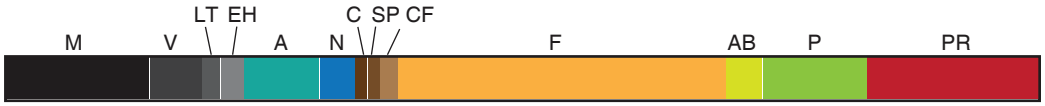
Figure S1



No.	Species	Branch	Branch name	Code
1	Homo sapiens	1	Mammals	M
2	Pan troglodytes	1	Mammals	M
3	Gorilla gorilla gorilla	1	Mammals	M
4	Pongo abelii	1	Mammals	M
5	Nomascus leucogenys	1	Mammals	M
6	Macaca mulatta	1	Mammals	M
7	Papio anubis	1	Mammals	M
8	Saimiri boliviensis	1	Mammals	M
9	Callithrix jacchus	1	Mammals	M
10	Otolemur garnettii	1	Mammals	M
11	Mus musculus	1	Mammals	M
12	Cricetulus griseus	1	Mammals	M
13	Rattus norvegicus	1	Mammals	M
14	Cavia porcellus	1	Mammals	M
15	Oryctolagus cuniculus	1	Mammals	M
16	Canis lupus familiaris	1	Mammals	M
17	Ailuropoda melanoleuca	1	Mammals	M
18	Equus caballus	1	Mammals	M
19	Sus scrofa	1	Mammals	M
20	Bos taurus	1	Mammals	M
21	Trichechus manatus latirostris	1	Mammals	M
22	Loxodonta africana	1	Mammals	M
23	Monodelphis domestica	1	Mammals	M
24	Sarcophilus harrisii	1	Mammals	M
25	Ornithorhynchus anatinus	1	Mammals	M
26	Gallus gallus	2	Other vertebrates	V
27	Taeniopygia guttata	2	Other vertebrates	V
28	Anolis carolinensis	2	Other vertebrates	V
29	Chelonia mydas	2	Other vertebrates	V
30	Xenopus (Silurana) tropicalis	2	Other vertebrates	V
31	Danio rerio	2	Other vertebrates	V
32	Oreochromis niloticus	2	Other vertebrates	V
33	Oryzias latipes	2	Other vertebrates	V
34	Tetraodon nigroviridis	3	Lancelets/tunicates	LT
35	Branchiostoma floridae	3	Lancelets/tunicates	LT
36	Ciona intestinalis	3	Lancelets/tunicates	LT
37	Oikopleura dioica	3	Lancelets/tunicates	LT
38	Strongylocentrotus purpuratus	4	Echinoderms/hemichordata	EH
39	Crassostrea gigas	4	Echinoderms/hemichordata	EH
40	Saccoglossus kowalevskii	4	Echinoderms/hemichordata	EH
41	Capitella teleta	4	Echinoderms/hemichordata	EH
42	Acyrthosiphon pisum	5	Arthropods	A
43	Pediculus humanus corporis	5	Arthropods	A
44	Ixodes scapularis	5	Arthropods	A
45	Daphnia pulex	5	Arthropods	A
46	Drosophila melanogaster	5	Arthropods	A
47	Tribolium castaneum	5	Arthropods	A
48	Apis mellifera	5	Arthropods	A
49	Bombus terrestris	5	Arthropods	A
50	Anopheles gambiae str. PEST	5	Arthropods	A
51	Bombyx mori	5	Arthropods	A
52	Aedes aegypti	5	Arthropods	A
53	Culex quinquefasciatus	5	Arthropods	A
54	Nasonia vitripennis	5	Arthropods	A
55	Caenorhabditis elegans	6	Nematodes	N
56	Schistosoma mansoni	6	Nematodes	N
57	Brugia malayi	6	Nematodes	N
58	Clonorchis sinensis	6	Nematodes	N
59	Loa loa	6	Nematodes	N
60	Trichinella spiralis	6	Nematodes	N
61	Nematostella vectensis	7	Cnidaria	C
62	Hydra magnipapillata	7	Cnidaria	C
63	Trichoplax adhaerens	8	Sponge/Placozoa	SP
64	Amphimedon queenslandica	8	Sponge/Placozoa	SP
65	Salpingoeca sp.atcc50818	9	Choanoflagellates	CF
66	Capsaspora owczarzaki	9	Choanoflagellates	CF
67	Monosiga brevicollis mx1	9	Choanoflagellates	CF
68	Batrachochytrium dendrobatidis jam81	10	Fungi	F
69	Rhizopus delemar	10	Fungi	F
70	Eremothecium cymbalariae DBVPG#7215	10	Fungi	F
71	Ashbya gossypii ATCC 10895	10	Fungi	F
72	Torulaspora delbrueckii	10	Fungi	F
73	Encephalitozoon cuniculi GB-M1	10	Fungi	F
74	Podospora anserina S mat+	10	Fungi	F
75	Verticillium albo-atrum VaMs.102	10	Fungi	F
76	Candida albicans	10	Fungi	F
77	Neurospora crassa OR74A	10	Fungi	F
78	Sordaria macrospora	10	Fungi	F
79	Zygosaccharomyces rouxii CBS 732	10	Fungi	F
80	Rhodosporidium toruloides np11	10	Fungi	F
81	Myceliophthora thermophila ATCC 42464	10	Fungi	F
82	Naumovozyma castellii CBS 4309	10	Fungi	F
83	Vanderwaltozyma polyspora DSM 70294	10	Fungi	F
84	Talaromyces stipitatus ATCC 10500	10	Fungi	F
85	Pneumocystis jirovecii	10	Fungi	F
86	Botryotinia fuckeliana	10	Fungi	F
87	Puccinia graminis f. sp. tritici	10	Fungi	F
88	Thielavia terrestris NRRL 8126	10	Fungi	F
89	Nosema ceranae BRL01	10	Fungi	F
90	Melampsora larici-populina 98ag31	10	Fungi	F
91	Gibberella zeae PH-1 (Fusarium)	10	Fungi	F
92	Chaetomium globosum CBS 148.51	10	Fungi	F
93	Magnaporthe oryzae 70-15	10	Fungi	F
94	Tuber melanosporum	10	Fungi	F
95	Schizosaccharomyces pombe	10	Fungi	F
96	Sclerotinia sclerotiorum 1980 UF-70	10	Fungi	F
97	Schizophyllum commune H4-8	10	Fungi	F
98	Neosartorya fischeri NRRL 181	10	Fungi	F
99	Phaeosphaeria nodorum SN15	10	Fungi	F
100	Postia placenta	10	Fungi	F
101	Ustilago maydis	10	Fungi	F
102	Enterocytozoon bieneusi h348	10	Fungi	F
103	Malassezia globosa	10	Fungi	F
104	Pyrenophora teres f. teres 0-1	10	Fungi	F
105	Aspergillus nidulans FGSC A4	10	Fungi	F
106	Moniliophthora perniciosa FA553	10	Fungi	F
107	Laccaria bicolor s238n-h82	10	Fungi	F
108	Cryptococcus neoformans jec21	10	Fungi	F
109	Ajellomyces dermatitidis SLH14081	10	Fungi	F
110	Uncinocarpus reesii 1704	10	Fungi	F
111	Nectria haematococca mpVI 77-13-4	10	Fungi	F
112	Arthroderma gypseum CBS 118893	10	Fungi	F
113	Trichophyton rubrum CBS 118892	10	Fungi	F
114	Paracoccidioides brasiliensis Pb01	10	Fungi	F
115	Clavispora lusitaniae ATCC 42720	10	Fungi	F
116	Coccidioides posadasii C735 delta SOWgp	10	Fungi	F
117	Komagataella pastoris GS115	10	Fungi	F
118	Meyerozyma guilliermondii ATCC 6260	10	Fungi	F
119	Tetrapisispora phaffii CBS 4417	10	Fungi	F
120	Lachancea thermotolerans CBS 6340	10	Fungi	F
121	Scheffersomyces stipitis CBS 6054	10	Fungi	F
122	Yarrowia lipolytica	10	Fungi	F
123	Saccharomyces cerevisiae s288c	10	Fungi	F
124	Debaryomyces hansenii CBS767	10	Fungi	F
125	Polysphondylium pallidum pn500	11	Amoebozoa	AB
126	Entamoeba dispar SAW760	11	Amoebozoa	AB
127	Acanthamoeba castellanii str.neff	11	Amoebozoa	AB
128	Entamoeba histolytica hm-1:imss	11	Amoebozoa	AB
129	Dictyostelium discoideum ax4	11	Amoebozoa	AB
130	Medicago truncatula	12	Plantae	P
131	Vitis vinifera	12	Plantae	P
132	Glycine max	12	Plantae	P
133	Brachypodium distachyon	12	Plantae	P
134	Zea mays	12	Plantae	P
135	Populus trichocarpa	12	Plantae	P
136	Oryza sativa japonica group	12	Plantae	P
137	Ricinus communis	12	Plantae	P
138	Sorghum bicolor	12	Plantae	P
139	Arabidopsis thaliana	12	Plantae	P
140	Selaginella moellendorffii	12	Plantae	P
141	Physcomitrella patens subsp.patens	12	Plantae	P
142	Chlamydomonas reinhardtii	12	Plantae	P
143	Volvox carteri	12	Plantae	P
144	Coccomyxa subellipsoidea c-169	12	Plantae	P
145	Ostreococcus tauri	12	Plantae	P
146	Micromonas pusilla ccmp1545	12	Plantae	P
147	Chlorella variabilis	12	Plantae	P
148	Naegleria gruberi	13	Other protists	PR
149	Ichthyophthirius multifiliis	13	Other protists	PR
150	Hemiselmis andersenii	13	Other protists	PR
151	Ectocarpus siliculosus	13	Other protists	PR
152	Oxytricha trifallax	13	Other protists	PR
153	Bigelowiella natans	13	Other protists	PR
154	Guillardia theta ccmp2712	13	Other protists	PR
155	Aureococcus anophagefferens	13	Other protists	PR
156	Blastocystis hominis	13	Other protists	PR
157	Neospora caninum liverpool	13	Other protists	PR
158	Babesia bovis	13	Other protists	PR
159	Theileria parva	13	Other protists	PR
160	Phaeodactylum tricornutum	13	Other protists	PR
161	Thalassiosira oceanica	13	Other protists	PR
162	Cyanidioschyzon merolae	13	Other protists	PR
163	Cryptomonas paramecium	13	Other protists	PR
164	Galdieria sulphuraria	13	Other protists	PR
165	Phytophthora infestans	13	Other protists	PR
166	Leishmania major	13	Other protists	PR
167	Giardia lamblia	13	Other protists	PR
168	Tetrahymena thermophila	13	Other protists	PR
169	Perkinsus marinus	13	Other protists	PR
170	Cryptosporidium parvum	13	Other protists	PR
171	Toxoplasma gondii me49	13	Other protists	PR
172	Plasmodium falciparum 3d7	13	Other protists	PR
173	Nematocida parisii	13	Other protists	PR
174	Paramecium tetraurelia	13	Other protists	PR
175	Trichomonas vaginalis	13	Other protists	PR
176	Trypanosoma cruzi	13	Other protists	PR
177	Paulinella chromatophora	13	Other protists	PR

