]. Querying the database with positions 1-71 (and a slightly modified e-value threshold of 10^{-05}) returns insect homologs (iteration 3), a homolog from the plant *Medicago truncatula* (iteration 4) two homologs from the hemichordate *Saccoglossus kowalevskii* and *Branchiostoma floridae* (amphioxus)[34] (iteration 5) and subsequently vertebrate homologs (iterations 6-7), with 70 homologs in total (**Supplementary Table 1**). All homologs have ~300-residue sequences, without complex domain patterns. This holds for further iterations (up to 15, not shown), which reveal the presence of Nup37 in all metazoa including cnidaria (sea anemone) as well as plants[26].

SEH1 [FBpp0087938]

SEH1 – encoded in *D. melanogaster* by the CG8722 gene – also known as nup44A from its chromosomal location, is a central but least stable component of the Y-complex[25], also being part of the SEA complex[35]. However, it confers compositional integrity to the complex as depletion by RNA interference of this single subunit depletes the rest of the subunits from kinetochores[36]. The 356-residue SEH1 contains WD40 repeats throughout its length resulting in an open six-bladed β-propeller, with a seventh blade provided by its partner Nup85/75[13].

Following CAST masking (score≥10; instead of 15), during iteration 1 the search already returns SEH1 homologs as well as their closest relatives, those in SEC13 family. Despite the WD40 repeat structure, the search identifies the *bona fide* homologs with specificity (**Supplementary Table 1**). [Case 22:] The only notable exception is a 794-residue

phosphoglucomutase from *Brugia malayi*[37], which contains SEH1/SEC13 repeats at its N-terminus (residues 1-347) and the enzyme domain at the C-terminus (398-794C). It should be noted that *B. malayi* has another gene coding for SEC13 (GI:170588105), thus excluding the possibility that the PGM gene is a sequencing artifact. In total, there are (at least) 500 sequences detected at iteration 6 (including 7 proteins of known structure), without false positives (**Data Supplement DS09**).

[Case 23:] A 878-residue protein from *Schistostoma mansoni* (GI:256087901) contains the SEH1-specific region (positions 500-878C), while exhibiting similarity to MFS at the N-terminus (positions 45-499). [Case 24:] Similarly, there is a 936-residue protein in *Anopheles darlingi* (GI:312376549), which contains a SEH1 region at positions 26-270 and a catalytic domain of protein kinases (residues 500-701). [Case 25:] Another 634-residue protein from the stramenopile *Blastocystis hominis* (GI:300121922) contains a SEH-1 specific region at positions 336-624 and a HCO3 cotransporter domain (Pfam_00955) at positions 68-319 – also available as a single-domain gene in the database (GI:300121560). These hits are not supported by linker searches (see **Methods** and **Supplementary Table 2**, **Data Supplement DS06**).

Yet, two puzzling cases of domain coexistence emerge – e.g. [Case 26:] *Ogataea parapolymorpha* DL-1 (length 1039) (GI:320581285) TAF9-containing protein at positions 11-137 (Pfam_02291) and Chs5p-Arf1p-binding domain (Pfam_09295) at positions 187-571 with SEH1-specific region at the C-terminus (residues 749-1037), and [Case 27:] centrosomal protein 192 in rat Da1-6 (length 2377) (GI:33086682) (SEH1 positions 385-836) and rCG46902 (length 3259) (GI:149064540) (SEH1 positions 314-765), *Taeniopygia guttata* protein (length 3207) (SEH1 positions 1-324) (GI:224046042), *Callithrix jacchus* (SEH1 positions 1-324) (GI:296222198) (length 2839), *Ailuropoda melanoleuca* (GI:301775003) (length 2971) and *Monodelphis domestica* (GI:334325923) (length 2865).

SEC13 [FBpp0083801]

SEC13 – encoded in *D. melanogaster* by the CG6773 gene – is a key component of the COPII membrane coat[38] of the SEA complex[35], which are all members of a family of membrane and vesicle coating assemblies – in addition of being a stable component of the Y-complex[39]. SEC13 has a central structural position and forms a trimeric complex with Nup145C, and Nup84[16], while it shuttles between cytoplasm and nucleus[40]. Similarly to SEH1, SEC13 is unique in that it forms an open six-bladed β-propeller[13], with a seventh blade provided by its partner Nup96/Nup145C[41]. Sequence comparison results are included in the above case of SEH1 (**Supplementary Table 1**, **Data Supplement DS12**, **Supplementary Table 2**). As SEH1 and SEC13 identify each other, all observations for SEH1 are also valid for SEC13 (see text for details).

Data availability

All results (in 12 **Data Supplements**) are available as a ZIP archive (58.3 MBytes).

Data Supplement 1 (DS01): CAST-filtered Y-Nup sequences from *Drosophila melanogaster* used in this study (referred as the query set), in FASTA format (see also **Supplementary Table 1**).

Data Supplement 2 (DS02): CAST-filtered regions for the query set providing detailed statistics per region masked, as CAST output text format.

Data Supplement 3 (DS03): NCBI PSI-BLAST results at various iteration levels for each of the nine sequences of the query set, and resulting multiple sequence alignments with specified redundancy levels, in ZIP archive format. Alignments define implicit profile hidden Markov models (HMMs)[42], that can be built with appropriate tools.

Data Supplement 4 (DS04): Sequence identity distributions in this analysis, see also **Supplementary Figure 2** of the manuscript, in Microsoft Excel® format (XLS).

Data Supplement 5 (DS05): Linker sequences of the multi-domain proteins detected in **DS03** used for the validation of genuine domain associations, in FASTA format.

Data Supplement 6 (DS06): NCBI TBLASTN results against nucleotide databases NR/HTGS/EST (and BLASTP against nrdb) with the linker sequences of multi-domain proteins in **DS05** used as queries, in ZIP archive format.