Figure S2

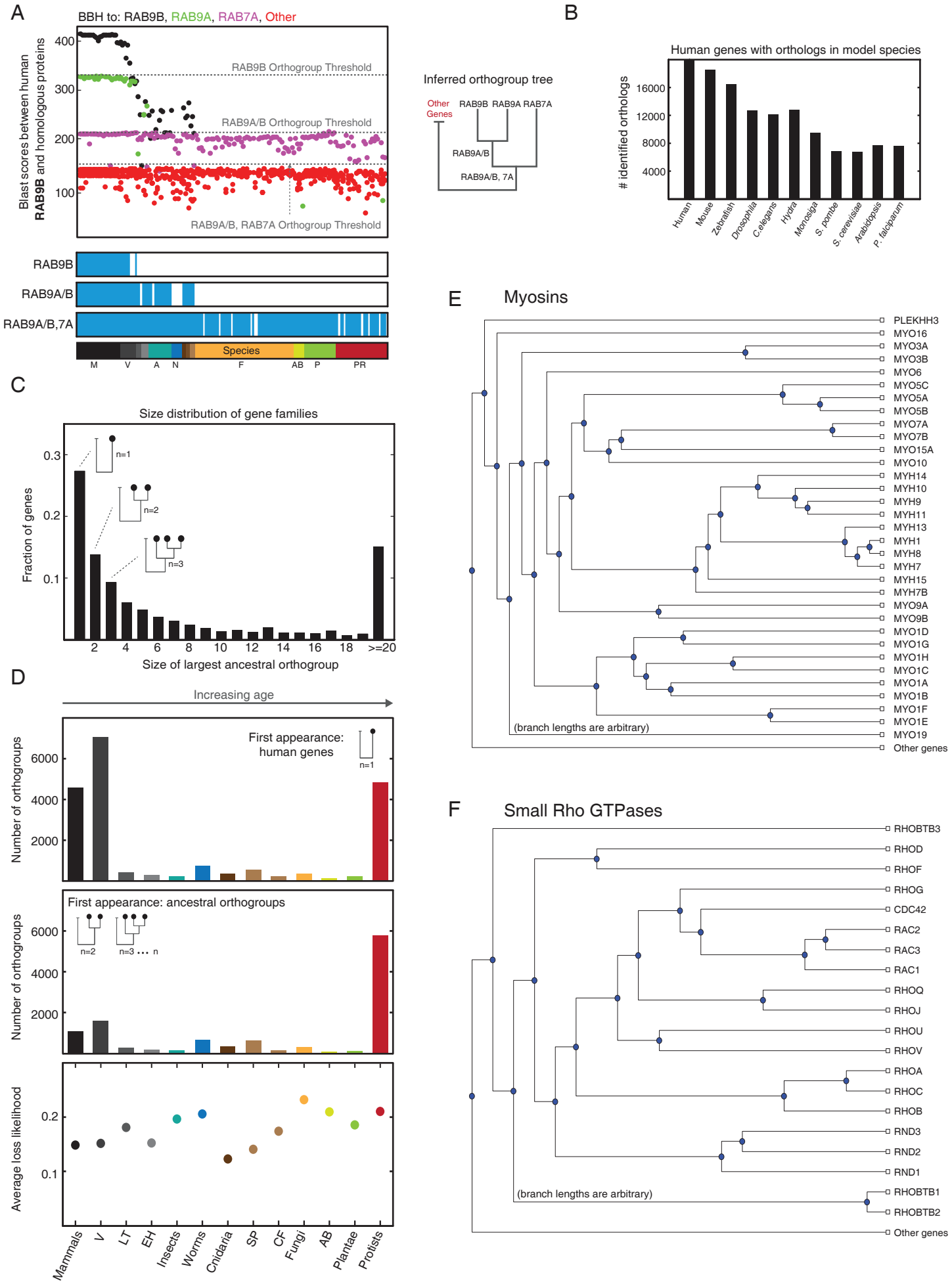
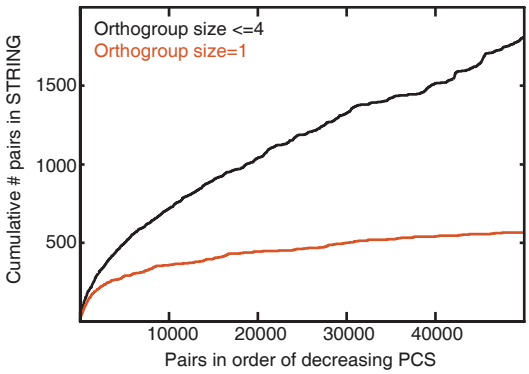
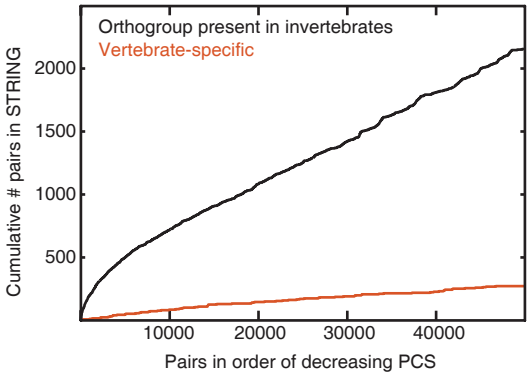

A
BBH to: RAB9B, RAB9A, RAB7A, Other
Blast scores between human RAB9B and homologous proteins
400
300
200
100
RAB9B Orthogroup Threshold
RAB9A/B Orthogroup Threshold
RAB9A/B, RAB7A Orthogroup Threshold
RAB9B
RAB9A/B
RAB9A/B,7A
Species
M V A N F AB P PR
Inferred orthogroup tree
Other Genes
RAB9B RAB9A RAB7A
RAB9A/B
RAB9A/B, 7A
B
Human genes with orthologs in model species
identified orthologs
16000
12000
8000
4000
Human
Mouse
Zebrafish
Drosophila
C.elegans
Hydra
Monosiga
S. pombe
S. cerevisiae
Arabidopsis
P. falciparum
C
Size distribution of gene families
Fraction of genes
0.3
0.2
0.1
n=1
n=2
n=3
2 4 6 8 10 12 14 16 18 >=20
Size of largest ancestral orthogroup
D
Increasing age
Number of orthogroups
6000
4000
2000
First appearance: human genes
n=1
First appearance: ancestral orthogroups
n=2
n=3 ... n
Average loss likelihood
0.2
0.1
Mammals V LT EH Insects Worms Cnidaria SP CF Fungi AB Plantae Protists
E
Myosins
PLEKHH3
MYO16
MYO3A
MYO3B
MYO6
MYO5C
MYO5A
MYO5B
MYO7A
MYO7B
MYO15A
MYO10
MYH14
MYH10
MYH9
MYH11
MYH13
MYH1
MYH8
MYH7
MYH15
MYH7B
MYO9A
MYO9B
MYO1D
MYO1G
MYO1H
MYO1C
MYO1A
MYO1B
MYO1F
MYO1E
MYO19
Other genes
(branch lengths are arbitrary)
F
Small Rho GTPases
RHOBTB3
RHOD
RHOF
RHOG
CDC42
RAC2
RAC3
RAC1
RHOQ
RHOJ
RHOU
RHOV
RHOA
RHOC
RHOB
RND3
RND2
RND1
RHOBTB1
RHOBTB2
Other genes
(branch lengths are arbitrary)

Figure S3

A



B



C

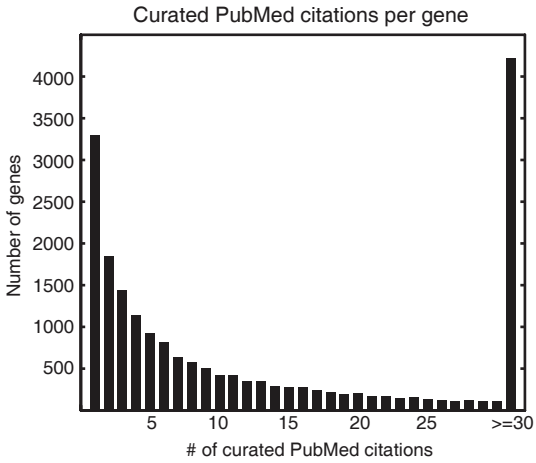


Figure S4

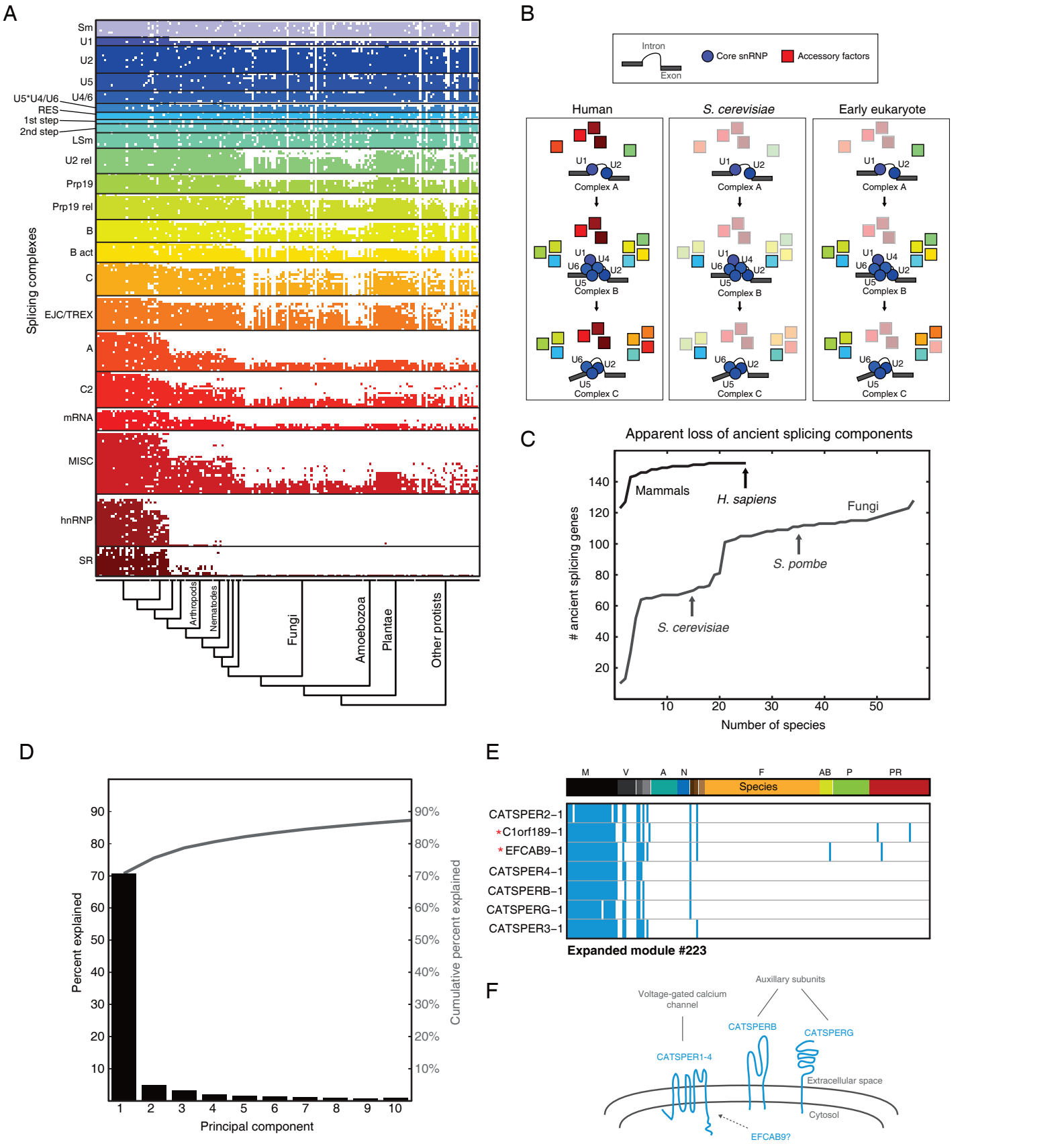

A
Sm
U1
U2
U5
U5*U4/U6
U4/6
RES
1st step
2nd step
LSm
U2 rel
Prp19
Prp19 rel
B
B act
C
EJC/TREX
A
C2
mRNA
MISC
hnRNP
SR
Splicing complexes
Arthropods
Nematodes
Fungi
Amoebozoa
Plantae
Other protists
B
Intron
Exon
Core snRNP
Accessory factors
Human
S. cerevisiae
Early eukaryote
Complex A
Complex B
Complex C
C
Apparent loss of ancient splicing components
Mammals
H. sapiens
Fungi
S. pombe
S. cerevisiae
ancient splicing genes
Number of species
D
Percent explained
Cumulative percent explained
Principal component
E
M
V
A
N
F
AB
P
PR
Species
CATSPER2−1
*C1orf189−1
*EFCAB9−1
CATSPER4−1
CATSPERB−1
CATSPERG−1
CATSPER3−1
Expanded module #223
F
Auxillary subunits
Voltage-gated calcium channel
CATSPERB
CATSPERG
CATSPER1-4
Extracellular space
Cytosol
EFCAB9?