Data Supplement 7 (DS07): Non-nucleoporin domains matching multi-domain members of the query set, manually collected and labeled accordingly, to assist validated automatic clustering (corresponding to the entries of **Supplementary Table 2**), in FASTA format.

Data Supplement 8 (DS08): Taxonomy distribution of the 27 'external' domains detected in this study (see **DS06**), from NCBI BLAST output, in text format.

Data Supplement 9 (DS09): Clustering results using MCL for the homologies detected by the query set (**Supplementary Table 1** and **Figure 1**), in ZIP archive format.

Data Supplement 10 (DS10): Expression data files for human and mouse across multiple tissues, in Microsoft Excel® Open XML format (XLSX).

Data Supplement 11 (DS11): Protein interaction information from public databases including nucleoporins, known interactions and the domain associations reported herein, in BioLayout format; annotation embedded in the text file.

Data Supplement 12 (DS12): All sequences with significant homology to the query set discovered in this study, including multi-domain proteins, in FASTA format as well as CAST-filtered versions, in ZIP archive format.

REFERENCES

1 Tang, S. & Presgraves, D. C. Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. *Science* **323**, 779-782 (2009).
2 Presgraves, D. C. The molecular evolutionary basis of species formation. *Nat Rev Genet* **11**, 175-180 (2010).
3 Gao, Q. *et al.* Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. *PLoS Genet* **7**, e1001264 (2011).
4 Amlacher, S. *et al.* Insight into structure and assembly of the nuclear pore complex by utilizing the genome of a eukaryotic thermophile. *Cell* **146**, 277-289 (2011).
5 Denoeud, F. *et al.* Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381-1385 (2010).
6 Whittle, J. R. & Schwartz, T. U. Architectural nucleoporins Nup157/170 and Nup133 are structurally related and descend from a second ancestral element. *J Biol Chem* **284**, 28442-28452 (2009).
7 Boehmer, T., Jeudy, S., Berke, I. C. & Schwartz, T. U. Structural and functional studies of Nup107/Nup133 interaction and its implications for the architecture of the nuclear pore complex. *Mol Cell* **30**, 721-731 (2008).
8 Cock, J. M. *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617-621 (2010).
9 Mitreva, M. *et al.* The draft genome of the parasitic nematode *Trichinella spiralis*. *Nat Genet* **43**, 228-235 (2011).
10 Katsani, K. R., Karess, R. E., Dostatni, N. & Doye, V. In vivo dynamics of *Drosophila* nuclear envelope components. *Mol Biol Cell* **19**, 3652-3666 (2008).
11 Boehmer, T., Enninga, J., Dales, S., Blobel, G. & Zhong, H. Depletion of a single nucleoporin, Nup107, prevents the assembly of a subset of nucleoporins into the nuclear pore complex. *Proc Natl Acad Sci U S A* **100**, 981-985 (2003).
12 Walther, T. C. *et al.* The conserved Nup107-160 complex is critical for nuclear pore complex assembly. *Cell* **113**, 195-206 (2003).
13 Brohawn, S. G., Leksa, N. C., Spear, E. D., Rajashankar, K. R. & Schwartz, T. U. Structural evidence for common ancestry of the nuclear pore complex and vesicle coats. *Science* **322**, 1369-1373 (2008).
14 Dokudovskaya, S., Veenhoff, L. M. & Rout, M. P. Cleave to leave: structural insights into the dynamic organization of the nuclear pore complex. *Mol Cell* **10**, 221-223 (2002).
15 Wang, G. G., Cai, L., Pasillas, M. P. & Kamps, M. P. NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis. *Nat Cell Biol* **9**, 804-812 (2007).
16 Nagy, V. *et al.* Structure of a trimeric nucleoporin complex reveals alternate oligomerization states. *Proc Natl Acad Sci U S A* **106**, 17693-17698 (2009).
17 Jeudy, S. & Schwartz, T. U. Crystal structure of nucleoporin Nic96 reveals a novel, intricate helical domain architecture. *J Biol Chem* **282**, 34904-34912 (2007).
18 Debler, E. W. *et al.* A fence-like coat for the nuclear pore membrane. *Mol Cell* **32**, 815-826 (2008).
19 Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86 (2004).
20 Bonasio, R. *et al.* Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*. *Science* **329**, 1068-1071 (2010).
21 Fritz-Laylin, L. K. *et al.* The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140**, 631-642 (2010).
22 Blanc, G. *et al.* The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* **22**, 2943-2955 (2010).
23 DiGuistini, S. *et al.* Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. *Proc Natl Acad Sci U S A* **108**, 2504-2509 (2011).
24 Cronshaw, J. M., Krutchinsky, A. N., Zhang, W., Chait, B. T. & Matunis, M. J. Proteomic analysis of the mammalian nuclear pore complex. *J Cell Biol* **158**, 915-927 (2002).
25 Loiodice, I. *et al.* The entire Nup107-160 complex, including three new members, is targeted as one entity to kinetochores in mitosis. *Mol Biol Cell* **15**, 3333-3344 (2004).
26 Neumann, N., Lundin, D. & Poole, A. M. Comparative genomic evidence for a complete nuclear pore complex in the last eukaryotic common ancestor. *PLoS One* **5**, e13241 (2010).
27 Xu, X. M. & Meier, I. The nuclear pore comes to the fore. *Trends Plant Sci* **13**, 20-27 (2008).
28 Wang, J. *et al.* Deep small RNA sequencing from the nematode *Ascaris* reveals conservation, functional diversification, and novel developmental profiles. *Genome Res* **21**, 1462-1477 (2011).
29 Vermeire, J. J., Cho, Y., Lolis, E., Bucala, R. & Cappello, M. Orthologs of macrophage migration inhibitory factor from parasitic nematodes. *Trends Parasitol* **24**, 355-363 (2008).
30 Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 86-90 (1999).
31 Kaur, S., White, T. E., DiGuilio, A. L. & Glavy, J. S. The discovery of a Werner Helicase Interacting Protein (WHIP) association with the nuclear pore complex. *Cell Cycle* **9**, 3106-3111 (2010).
32 Liu, X., Mitchell, J. M., Wozniak, R. W., Blobel, G. & Fan, J. Structural evolution of the membrane-coating module of the nuclear pore complex. *Proc Natl Acad Sci U S A* **109**, 16498-16503 (2012).
33 Bilokapic, S. & Schwartz, T. U. Molecular basis for Nup37 and ELY5/ELYS recruitment to the nuclear pore complex. *Proc Natl Acad Sci U S A* **109**, 15241-15246 (2012).
34 Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071 (2008).
35 Dokudovskaya, S. *et al.* A conserved coatomer-related complex containing Sec13 and Seh1 dynamically associates with the vacuole in *Saccharomyces cerevisiae*. *Mol Cell Proteomics* **10**, M110 006478 (2011).
36 Zuccolo, M. *et al.* The human Nup107-160 nuclear pore subcomplex contributes to proper kinetochore functions. *EMBO J* **26**, 1853-1864 (2007).
37 Ghedin, E. *et al.* Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**, 1756-1760 (2007).
38 Stagg, S. M. *et al.* Structure of the Sec13/31 COPII coat cage. *Nature* **439**, 234-238 (2006).
39 Siniossoglou, S. *et al.* A novel complex of nucleoporins, which includes Sec13p and a Sec13p homolog, is essential for normal nuclear pores. *Cell* **84**, 265-275 (1996).
40 Enninga, J., Levay, A. & Fontoura, B. M. Sec13 shuttles between the nucleus and the cytoplasm and stably interacts with Nup96 at the nuclear pore complex. *Mol Cell Biol* **23**, 7271-7284 (2003).
41 Hsia, K. C., Stavropoulos, P., Blobel, G. & Hoelz, A. Architecture of a coat for the nuclear pore membrane. *Cell* **131**, 1313-1326 (2007).
42 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
43 Onischenko, E. & Weis, K. Nuclear pore complex-a coat specifically tailored for the nuclear envelope. *Curr Opin Cell Biol* **23**, 293-301 (2011).
44 Schneiter, R. *et al.* A yeast acetyl coenzyme A carboxylase mutant links very-long-chain fatty acid synthesis to the structure and function of the nuclear membrane-pore complex. *Mol Cell Biol* **16**, 7161-7172 (1996).
45 Joshi, P. *et al.* The functional interactome landscape of the human histone deacetylase family. *Molecular Systems Biology* **9**, 672 (2013).
46 Cline, M. S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2**, 2366-2382 (2007).
47 Brohawn, S. G. & Schwartz, T. U. Molecular architecture of the Nup84-Nup145C-Sec13 edge element in the nuclear pore complex lattice. *Nat Struct Mol Biol* **16**, 1173-1177 (2009).