Figure S5

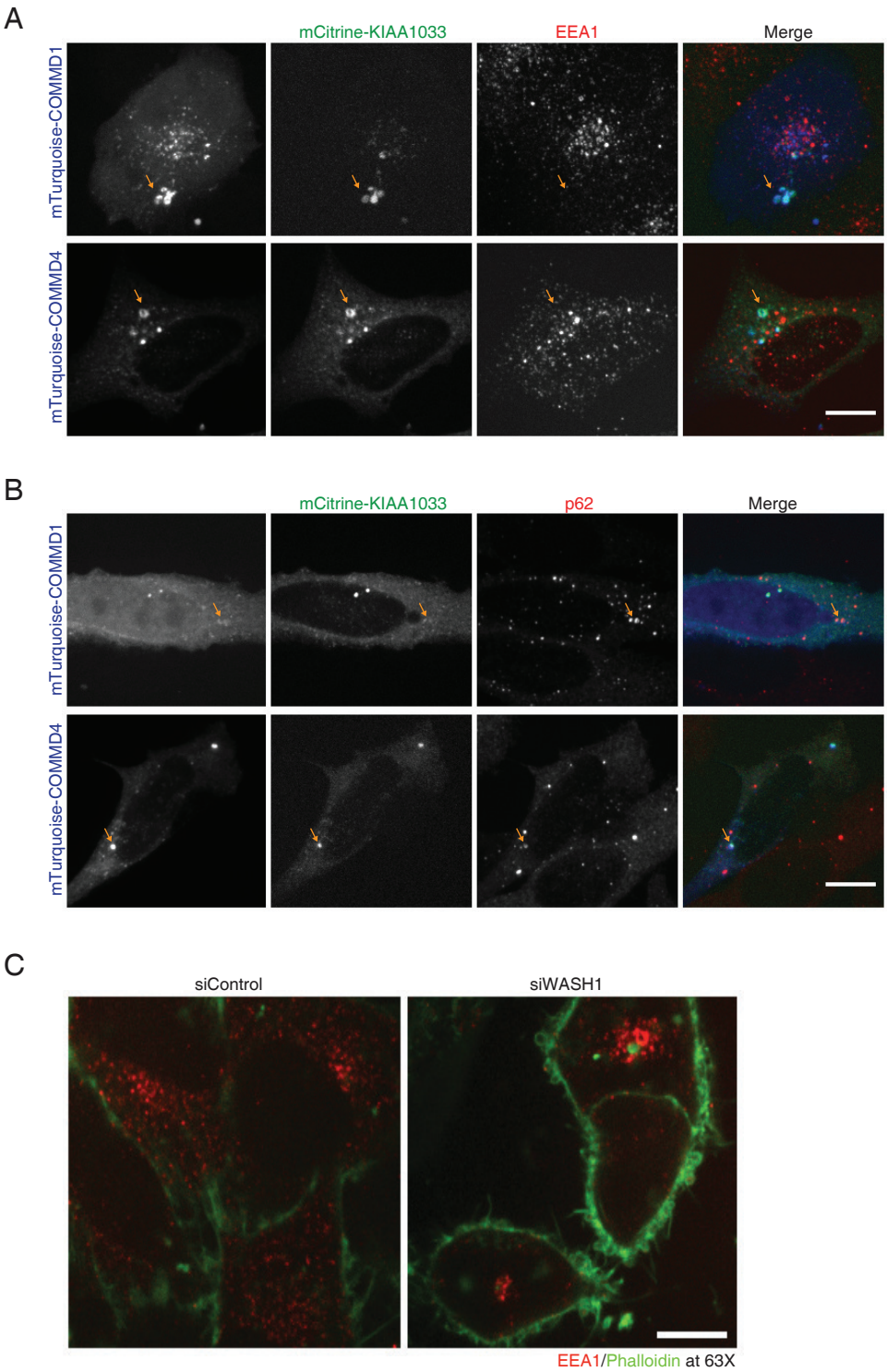
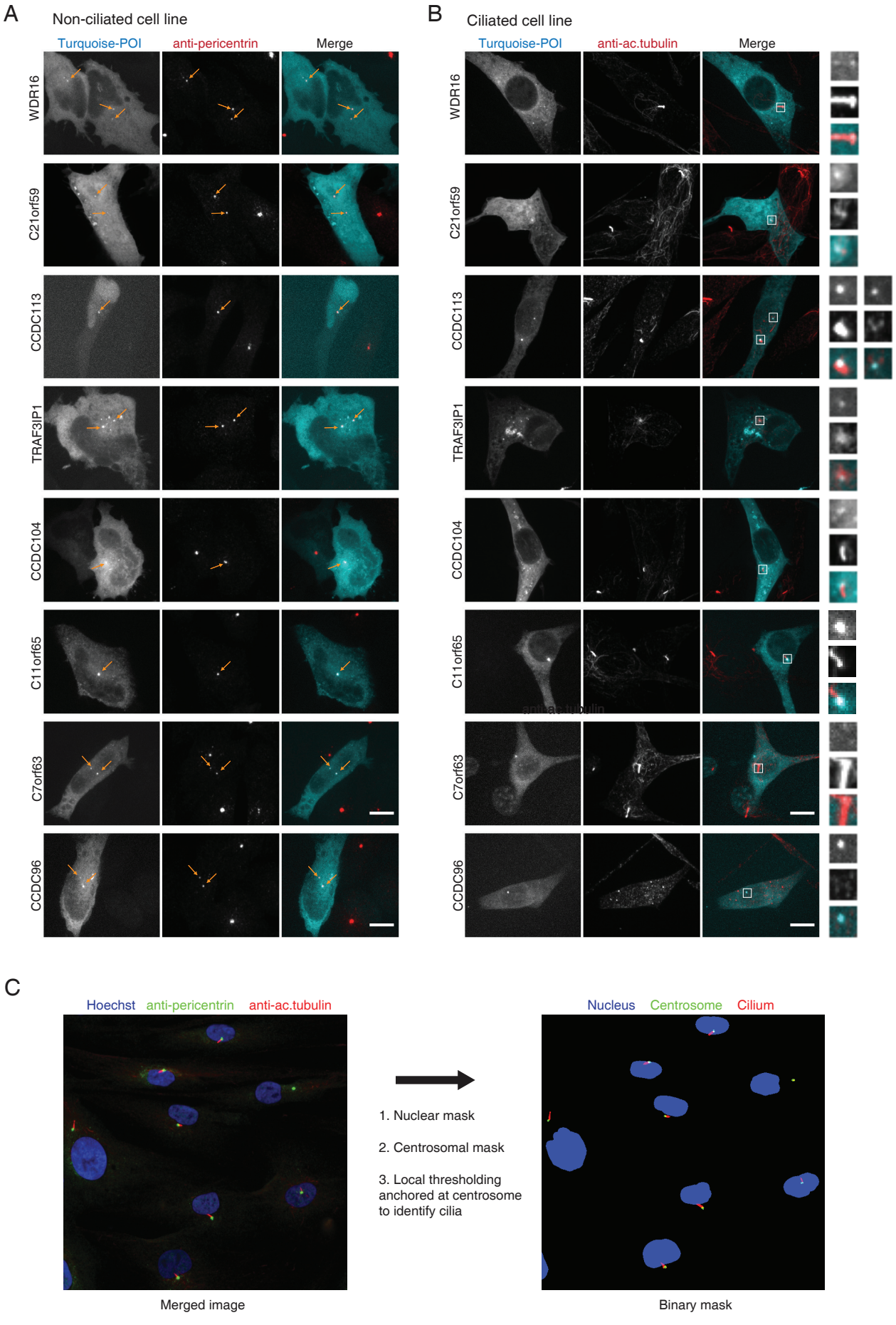

A
mCitrine-KIAA1033
EEA1
Merge
mTurquoise-COMMD1
mTurquoise-COMMD4
B
mCitrine-KIAA1033
p62
Merge
mTurquoise-COMMD1
mTurquoise-COMMD4
C
siControl
siWASH1
EEA1/Phalloidin at 63X

Figure S6


A
Non-ciliated cell line
Turquoise-POI
anti-pericentrin
Merge
B
Ciliated cell line
Turquoise-POI
anti-ac.tubulin
Merge
WDR16
C21orf59
CCDC113
TRAF3IP1
CCDC104
C11orf65
C7orf63
CCDC96
C
Hoechst
anti-pericentrin
anti-ac.tubulin
Merged image
1. Nuclear mask
2. Centrosomal mask
3. Local thresholding anchored at centrosome to identify cilia
Nucleus
Centrosome
Cilium
Binary mask

SUPPLEMENTAL FIGURE LEGENDS

Figure S1, Related to Figure 1. Species used for phylogenetic profiling.

177 eukaryotic species used for phylogenetic profiling, in the order used for the binary vectors throughout the study. Each species belongs to one of 13 branches determined by the NCBI taxonomy, colored in accordance with the bar at the top of the figure, and each branch was assigned a name and an abbreviated code used throughout the study. M-Mammals, V- other Vertebrates, LT- Lancelets/Tunicates, EH- Echinoderms/Hemichordates, A- Arthropods, N-Nematodes, C- Cnidaria, SP- Sponges/Placozoa, CF- Choanoflagellates, F- Fungi, A-Amoebozoa, P-Plantae, PR- Other protists.

Figure S2, Related to Figure 2. Orthogroup creation and analysis of gain/loss.

(A) The top BLAST hit against human RAB9B in each species colored according to a BBH match for the corresponding protein against RAB9B (black), RAB9A (green), RAB7A (violet), or any other human gene (red). Gray lines highlight inferred orthogroup thresholds. Below, resulting phylogenetic profiles (blue/white) and inferred orthogroup tree. **(B)** Number of human genes with orthologs in several commonly used model organisms. **(C)** Histogram of gene family size, estimated using the largest (most ancestral) orthogroup identified for each gene in the genome, with representative orthogroup trees to illustrate the contents of the first three bins. **(D)** Distribution of gene first appearances sorted by the branch with the first detectable ortholog for each gene (orthogroup size=1, top plot) and each orthogroup corresponding to a gene family or part of a gene family (orthogroup size >=2, middle plot). The lower plot shows the estimated average loss frequency in each branch, calculated using all genes with first appearance in that branch or older. Here the branches are used as a proxy for increasing age (gray arrow) based on their estimated divergence from the human trunk (NCBI). V= Other Vertebrates, LT= Lancelets/Tunicates, EH= Echinoderms/Hemichordates, SP= Sponges/Placozoa, CF= Choanoflagellates, AB= Amoebozoa. **(E)** Myosin gene family tree reconstructed using hierarchical orthogroup relationships generated by the algorithm and using the gene MYH1 as an initial seed. Branch lengths are arbitrary with each branch set to a single unit. **(F)** Gene family tree for small Rho GTPases using the gene RAC1 as a starting point. Branch lengths are arbitrary with each branch set to a single unit.

Figure S3, Related to Figure 3. Benchmarking using STRING and curated PubMed citations.

(A) Number of known functionally interacting pairs (STRING) as a function of decreasing PCS, for all orthogroups (<= 4 members, black), or restricted to orthogroups of size=1 (red). **(B)** Number of known functionally interacting pairs (STRING) as a function of decreasing PCS, for orthogroups (<=4 members) present in invertebrates (black) and orthogroups (<=4 members) present only in vertebrate species (red). **(C)** Histogram of curated PubMed citation counts for 19973 human genes (NCBI, see Supplemental Experimental Procedures for details).

Figure S4, Related to Figures 4 and 5. In-depth analysis of splicing genes, PCA of modules and analysis of expanded modules.

(A) Orthogroups belonging to various splicing complexes, each labeled in a different color, with the most conserved 'core' snRNP complexes at the top. For simplicity graph contains only singleton orthogroups. Within each complex, orthogroups are ordered by age with youngest first. **(B)** Schematic illustrating the composition of the spliceosome in humans, *S. cerevisiae*, and the inferred composition in an early eukaryote. The core snRNPs are denoted with circles and additional complexes with squares, color-coded as in (A). Ghosted boxes or squares denote complexes that have suffered significant losses (or have not expanded yet) **(C)** Species ordered according to the number of genes shared with an inferred ancestral eukaryote, separately for fungal species (gray) and mammals (black). Arrows mark the positions of *S. cerevisiae*, *S. pombe*, and *H. sapiens* in this plot. **(D)** Percent of variance explained by the first ten principal components obtained from a PCA of 334 consensus profiles, as well as the cumulative percentage (gray line and axis). **(E)** Example of an expanded hOP-module containing members of the germline-specific CATSPER ion channel complex using a relaxed threshold (Supplemental Experimental Procedures). Red stars mark uncharacterized genes. **(F)** Schematic illustrating the CATSPER protein complex with auxillary subunits and highlighting the potential link to the EF-hand containing candidate EFCAB9.

Figure S5, Related to Figure 6. Additional co-localization and siRNA experiments.

(A), (B) mTurquoise-COMMD1 and mTurquoise-COMMD4 (blue in merge) were co-expressed with the WASH complex marker KIAA1033 (green in merge) in HeLa cells and colocalized with a marker for early endosomes (EEA1, (A)) or for autophagosomes/aggresomes (p62, (B)). Yellow arrows highlight large puncta that appear to be EEA1-negative and p62-positive. Scale bars= 10 μm. **(C)** Depletion of WASH1 causes endosomal collapse (EEA1, red) to an area near the center of the cell visible at high magnification. Images are single slices of a confocal stack acquired at 63X. Scale bar= 10 μm.

Figure S6, Related to Figure 7. Additional co-localization experiments and image segmentation.

(A) Colocalization of mTurquoise-tagged (cyan in merge) candidate proteins with antibody to pericentrin (red in merge) in HeLa cells, fixed 18 hours after transfection. Yellow arrows mark the position of centrosomes. Scale bar= 10 μm. **(B)** Colocalization of mTurquoise-tagged (cyan in merge) candidate proteins with antibody to acetylated tubulin (red in merge) in NIH3T3 cells, fixed 18 hours after transfection. White boxes (zoom on right) highlight the cilia and surrounding area. Scale bar= 10 μm. **(C)** Images of nuclei (Hoechst, blue), centrosomes (antibody to pericentrin, green) and cilia (acetylated tubulin) were processed with a custom-written MATLAB routine (see Materials and Methods for details). Briefly, a nuclear mask was created through semi-local background subtraction, and centrosomal foci identified using a puncta identification routine. Each identified centrosome was linked to the closest nucleus and then utilized to partition the image (watershed) and subtract background locally in the acetylated tubulin channel, a required step because of the high cell-to-cell variance in acetylated tubulin levels.

Objects were then identified in the tubulin channel and discarded if they could not be associated with a centrosome. Masked cilia were thereby linked to individual cells and used to quantify various parameters (size, shape, intensity) on a per-cell basis. Also see Supplemental Experimental Procedures for details.

Table S4

Gene	Suggestive evidence	Chlamydomonas FBB[a]	HPA[b]	Specific prediction
ATAT1	PMID: 23748901	-	Basal body	Cilia biogenesis/stability
CCDC176	-	XP_001701007	Cytoplasm	Basal body
C21orf59	PMID: 24094744	XP_001699200	Cilia	Dynein arm assembly
CCDC37	PMID: 21289087, PMID: 23569216	XP_001699777	Basal body	Cilia motility
C7orf63	-	XP_001703508	Basal body	Basal body
WDR16	PMID: 17394468	XP_001690930	Cilia	-
DPY30	PMID: 22851692	-	Cilia	Radial spoke complexes
C11orf65	-	-	Basal body	Basal body
CCDC104	PMID: 22085962	XP_001694490	Cilia	Transport to cilium/basal body
CCDC96		XP_001697427	Cilia	Basal body
CCDC113		XP_001703742	Basal body	Basal body
C20orf26	PMID: 17967944	XP_001703513	Cilia	Dynein activity

a: FBB: Flagellum/ basal body proteome; **b**: Human Protein Atlas, staining fallopian tube glandular cells

SUPPLEMENTAL TABLE LEGENDS

Table S1, Related to Figure 3. Functional enrichment for top pairs.