SUPPLEMENTARY TABLES

Supplementary Table 1 | Statistics of Y-Nup sequence searches

Name	Query Length	Domain Threshold	Iteration	Structures	Homologs	In database	Redundancy threshold	Above/below minimal support
Nup160	1411	1680	6	3	316	492	77	3/2
Nup133	1200	1500	4	5	320	648	80	1/2
Nup107	845	1200	7	5	340	512	97	2/2
Nup98	1960	2100	10	9	500	733	95	1/1
Nup75	668	1190	4	3	314	489	92	1/4
Nup43	358	500	4	*	102	79	97	1/1
Nup37	320	400	7	*	70	66	90	0/0
Seh1	354	600	6	7	500	228	90	2/4
Sec13	356	600	6	7	500	567	90	–"–
mean | (∑)	830	1085	6	∑=39	∑=2962	∑=3814	90	11/16

Name: name of molecule; Query length: length of query; Domain threshold: length for assessing multi-domain structure of query molecule; Iteration: number of PSI-BLAST iterations until convergence or detection of known false positives; Structures: number of detected similarities to proteins of known structure in PDB (*not in compendium, structures appeared recently); Homologs: number of identified homologs in NRDB, set detection number at 500 for those entries with >500 homologs (underestimate, redundancy-reduced entries are excluded); In database: number of entries when queried for Pfam identifier, includes multiple false assignments; Redundancy threshold: level of sequence identity below which multiple sequence alignments are shown (available in **Data Supplement DS03**); Above/below minimal support: cases supported (or not) by independent evidence – Sec13 entries identical to those for Seh1 (see **Supplementary Table 2**). Mean, average of numerical values; ∑, sum where relevant; not applicable for last column.

Supplementary Table 2 | Support for Y-Nup domain associations

Case	Domain GI	Pfam	by Frequency	by Domain	by Genome	by Expression	Supported by
>01 RAD51–NUP160	322698012	PF11715	2	√		√	
>02 PPOX–NUP160	254572704	–		√			
>03 CCME–NUP160	118587747	PF11715	4	√	√	–	
>04 DUF1668–NUP160	125538033	–		√		–	
>05 AMINOPEP–NUP160	270014242	–		√	√		
>06 ARGI–NUP133	336371647	PF03177		√			
>07 STF2LIKE–NUP133	330922163	–		√	√	–	
>08 AMYLASE–NUP133	66808415	PF03177		√		–	
>09 TPR–NUP107	330890398	–		√		–	
>10 PBPGOBP–NUP107	323520027	PF04121		√	√	–	
>11 ARFSEC7PH–NUP107	353235418	PF04121		√			
>12 DUF1767–NUP107	29841332	PF04121		√		√	
>13 SET–NUP98	302917798	PF00856	3	√	√	–	
>14 DUF1680LECT–NUP98	373462448	PF12110		√		–	
>15 FGREPEATS–NUP75	354546804	–		√		–	
>16 ACETYLCOACBC–NUP75	353238039	PF07575		√			*
>17 ACONITASE–NUP75	322710095	PF07575		√			
>18 SMCX2–NUP75	301093748	PF07575		√		√	**
>19 VITHT1MFS–NUP75	340514810	–		√		–	
>20 MIF–NUP43	49257069	–		√			
>21 DHX15HEL–NUP43	332019512	–	6	√	√	√	
>22 PGM–SEH1	312072423	–		√			
>23 MFS–SEH1	358334778	–		√		–	
>24 PKDOM–SEH1	119113819	–		√			
>25 HCO3–SEH1	300121560	–		√			
>26 TAF9CHS5PARF1P–SEH1	EFW95506.1	–		√	√	√	
>27 CEN192–SEH1	351712025	–	6	√	√	√	

Case: encoding of domains identified (available in **Data Supplement DS07**); Domain GI: GI number of single-domain molecule; Pfam: Pfam identifier of Y-Nup; by Frequency: number of hits – listed only if there are >1 unique GI hits (in **Data Supplement DS09**, filename: "all3.parsed2"), sequence assembly error propagation cannot be excluded; by Domain: support by domain analysis (all); by Genome: support by genomic evidence (see **Methods**); by Expression: support by expression evidence; Supported by: level of support according to the four previous columns (vertical bars). Minimal level of support is represented by single frequency counts (not listed). *Case 16 has minimal support; however, there is some evidence for the involvement of lipids in nuclear pore formation[43,44]. **Case 18 has been recently documented within the context of the histone deacetylase interactome (HDAC11)[45]. Grey dashes signify absence of corresponding homologs in the NGS expression datasets. Light pink background signifies support above minimal level.

Supplementary Table 3 | Rank correlation statistics for NGS of Y-Nups and associated domains

Query	correlates with	Y–Nup	(by genome)	NUP160	NUP133	NUP37	SEH1L	NUP98	NUP107	NUP43	NUP75	SEC13	expression supported
NUP160				1.0	5.0	3.0	11.9	4.0	3.0	6.0	4.0	22.2	
NUP133				6.0	1.0	4.0	2.0	2.0	5.0	2.0	10.0	28.3	
NUP37				4.0	4.0	1.0	16.9	5.0	2.0	4.0	8.0	24.2	
SEH1L				22.0	15.0	21.9	1.0	27.0	18.0	14.9	32.1	53.2	
NUP98				19.0	18.0	18.0	27.0	1.0	20.9	17.9	24.1	19.2	
NUP107				2.0	3.0	2.0	4.0	7.0	1.0	5.0	9.0	35.3	
NUP43				7.0	2.0	7.0	3.0	3.0	13.0	1.0	19.0	32.3	
NUP85				16.0	16.0	17.0	21.9	11.0	17.0	18.9	1.0	7.0	
SEC13				50.0	53.0	47.0	65.9	40.9	56.0	52.9	31.1	1.0	
RAD51	RAD51	NUP160	T+	11.0	9.0	9.0	13.9	6.0	8.0	7.0	7.0	25.2	√
PPOX	PPOX	NUP160	T–	66.0	59.0	60.0	70.8	60.9	67.0	64.8	36.1	6.0	
ERAP2	AMINOPEP	NUP160	F+	128.8	129.8	125.8	131.0	125.8	129.8	128.8	130.8	102.0	
LNPEP	AMINOPEP	NUP160	F+	107.7	112.7	109.6	108.8	107.6	109.7	107.7	122.8	130.9	
AGMAT	ARGI	NUP133	T–	58.0	69.0	56.0	75.8	93.6	75.8	67.8	55.0	30.3	
ARG2	ARGI	NUP133	F–	56.0	43.9	46.0	47.0	53.0	48.9	42.9	51.1	23.2	
ARG1	ARGI	NUP133	T–	122.8	128.8	121.7	122.8	132.8	120.8	125.7	127.8	135.9	
PSD	ARFSEC7PH	NUP107	T–	100.7	86.7	101.8	84.9	95.6	95.6	89.7	79.9	70.1	
PSD4	ARFSEC7PH	NUP107	T–	110.7	118.7	103.7	123.8	124.8	115.7	120.7	98.0	59.2	
PSD2	ARFSEC7PH	NUP107	T–	86.0	78.9	92.7	51.0	99.5	76.8	83.7	75.8	88.1	
TDRD3	DUF1767	NUP107	T+	42.0	45.9	40.0	53.9	43.9	38.9	47.9	72.0	111.0	√
MCCC1	ACETYLCOACBC	NUP75	T–	92.8	88.7	76.8	87.8	135.9	85.9	97.6	90.0	82.0	
ACACA	ACETYLCOACBC	NUP75	F–	25.9	22.9	26.0	14.9	31.0	23.9	25.0	21.0	21.2	
ACACB	ACETYLCOACBC	NUP75	T–	136.9	137.0	135.9	134.9	137.8	136.9	134.9	137.8	120.8	
ACO2	ACONITASE	NUP75	T–	126.8	111.7	122.7	111.7	130.8	124.8	116.7	116.7	43.1	
ACO1	ACONITASE	NUP75	T–	89.9	96.7	82.9	105.8	94.6	94.6	81.7	109.9	98.0	
SMC4	SMCX2	NUP75	F–	5.0	11.0	6.0	17.9	10.0	6.0	8.0	13.1	31.3	√
SMC6	SMCX2	NUP75	F–	14.0	13.0	8.0	18.9	16.9	7.0	9.0	16.1	42.1	√
SMC2	SMCX2	NUP75	F–	9.0	6.0	10.0	5.9	18.9	12.0	11.0	17.1	34.3	√
SMC3	SMCX2	NUP75	F–	12.0	12.0	13.0	10.9	15.9	11.0	15.9	30.1	56.2	√
SMC1A	SMCX2	NUP75	F–	8.0	10.0	14.0	15.9	19.9	14.0	12.0	2.0	14.1	√
MIF	MIF	NUP43	T–	75.8	76.0	78.8	106.8	85.8	92.6	88.7	41.0	2.0	
DHX15	DHX15HEL	NUP43	T+	3.0	7.0	5.0	6.9	9.0	4.0	3.0	3.0	29.3	√
PGM2	PGM	SEH1	F–	24.0	27.0	20.0	46.0	26.0	29.0	31.0	26.0	46.2	
PGM2L1	PGM	SEH1	F–	55.0	50.9	57.0	34.9	59.0	47.9	48.9	63.0	110.0	
PGM1	PGM	SEH1	T–	114.7	116.7	112.7	125.9	116.7	118.8	111.7	128.8	97.0	
PXK	PKDOM	SEH1	T–	124.8	117.7	127.8	109.8	108.6	122.8	123.7	118.8	94.0	
SLC4A8	HCO3	SEH1	F–	65.0	57.0	66.9	37.8	74.9	57.0	58.0	56.0	91.0	
SLC4A10	HCO3	SEH1	T–	105.7	106.6	110.6	80.9	118.7	101.8	100.6	105.9	132.9	
TAF9	TAF9CHS5PARF1P	SEH1	F–	18.0	20.0	16.0	9.9	12.0	16.0	22.0	15.1	33.3	√
TAF9B	TAF9CHS5PARF1P	SEH1	F–	41.0	44.9	42.9	20.9	44.9	39.9	41.9	65.0	121.8	√
CEP192	CEN192	SEH1	T+	10.0	8.0	11.0	7.9	21.9	10.0	10.0	5.0	39.2	√