Excel file containing functional enrichment calculations for the top co-evolving gene pairs (Experimental Procedures, Figure 3) using Reactome Pathways. P-values were estimated using the hypergeometric test and converted to False Discovery Rates (FDR) to account for multiple hypothesis testing (in MATLAB, See Supplemental Experimental Procedures for additional details).

Table S2, Related to Figures 4 and 5. Complete hOP-module details.

Details for all 1074 hOP-modules, including seed pair proximity scores, component orthogroups (Sheet 1), enriched Reactome Pathways (Sheet 2) and enriched protein complexes (COMPLEAT, Sheet 3). P-values were estimated using the hypergeometric test and reported in terms of the False Discovery Rate to account for multiple hypothesis testing (See Supplemental Experimental Procedures for additional details). Given the small size of most modules, enrichment was reported at a relatively high FDR threshold of 0.1 in the interests of a thorough characterization of hOP-module function.

Table S3, Related to Figure 7. Ciliome 'super-module'.

Orthogroups predicted to have cilia/basal body function, listing constituent genes and compared to 2 independent annotation resources and a recent proteomic characterization of multi-ciliated cells (Supplemental Experimental Procedures).

Table S4, Related to Figure 7. Selection of cilia/basal body candidates for experimental investigation.

Genes shortlisted for testing were first investigated in detail using two sources of existing information in the literature. First, a detailed PubMed search was carried out to identify papers that might have been missed by existing databases or that are more recent that provide evidence for CBB function in some other system ("suggestive evidence" column); second, we asked if the *Chlamydomonas* ortholog (if it exists) is annotated to have CBB function ("*Chlamydomonas* FBB" column); third, we assessed if tissue staining data from the Human Protein Atlas was suggestive of CBB localization in human tissues "HPA" column. Based on these various sources of information, we made initial specific predictions for each candidate gene (final column).

Table S5, Related to Experimental Procedures. siRNA and cDNA construct details.

Details of commercial siRNAs used in this study, including sequences and catalog numbers (Sheet 1), and sources for cDNA constructs used (primarily human ORFeome collection). See Supplemental Experimental Procedures for additional details.

File S1, Related to Figures 2 and 3. Raw data associated with study (compressed).

This file contains tab-delimited text files with all the data required to reproduce the bulk of the analyses carried out in this study. This is intended for users interested in further analysis of the data, as a complement to the online web server. There are 5 .txt files containing keys to the gene identifiers, orthogroup identifiers and species identifiers (Files 1-3), the full binary hOPMAP (31406*177, File 4), and all pair-wise phylogenetic co-occurrence scores (PCS, file 5). The zipped folder contains a README file with column labels and further details. See Figures 2 and 3, Experimental Procedures and Supplemental Experimental Procedures for additional details on the calculations used to generate these files.

SUPPLEMENTAL EXPERIMENTAL PROCEDURES

Brief glossary of terms

Orthology and paralogy: Orthology and paralogy were originally defined to separate two classes of observed homology based on descent- orthologs (vertical) are genes originating from a single common ancestor and paralogs are genes (parallel) originating from a duplication. However with the advent of genomics and the ensuing comparisons across a large number species assigning orthology and paralogy relationships became a lot more complicated. In one example, lineage-specific duplications will give rise to paralogs but also orthologs after a speciation event in that lineage (co-orthologs in the new genome). As a result, orthologous groups or orthogroups, referring to the full set of homologous genes that evolved from a single ancestral gene, may contain orthologs, co-orthologs as well as paralogs depending on the reference point. To simplify this confusing situation we do not address paralogs specifically in our study but instead restrict ourselves purely to the analysis of orthogroups and the above-mentioned broad definition of orthology. (Koonin, 2005)

Best Bidirectional Hit: Best bidirectional hit or BBH refers to a commonly used strategy to identify orthologs. If a gene A in genome X returns a top BLAST hit B in genome Y, and reciprocally gene B returns gene A as its top BLAST hit from genome X, A and B are defined to be orthologs by a BBH criterion.

Loss and gain events: We use the terms ‘loss’, ‘gain’ and ‘appearance’ throughout our study. It is important to point out that, since we do not actually build specific models of evolution or reconstruct full gene trees, losses and gains are inferred indirectly from the phylogenetic profiles using a parsimony principle. In other words, if older-branching species contain a copy of a gene, we consider gaps in the phylogenetic profile to represent ancestral loss events in the relevant lineages.

Branch naming: The names/abbreviations for individual branches were determined by an examination of the current taxonomy literature (Burki, 2014) and selected to be reasonably informative for the lay reader, especially because kingdom and phyla designations have changed frequently over the last few years with the advent of systematic genomics (and are expected to change further).

Co-evolution: Co-evolution in this study, unless specifically mentioned, refers to the specific phenomenon of co-evolution through shared gene gains and losses, and is used interchangeably with the term ‘co-occurrence’. Co-evolution between genes or gene families can occur in many ways in addition to gene loss, often in a sequence or structure-specific manner- those modes of co-evolution are beyond the scope of this study.

Generation of hOP-profiles

Identifying homologs of human protein coding genes. We investigated 19973 protein coding human genes that yielded homologs in a search of the NCBI online BLAST database (NCBI, September 2013; using BLASTp bit scores). We used the longest annotated human protein for our query and retrieved all significant homologs across 177 eukaryotic species (using the online query option and a custom made MATLAB program). These 177 species were selected based on completion or near completion of their genomic sequencing and annotation as well as on the availability of their proteomes in the NCBI BLAST database. The automated query of 19973 human proteins retrieved ~80 million homology scores for ~4 million unique proteins in the 177 species. Our subsequent analysis made use of a four-column matrix of these 80 million matches (M4 matrix), with the human NCBI gene ID corresponding to the queried protein, the RefSeq protein ID of the match, the species corresponding to the protein match and the raw BLAST homology score.

Orthogroup merging and orthogroup thresholds. Orthogroups of homologous human genes were generated using the M4 matrix derived above. Our identification of orthogroups only considered those inferred gene duplications that led to additional human genes (human-centric orthogroups). Since it is hard to predict a priori which subset of human genes from a gene family should be combined into a relevant orthogroup for phylogenetic profile comparisons, all possible orthogroups need to be tested for potential phylogenetic matches. The key goal of the analysis was to identify an optimal threshold to establish whether a homolog in a particular species belongs to a particular orthogroup generated by sequential merging of human genes with shared ancestry. To pursue this strategy, we first determined the BLAST scores of all proteins in any species against each human protein (query) that are a better match (BBH) to proteins encoded by other human genes (target). This generates effective 'link' values between all query-target pairs of homologous human genes that we represented by an asymmetric sparse matrix (BLAST distance value as a function of the query gene and the connected gene). We termed the maximal of these link values an "orthogroup threshold" that was derived for each gene and orthogroup (we noted occasional outlier BLAST scores resulting from genes annotated to the wrong species and selected the third highest instead of the highest score to eliminate these). A procedure was implemented to iteratively merge the two human genes or orthogroups that are linked by this orthogroup threshold, following which the link values were pooled for the new orthogroup. The iterative procedure was halted when the BLAST values dropped below 50 (the threshold below which we observed frequent merging between orthogroups with clearly different functions- based on inspection of known gene families like GTPases or kinases- often due to shared functional domains) or the orthogroup size reached 100. This size cutoff was selected given the limited practical value of making specific functional predictions for groups of genes of that size or larger. This procedure resulted in 31406 orthogroups and orthogroup thresholds.