Query: first molecule under consideration; correlates with: signifies the Y-Nup or corresponding domain found in query molecule; Y-Nup: second molecule (always a Y-Nup) under consideration; by genome: support by genome comparison, T+ (dual support), F- (not supported by genome comparison, only by expression); names of Y-Nups: ranging from NUP160 to SEC13 (order modified according to clustering, does not follow rest of analysis); Expression supported: final verdict for expression support of associations (all eleven instances are shown in **Figure 3B** with correlation values). *Color scheme*: Dark green – self-comparison of Y-Nups; Light green – cross-comparison of Y-Nups; Yellow – marks most distant Y-Nup query in cross-comparison, serves as threshold for bottom section of table i.e. domain associations; Grey – query ranks above threshold except domain-supported cases; White – query ranks below threshold, not considered further; Blue shades signify levels of support by expression for query: petrol blue – supported as ranking above threshold for human ortholog, sky blue – supported as ranking above threshold for all paralogs, ice blue – not supported as ranking below threshold (grey font) or above threshold (purple font) for one paralog only. For details, see **Methods**.

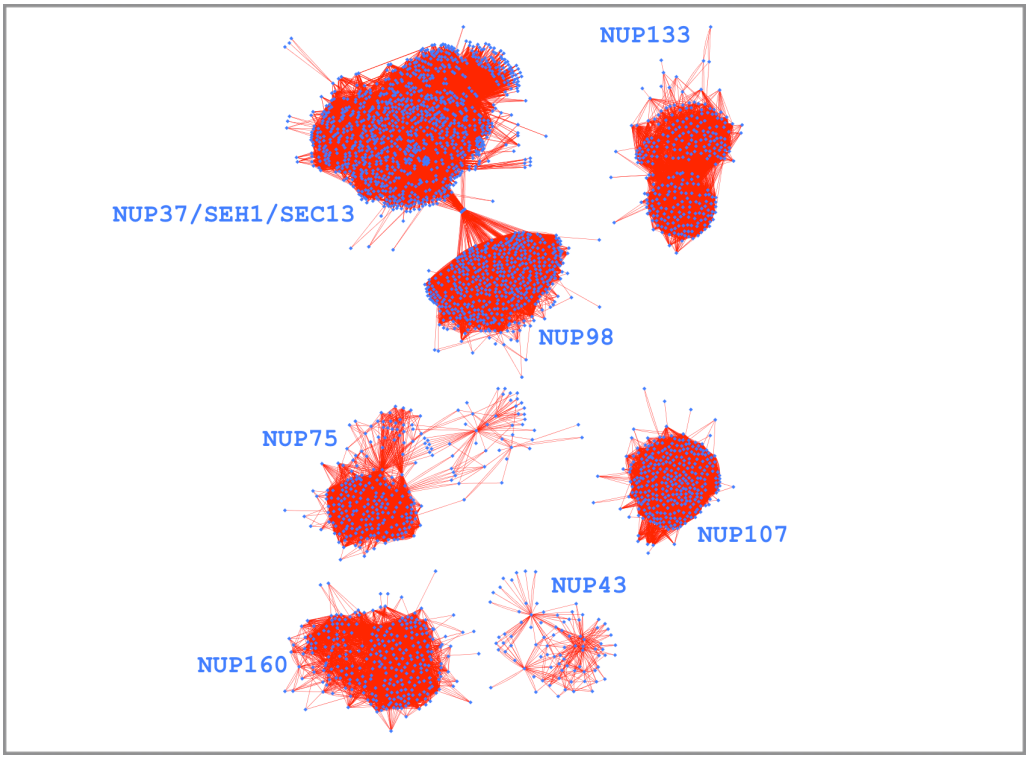
Supplementary Table 4 | Next-generation sequencing (NGS) data used in this study

Type	#	SAMPLE	Read-Length	# reads	Source (SRA)	Reference/GEO
HUMAN						
ESC	1	H1-a	50	112,504,744	SRR065492,3; SRR065504; SRR065526; SRR066678	GSM591680
	2	H1-b	50	189,880,016	SRR031628-34	PMID: 20944595
	3	H9-a	50	79,225,971	SRR345211-3	PMID: 21924763
	4	H9-b	50	39,346,501	SRR067388-90	PMID: 21324177
	5	hESC2	50 (SOLiD)	270,903,690	SRR094594-600	PMID: 22042643
iPS	1	iPS-a	50	134,836,243	SRR350717	PMID: 21915259
	2	iPS-b	50 (SOLiD)	146,307,181	SRR179588-90	GSM706050
Lines	1	HNEK	50	136,125,754	SRR315327-9	GSM765401
	2	HUVEK	50	64,464,576	SRR307905,6	GSM758563
	3	HepG2	50	63,046,870	SRR065512,3,23	GSM591677
	4	MCF7	50	87,432,105	SRR315301,2	GSM765388
	5	Gm12878	50	118,526,691	SRR065510,14,15,32	GSM591664
	6	HeLa	50	85,935,538	SRX255057	GSE45505
	7	293T	50	67,568,978	SRX255055	GSE45505
Tissues	1	Whole Brain	50	137,826,251	ERR030882,90	Human BodyMap2
	2	Cortex	50	115,601,548	SRR306838-43	PMID: 22012392
	3	Cerebellum	50	79,453,779	SRR306844-6	PMID: 22012392
	4	Liver-a	50	157,502,500	ERR030887,95	Human BodyMap2
	5	Liver-b	50	67,013,560	SRR306854-6	PMID: 22012392
	6	Kidney-a	50	160,169,730	ERR030885,93	Human BodyMap2
	7	Kidney-b	50	74,564,889	SRR306851-3	PMID: 22012392
	8	Heart-a	50	159,685,646	ERR030886,94	Human BodyMap2
	9	Heart-b	50	80,222,268	SRR306847-50	PMID: 22012392
	10	Muscle	50	164,975,775	ERR030876,99	Human BodyMap2
	11	Testis	50	163,880,518	ERR030873,903	Human BodyMap2
	13	Adipose	50	153,569,297	ERR030880,8	Human BodyMap2
	14	Adrenal	50	150,644,440	ERR030881,9	Human BodyMap2
	15	Breast	50	153,057,475	ERR030883,91	Human BodyMap2
	16	Colon	50	162,695,200	ERR030884,92	Human BodyMap2
	17	Lung	50	160,552,343	ERR030879,96	Human BodyMap2
	18	Lymph node	50	163,994,617	ERR030878,97	Human BodyMap2
	19	Ovary	50	161,949,312	ERR030874,901	Human BodyMap2
	20	Prostate	50	165,653,978	ERR030877,98	Human BodyMap2
	21	Thyroid	50	162,159,544	ERR030972,903	Human BodyMap2
	22	WBC	50	164,002,821	ERR030875,900	Human BodyMap2
MOUSE						
ESC	1	v6.5	50	58,065,395	SRR039999-40001	PMID: 20436462
	2	OS25	50	79,984,743	SRR391029-31	PMID: 22305566
	3	J1	50	43,040,267	SRR210861	PMID: 21624812
	4	E14	50	69,972,477	SRR414942,3	GSM881355
	5	D3	50	62,345,902	SRR331051,3	PMID: 21884934
iPS	1	iPS-a	50	19,262,401	SRR611665	PMID: 23217423
	2	iPS-b	50	47,314,841	SRR611677	PMID: 23217423
Lines	1	MEF	50	55,325,338	SRR611704	PMID: 23217423
	2	C2C12	50	42,184,539	SRR037945,6	PMID: 20436464
	3	N2A	36	106,434,928	Upon request	GSE45505
	4	3T3	36	133,878,952	SRR149235-7, SRR149707-9, SRR390297	PMID: 21593866
Embryonic	1	Embr_Limb_14dpc	50	170,903,264	SRR567490	GSM1000568
	2	Embr_CNS_14dpc	50	221,581,975	SRR567493	GSM1000569
	3	Embr_CNS_11dpc	50	204,144,434	SRR567500	GSM1000573
Tissues	1	Adrenal	50	148,300,000	SRR453116-21	GSM900188
	2	Bladder	50	161,644,204	SRR567482	GSM1000564
	3	Brain_FrontalLobe	50	185,745,131	SRR567478	GSM1000562
	4	Brain-a	50	141,111,481	SRR579545,6	PMID: 23258890
	5	Brain-b	50	120,991,273	SRR306757-62	PMID: 22012392
	6	Cerebellum	50	90,998,774	SRR306763-5	PMID: 22012392
	7	Colon	50	131,005,753	SRR453166-71	GSM900198
	8	Duodenum	50	155,900,000	SRR453109-15	GSM900187
	9	GenitalFatPad	50	148,000,000	SRR453126-9	GSM900190
	10	Heart-a	50	38,701,291	SRR579549	PMID: 23258890
	11	Heart-b	50	95,066,626	SRR306766-8	PMID: 22012392
	12	Kidney-a	50	42,182,481	SRR579548	PMID: 23258890
	13	Kidney-b	50	98,411,330	SRR306769-71	PMID: 22012392
	14	Liver-a	50	68,224,409	SRR579547	PMID: 23258890
	15	Liver-b	50	100,761,351	SRR306772-4	PMID: 22012392
	16	Lung	50	133,066,159	SRR453156-9	GSM900196
	17	MammGland	50	147,000,000	SRR453087-92	GSM900184
	18	Muscle	50	79,026,285	SRR579550	PMID: 23258890
	19	Ovary	50	105,700,000	SRR453077-86	GSM900183
	20	Placenta	50	237,925,260	SRR567485	GSM1000565
	21	Retina	50	104,168,856	SRR327045,7, SRR342457,8	PMID: 21775302
	22	SmIntest	50	138,100,000	SRR453099-108	GSM900186
	23	Spleen	50	152,594,296	SRR453160-5	GSM900197
	24	Stomach	50	168,800,000	SRR453093-8	GSM900185
	25	SubcFatPad	50	157,200,000	SRR453130-3	GSM900191
	26	T-cells	36	157,379,633	SRR210844-8	PMID: 21765417
	27	Testis	50	49,736,678	SRR306775-6	PMID: 22012392
	28	Thymus	50	181,600,000	SRR453134-9	GSM900192