Species and orthogroup correction terms to derive hOP-profiles. We next generated a BBH (best bidirectional hit) homology matrix of the highest BLAST scores for the 19973 human genes against the 177 species. For each of the orthogroups, we used the maximal homology values in each species to sequentially generate the 31406 x 177 orthogroup homology matrix and applied the orthogroup thresholds to gain initial binary profiles. We next considered that species branching off from the same sites with respect to the human lineage can significantly

diverge from each other, meaning that a neighboring species can exhibit a relatively lower homology score despite having clearly matching orthologs. To calculate correction terms, we investigated those species branching off at the same site (18 branch points have multiple species in the human-centered tree). For species branching from each of these points, we calculated a relative average BLAST score divergence by selecting orthogroups with corresponding gene orthologs in neighboring species. We used the slowest diverging species with the highest average BLAST score at each branch site as a reference (this reference species was not corrected). For each species s, the average deviation (D_s) from the reference species was multiplied by an additional factor (F_s) to correct for the sequence evolution rate of each individual orthogroup. The factor F was calculated by comparing the divergence rate of a particular orthogroup compared to the average divergence rate of all orthogroups. Specifically, the relative divergence was calculated using a linear fit to the BLAST scores from each orthogroup using the averaged BLAST score values for all orthogroups as the x-axis. For orthogroups with homologs in more than 24 species, the first and second half of the species in the list were corrected by separate linear fits to account for a frequently observed shift in sequence evolution rates. The BLAST scores for each species were then adjusted by adding the combined correction term (D_s*F_s). Empirical analysis showed that the correction terms elevated low BLAST scores of clearly conserved homologs close to those of the neighboring species. We applied the orthogroup thresholds to each of the corrected homology matrices to generate the 31406 binary hOP-profiles.

Orthogroup naming conventions. Throughout the study, orthogroups were designated by a 2-part naming convention (a-b), with the first part being the gene in the orthogroup with the most similarity to the inferred ancestor (a) and the second the number of genes in the orthogroup (b). For example, the orthogroup containing TTC21A and TTC21B would be named TTC21B-2 (Figure 2).

Relationship to other orthology inference methods. There is an extensive literature on orthology inference, widely considered one of the central challenges in comparative genomics. Orthology methods can be broadly divided into graph-based and tree-based methods. Graph-based methods (COG, eggNOG, OrthoMCL)(Li et al., 2003; Powell et al., 2014; Tatusov et al., 1997) use similarity relationships (most often BLAST) in a pair-wise or multi-species comparison to identify best BLAST matches, which are in turn used to build orthogroups. Tree-based methods (TreeFam, PhylomeDB, ENSEMBL Compara)(Huerta-Cepas et al., 2014; Schreiber et al., 2014; Vilella et al., 2009) usually contain a tree construction step (based on a multiple sequence alignment) followed by a tree reconciliation step (to superimpose the gene tree from step 1 on the consensus species tree). While arguably the most precise representation of orthologous relationships, tree-based methods are computationally intensive, especially as the number of species (and complexity of the resulting species tree) increases, and susceptible to noise (especially at the multiple sequence alignment step). On the other hand, graph-based methods (especially pair-wise) are more error-prone at larger evolutionary distances but more effective when large numbers of genes and species are involved (Trachana et al., 2011). Overall, however, studies have shown that many orthology inference methods produce comparable results (Trachana et al., 2011), with even the simple reverse-best BLAST sometimes outperforming far more sophisticated algorithms (Kristensen et al., 2011). These

findings and our unique requirements (genome-scale hierarchical orthology relationships defined specifically by human genes without taking into account gains/losses in other branches, no constraints on final orthogroup size, and an unprecedented number of species being profiled) led us to develop a graph-based approach that would be appropriate for this dataset. Conceptually, our method is very similar to other graph-based methods that use a BBH approach, with the additional criterion that each BBH is resolved across all BLAST scores from all the species being profiled.

Generation of the hOP-matrix

Pairwise interaction data used for optimization. The full set of STRING protein-protein interactions for *Homo sapiens* were parsed (http://string-db.org/, version 9.1). Interactions below a combined score of 0.4 and those for which the primary evidence involved co-occurence across genomes (average of all other sub-scores< 0.4) were filtered out. This data was then mapped onto the full list of orthogroups in binary form- a pair of orthogroups were said to interact if any of their member genes were found to interact. Orthogroups sharing ancestry were not given an interaction score. This allowed us to calculate a fractional rate at which predicted pairwise hOP-profile homologs match with known protein-protein interactions. A bootstrap with the identity of high-scoring pairs scrambled (100 randomized runs) provided a measure of the background. Overall, this strategy allows one to directly compare the effect of different optimization parameters in the metrics on the fraction of returned known interactors. We used this strategy to optimize the scoring metric for the hOP-matrix (see below).

Proximity score between pairs of hOP-profiles. The metric used to generate a genome-wide pairwise phylogenetic proximity (PPS) score between pairs of hOP-profiles was based on the linear representation of a eukaryotic species tree and by marking transitions form 1 to 0 and from 0 to 1 as loss events in individual species or neighboring species. We added an extra weight factor for transitions supported by two species on each side, e.g. 0011 or 1100 that have higher confidence (w). We also subtracted a penalty factor (p) when a transition is present for one profile but not for the other. For every pair, the phylogenetic proximity score was calculated as the linear sum of these weighted transitions and organized into a 31406*31406 sparse matrix.

In pseudo-code, given binary transition vectors T1 (0→1), T2 (1→0), T3 (00→11) and T4 (11→00) and a pair of phylogenetic profiles **i** and **j**:

$$PPS= [\ sum(T1_i == T1_j) + sum(T2_i == T2_j) + w^* (sum(T3_i == T3_j) + sum(T4_i == T4_j))\ ] - p^*[\ sum(T1_i \sim= T1_j) + sum(T2_i \sim= T2_j) + w^* (sum(T3_i \sim= T3_j) + sum(T4_i \sim= T4_j))\ ]$$

(In this pseudo-code formulation sum(X==Y) is 1 when X and Y match, sum (X~=Y) is 1 when X and Y do not match)

We tested combinations of different confidence and penalty factors to optimize the frequency at which a ranked list of predicted pairwise interactions based on phylogenetic profiles returned known protein-protein interactions (STRING). This analysis showed an improvement in the

return frequency when the confidence factor was set at 1.5. Higher penalty factors produced a tradeoff, reducing the false positive rate but decreasing the total number of STRING interactions identified at lower thresholds.

Exclusion of orthogroups for global analysis of pairs and modules

Two exclusion criteria were applied to the full set of 31406 orthogroups before carrying out any global analysis. First, all orthogroups present only in vertebrate species were excluded. This is because we found that they contributed very few useful predictions on account of having very few informative transitions (shared losses), as can be seen in Figure S3B. However, it should be noted that many genes represented in that set are also present in older orthogroups (Figure S2D) and were thus included in our analysis anyway. The second filter applied was to remove any orthogroups containing more than 4 genes. This was on account of the observation that the larger orthogroups tended to contain genes that had clearly diverged in function, making any pair-wise functional predictions less likely to be precise. These two filters resulted in a final set of 14412 orthogroups analyzed in Figures 3C-E, 4 and 5. Data for all 31406 orthogroups is available on our web server.

Generating hOP-modules

Seed pairs were simply selected as all pairs of orthogroups with a score >=threshold A. Agglomerative modules were created in a stepwise fashion starting with a seed pair of orthogroups. At each step, the top-scoring orthogroup from the weighted average of proximity scores between existing module components and the rest of the hOP-proximity matrix was included as long as it did not share any ancestry with any orthogroup already in the module. This stepwise process was repeated until the top-scoring orthogroup had a score of <threshold B. Instead of growing all modules simultaneously in a more typical centroid-based agglomerative clustering approach, this process was carried out one module at a time, starting with the strongest seed pair (highest PPS), and removing module components from the general pool before moving on to the next seed pair. This was done to give the strongest pairs (reflecting a more precise underlying model of co-evolution) the best chance of developing large modules. For our conservative analysis in this study, we set threshold A=5 and threshold B=5. With these parameters, we identified 1074 modules of which 740 remained as isolated pairs (Table S2).

Module expansion

We noted that in certain cases with weaker phylogenetic coupling and/or systematic sub-patterns, the stringent scoring criteria would exclude functionally related genes. In such cases we allowed each module component to draw in additional linked orthogroups at a threshold C, resulting in ‘expanded’ modules. For the expanded CATSPER module (Figure S4) and the ciliome analysis (Figure 7), we set this threshold C=3. Finally, to cast a wide net for additional

WASH module components (Figure 6), we used a hOP-proximity matrix with a mismatch penalty of 0.2 instead of 0.6 (available on web server).