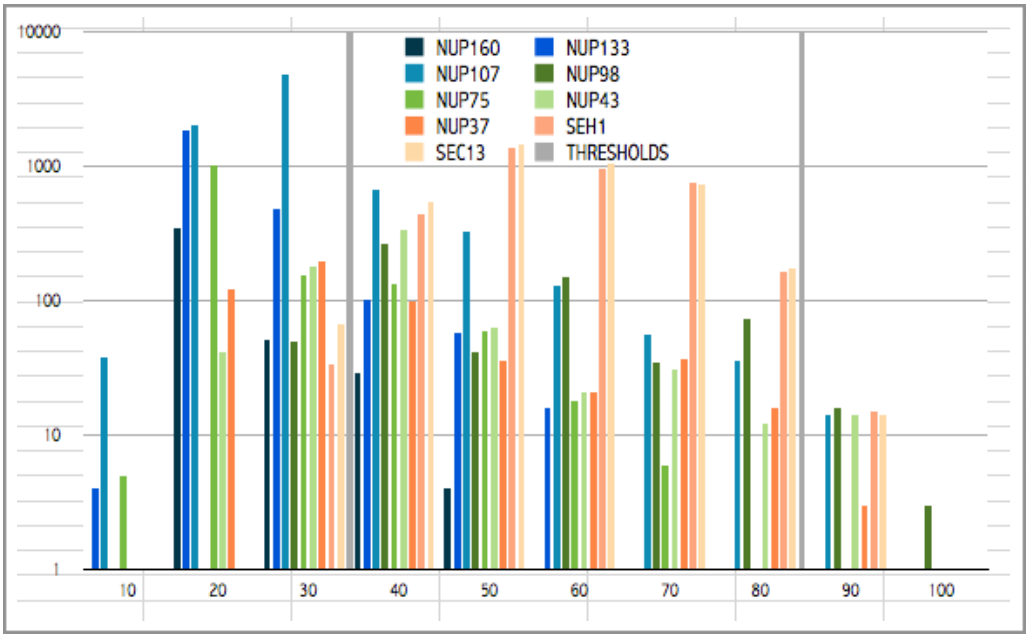
Type: for human and mouse, embryonic stem cells (ESC), induced pluripotent stem cells (iPS), cell lines, embryonic tissue (for mouse), tissues (various); #: sample number; Sample: description of sample; Read length: expected read length; # reads: total number of reads per sample; Source (SRA): short read archive accession number; Reference/GEO: PMID PubMed identifier, additional references from the GEO archive and the Human Body Map2 collection.

SUPPLEMENTARY FIGURES

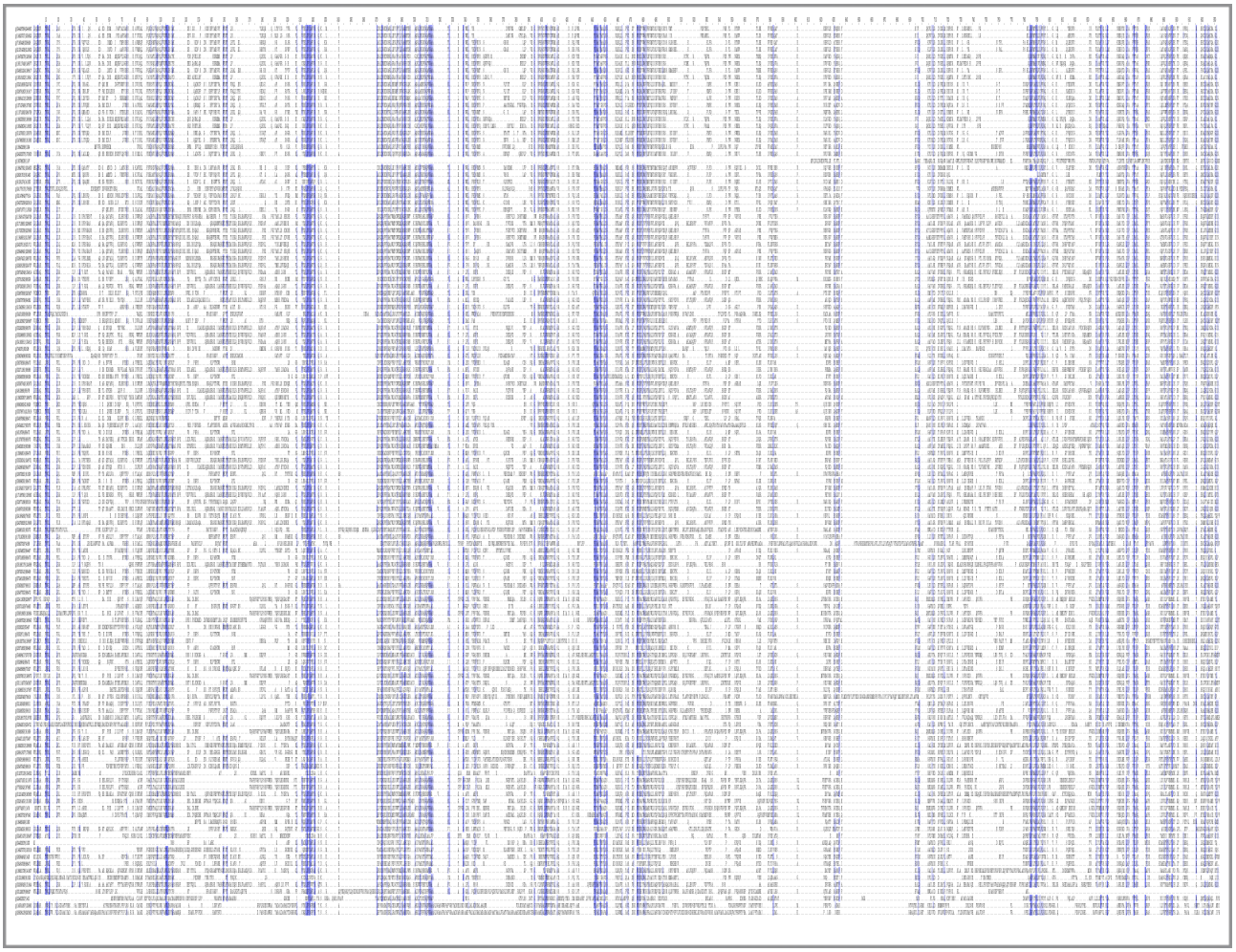
Supplementary Figure 1: Sequence similarity relationships of Y-Nups from the all-against-all comparison of detected homologs. For visualization purposes, only pairwise sequence identities <40% are depicted, using the organic layout diagram mode of Cytoscape[46] (**Data Supplement DS09**). It is evident that the most dispersed group is represented by NUP75, which results in two sub-groups following clustering and elimination of false positives using MCL (**Figure 1**). The node connecting the NUP37/SEH1/SEC13 and NUP98 clusters corresponds to the PDB database entry 3JRO[47], an artificial fusion protein of SEC13 and NUP98 of *Saccharomyces cerevisiae* (Nup145C).