Functional annotations and enrichment

Functional annotations were obtained from sources described in the Experimental Procedures. In general, orthogroups containing multiple genes were annotated with a certain term once if one or more genes in the orthogroup associated with that term. Hypergeometric probabilities and FDR were estimated in MATLAB with the help of the Bioinformatics Toolbox (release 2014a) functions *mafdr, hygepdf*. When required, gene annotations were converted between ENSembl and NCBI annotation using the NCBI ftp database (ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, last downloaded in July 2014). Linked PubMed citations were also obtained from the NCBI ftp database (downloaded July 2014). Papers linked to 20 or more human genes were excluded before specific citation counts were assessed.

Disease genes

A curated list of monogenic and polygenic disease genes was obtained from a recent study (Chen et al., 2013). A curated list of genes mutated in various human cancers was obtained from the Cancer Genome Census (Futreal et al., 2004) (http://cancer.sanger.ac.uk/cancergenome/projects/census/, downloaded July 2014). Due to the difficulty of assigning disease annotations to orthogroups containing more than one gene, we considered only singleton orthogroups at a PPS>=10 (1021 genes). We noted a strong enrichment of genes annotated to monogenic disorders (hypergeometric p-value <0.0001), and a smaller but significant depletion of genes involved in polygenic disorders (hypergeometric p-value=0.0075) and cancer (hypergeometric p-value=0.0013).

Selection of cilia/basal body candidates for experimental validation

We identified 206 candidate CBB (cilia/basal body) genes (Table S3) with approximately half overlapping with a gold standard ciliome (van Dam et al., 2013) (http://www.syscilia.org/goldstandard.shtml) and an independent database of cilia proteins (Inglis et al., 2006)(http://www.sfu.ca/~leroux/ciliome_database.htm). However inspection of the literature indicated that these data sources are not comprehensive enough to confirm that the remaining genes are completely uncharacterized. An example of this annotation issue is a comparison we carried out with a recent independent high-throughput proteomics dataset (Hoh et al., 2012) finding a fairly different overlapping set of genes (Table S3). As a result, we selected a shortlist (Table S4) based on an in-depth investigation of the literature and images from the Human Protein Atlas (Uhlen et al., 2010) (http://www.proteinatlas.org/).

siRNA reagents and expression constructs

siRNA reagents were obtained from Qiagen (pools of 3 individual siRNAs) and used at a final concentration of 25 nM. The catalog numbers and sequences of each siRNA are listed in Table S5. All constructs were generated using the Gateway cloning system (Invitrogen, https://www.lifetechnologies.com/us/en/home/life-science/cloning/gateway-cloning.html). Entry vectors encoding genes of interest were either generated by TOPO cloning from PCR products generated with HeLa or HUVEC cDNA libraries into the pENTR/D-TOPO vector (Invitrogen) or obtained directly from the human ORFeome collection (pDONR-223 entry vectors, ORFeome version 5.1, http://horfdb.dfci.harvard.edu/hv5/). Expression constructs were then generated by using LR-II recombination (Invitrogen) between these entry vectors and custom-designed destination vectors (pcDNA5/FRT/TO with a Gateway cassette and cDNA encoding mCitrine or mTurquoise inserted into the multiple cloning site). A full list of constructs used (and primers for PCR, if applicable) in this study can be found in Table S5.

siRNA and cDNA transfections

Glass-bottom plates were first coated overnight with 31ug/ml bovine collagen in PBS (Advanced BioMatrix, 5005-B). siRNA transfections in HeLa cells were carried out using Dharmafect I (GE/Dharmacon) 12-20 hours after cell plating using standard manufacturer protocols and a final siRNA concentration of 25 nM. The transfection mix was replaced with full growth medium 6 hours after transfection, and cells were assayed and/or fixed 48 hours later. For p62 puncta formation assays, cells were treated with 10 μM Bortezomib in full growth medium for 30 minutes before fixation. siRNA transfections in Hs68 cells were carried out with Lipofectamine RNAiMax (Life Technologies) using manufacturer-recommended ratios with reverse transfection. Cells were transferred to starvation medium (0.1% BSA in DMEM) 12 hours after transfection and fixed 48 hours later. cDNA transfections were carried out using Fugene HD (Promega) in HeLa cells and Lipofectamine 2000 (Life Technologies) in Hs68 cells. The transfection mix was replaced with full growth medium 6 hours after transfection and cells were assayed 18 hours later.

Immunofluorescence

Following fixation in 4% paraformaldehyde (in PBS, Ted Pella) for 20 minutes, cells were permeabilized using Triton-X (0.2% in PBS, Sigma) for 15 minutes and incubated in blocking buffer (3% BSA in PBS, Sigma-Aldrich) for 60 minutes. Primary antibody staining (antibodies used are listed in Experimental Procedures) was carried out at 4C overnight followed by incubation with the relevant Alexa Fluor-conjugated secondary antibodies for 1 hour at room temperature (Alexa-568, Alexa-647, Alexa-488, 1:1000, Life Technologies) and Hoechst 33342 (1:10000, Life Technologies).

Image analysis

Manual inspection of images was carried out using ImageJ (Fiji bundle). All image analysis was carried out using custom-written routines in MATLAB (Release 2014a). For the quantification of cilia and centrosomes, a nuclear mask was first created using the Hoechst signal. A second mask was created using the pericentrin channel and an iterative shrinking strategy (Gupta et al., 2014) to eliminate out-of-focus fluorescence, and pericentrin puncta above a minimum size threshold (3 pixels) were associated with the closest nucleus using a Euclidean distance metric. The nuclei with their associated centrosomes were used to partition the field of view using a watershed algorithm to carry out a local background subtraction of the acetylated tubulin image, a required step because of the high cell-to-cell variability in acetylated tubulin staining. Following local background subtraction, the tubulin image was used to create a third mask. In a final step, each discrete object within this mask that could not be associated with a centrosome was discarded. Different aspects of each centrosome (size, intensity, distance from nucleus) and each associated cilium (size, intensity, length) were then quantified and stored. For the analysis of p62 puncta, the Hoechst image was segmented as above. Puncta were identified in the p62 channel using a version of the same iterative shrinking strategy to reliably identify puncta of different intensities while eliminating out-of-focus fluorescence.

SUPPLEMENTAL REFERENCES

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. Cold Spring Harb. Perspect. Biol. *6*, a016147.

Chen, W.-H., Zhao, X.-M., van Noort, V., and Bork, P. (2013). Human monogenic disease genes have frequently functionally redundant paralogs. PLoS Comput. Biol. *9*, e1003073.

Van Dam, T.J., Wheway, G., Slaats, G.G., Huynen, M.A., and Giles, R.H. (2013). The SYSCILIA gold standard (SCGSv1) of known ciliary components and its applications within a systems biology consortium. Cilia *2*, 7.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Gupta, G.D., Dey, G., Swetha, M.G., Ramalingam, B., Shameer, K., Thottacherry, J.J., Kalappurakkal, J.M., Howes, M.T., Chandran, R., Das, A., et al. (2014). Population distribution analyses reveal a hierarchy of molecular players underlying parallel endocytic pathways. PLoS One *9*, e100554.

Hoh, R.A., Stowe, T.R., Turk, E., and Stearns, T. (2012). Transcriptional program of ciliated epithelial cells reveals new cilium and centrosome components and links to human disease. PLoS One *7*, e52166.

Huerta-Cepas, J., Capella-Gutiérrez, S., Pryszcz, L.P., Marcet-Houben, M., and Gabaldón, T. (2014). PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. Nucleic Acids Res. *42*, D897–D902.

Inglis, P.N., Boroevich, K.A., and Leroux, M.R. (2006). Piecing together a ciliome. Trends Genet. *22*, 491–500.

Koonin, E. V (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. *39*, 309–338.

Li, L., Stoeckert, C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. *13*, 2178–2189.

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldón, T., Rattei, T., Creevey, C., Kuhn, M., et al. (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. Nucleic Acids Res. *42*, D231–D239.

Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M., and Bateman, A. (2014). TreeFam v9: a new website, more species and orthology-on-the-fly. Nucleic Acids Res. *42*, D922–D925.

Tatusov, R.L., Koonin, E. V, and Lipman, D.J. (1997). A genomic perspective on protein families. Science *278*, 631–637.

Trachana, K., Larsson, T.A., Powell, S., Chen, W.-H., Doerks, T., Muller, J., and Bork, P. (2011). Orthology prediction methods: a quality assessment using curated protein families. Bioessays *33*, 769–780.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.