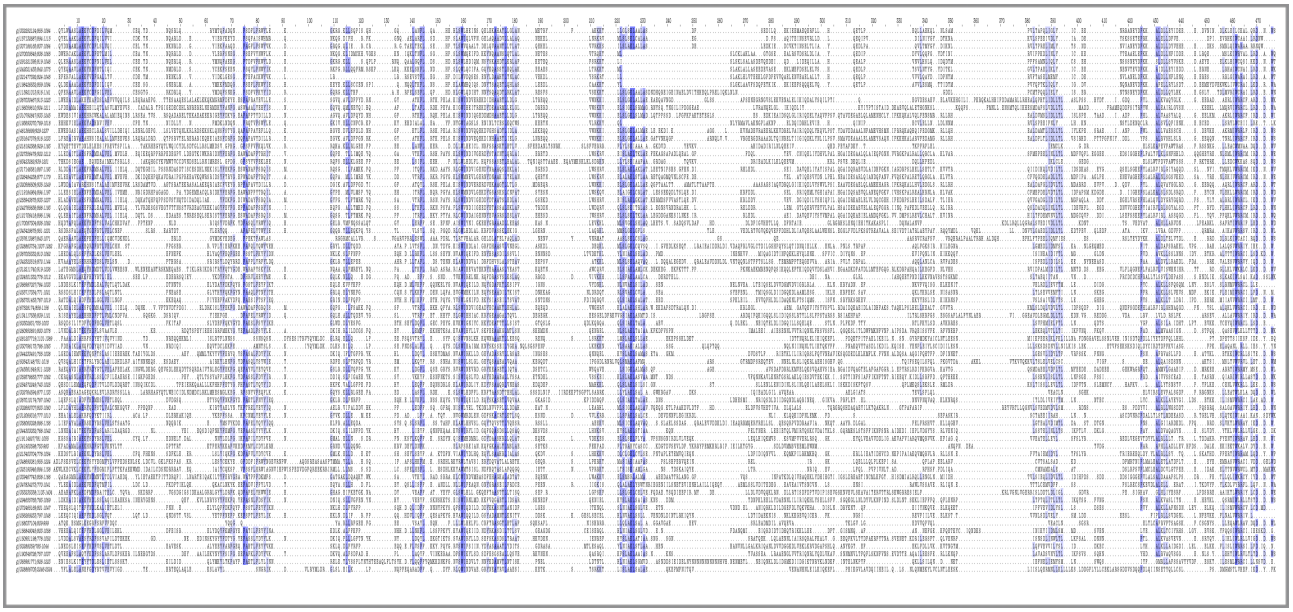
Supplementary Figure 2: Sequence identity frequency distributions of profile searches. Sequence identity bins (ranging from 0 to 100%) are shown on the x-axis, count of the nine Y-Nup families (color-coded) are shown on the y-axis (logarithmic scale). Thresholds at >30 and ≥80% sequence identity are also shown (grey). All identity counts refer to redundancy-reduced alignments (**Data Supplement DS03**), displayed data provided (**Data Supplement DS04**).



Supplementary Figure 3: Multiple sequence alignment of Nup107 homologs. Sequence redundancy threshold 97% (**Supplementary Table 1**), data provided (**Data Supplement DS03**).



Supplementary Figure 4: Multiple sequence alignment of Nup133 homologs. Sequence redundancy threshold 80% (**Supplementary Table 1**), data provided (**Data Supplement DS03**).



Supplementary Figure 5: Multi-domain architecture of Y-Nup homologs. Y-Nup domains identified by sequence searches (shown in light blue) found to be associated with protein domains with highly supported moonlighting functions (shown in green, support level 3+), other weakly supported functions (shown in yellow, support level 2), unrelated functions (shown in red, support level 1) – see also **Supplementary Table 2**. Grey boxes signify other protein regions. Scale is provided above and below the multi-domain diagram representations, on both panels.