natureresearch
https://doi.org/10.1038/s41559-020-01320-z

Supplementary information

Timing the origin of eukaryotic cellular complexity with ancient duplications

Supplementary information for:

Timing the origin of eukaryotic cellular complexity with ancient duplications

Julian Vosseberg[1*], Jolien J. E. van Hooff[1*§], Marina Marcet-Houben[2,3,4], Anne van Vlimmeren[1†], Leny M. van Wijk[1], Toni Gabaldón[2,3,4,5], Berend Snel[1]

[1]Theoretical Biology and Bioinformatics, Department of Biology, Faculty of Science, Utrecht University, Utrecht, The Netherlands
[2]Centre for Genomic Regulation, The Barcelona Institute of Science and Technology, Barcelona, Spain
[3]Life Sciences Department, Barcelona Supercomputing Center, Barcelona, Spain
[4]Mechanisms of Disease, Institute for Research in Biomedicine, Barcelona, Spain
[5]Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain

*These authors contributed equally to this work
§Current affiliation: Ecologie Systématique Evolution, CNRS, Université Paris-Saclay, AgroParisTech, Orsay, France
†Current affiliation: Department of Biological Sciences, Columbia University, New York City, United States of America

Correspondence to: toni.gabaldon@bsc.es (T.G.) or b.snel@uu.nl (B.S.)

Table of contents

Supplementary Methods

KOG-to-COG clusters analysis

Selecting sequences and generating clusters

In order to compare our phylogenomics approach to previously reported accounts of duplications during eukaryogenesis, we applied it to the clusters of homologous sequences established by Makarova *et al.*[11]. Briefly, they mapped eukaryotic orthologous groups (KOGs) to homologous prokaryotic orthologous groups (COGs). In many cases, multiple KOGs mapped to a single COG, which often reflects a duplication during eukaryogenesis. Furthermore, KOGs had been clustered together if they are homologous to each other but lack a homologous COG. We used these KOG-to-COG clusters to assess if we, using a phylogenomics approach, were able to recapture the prevalence of gene duplications during eukaryogenesis that Makarova *et al.* observed by calculating ratios of KOGs to their affiliated COGs. Moreover, we took advantage of the current wealth of sequenced biodiversity by using an alternative, more representative species and sequence dataset compared to the original study. The results of this KOG-to-COG analysis can be found in Supplementary Table 1.

To recreate the KOG-to-COG clusters we used the COG assignment of the non-Asgard archaeal prokaryotic sequences provided by eggNOG and performed sequence profile searches with the Asgard archaeal and eukaryotic sequences. For the Asgard archaea, we downloaded profile HMMs of all COGs from eggNOG 4.5[35] and assigned the Asgard protein sequences to COGs using hmmscan (HMMER v3.1b1[36]). For eukaryotes, we selected ten species to obtain a good representation of eukaryotic diversity: *Naegleria gruberi* and *Euglena gracilis* (Excavata), *Cladospihon okarmurans* and *Bigelowiella natans* (SAR+Haptista), *Guillardia theta* and *Klebsormidium flaccidum* (Archaeplastida+Cryptista), *Acanthamoeboa castellanii* and *Acytostelium subglobosum* (Amoebozoa), and *Capsaspora owczarzaki* and *Nuclearia* sp. (Obazoa). We specifically opted for these species, because they were often involved in BBHs in the Pfam sequence selection (see Methods, 'Reduction of sequences'). Subsequently, we downloaded profile HMMs for orthologue clusters at the level of eukaryotes from eggNOG 4.5[35]. These contained both the supervised KOGs and non-supervised orthologous groups (ENOGs). The original KOG-to-COG clusters from Makarova *et al.*[11] did not include these ENOGs, but instead included candidate orthologous groups (TWOGs). Because these TWOGs are now obsolete, we sought to find the best matching ENOG based on the original sequence members of each TWOG. We combined the profile HMMs of these ENOGs with those of the KOGs and created a profile database. We performed hmmscan to assign protein sequences from the eukaryotic species to these KOGs/ENOGs.

Subsequently, for all KOGs/ENOGs and COGs, we reduced the number of sequences with kClust v1.0[37], using a score per column of 3.53 (approximately 70% sequence identity). We subsequently merged homologous sequences from eukaryotes, prokaryotes and Asgard archaea according to the KOG-to-COG mapping, resulting in updated KOG-to-COG clusters comprising sequences from diverse and informative eukaryotic and prokaryotic clades.

Phylogenetic analyses

For each KOG-to-COG cluster, we generated phylogenetic trees using an in-house pipeline also used previously[10]. The sequences were aligned using MAFFT v6.861b[53], option –auto, and subsequently trimmed using trimAl v1.4[44] with a gap threshold of 0.1. From these alignments, we constructed phylogenetic trees using FastTree v2.1.8[48] with WAG as evolutionary model.

Tree analyses

For the annotation of nodes in KOG-to-COG trees a similar approach as for the Pfam-ScrollSaw trees was followed. Only the criteria for LECA and duplication nodes were slightly different. Because of the lower number of eukaryotic species we here simply annotated a node as a LECA node if it contained both Opimoda and Diphoda sequences, and instead of a consistency score, we used a species overlap criterion of two to annotate duplication nodes: if the daughters both fulfilled the LECA criterion and shared at least two out of the in total ten eukaryotic species, it was annotated as a duplication node.

Human phylome analysis

To validate the use of branch lengths to time gene duplications, we also applied this approach to the numerous duplications in chordates. We inferred these from the human phylome, which we downloaded from PhylomeDB[54] (Phylome ID 76: http://phylomedb.org/phylome_76). The results of this validation can be found in Extended Data Fig. 5f-h.

In this collection of phylogenetic trees we calculated the normalised vertebrate stem lengths by dividing the branch length between the common ancestors of chordates and vertebrates by the median branch length between the latter and present-day vertebrates. In case of duplications the stem length was included if the human seed protein was in the shortest possible stem length.

To obtain duplication lengths for duplications that occurred at different phylogenetic time points, we scanned in each tree the lineage of the human seed protein between the common ancestors of bilaterians and primates for the presence of duplications. Nodes connecting the seed with a human paralogue were annotated as duplication nodes. The phylogenetic time point ('age') of the duplication was obtained using the common ancestor of all species involved in the duplication event. Duplication lengths were calculated by dividing the branch length between the duplication node and the common ancestor of primates by the median branch length between the latter and present-day primates.

KOG functional categories were assigned to each protein in the phylome using emapper-2.0.1[51] based on eggNOG orthology data[55]. Functional annotation of the nodes in the trees were performed as described for duplication nodes before (see 'Functional annotation'). For each pair of duplications it was checked if they performed the same function and had the same age, performed the same function but had a different age or performed a different function but had the same age. For these pairs the difference in log-transformed duplication lengths was calculated.

Supplementary Discussion

Data sets used

We tested two different data sets. The KOG-to-COG gene family clusters[11] are a set specifically constructed to study duplications during eukaryogenesis and were therefore an ideal starting point. To get an even more complete picture of duplications we decided to use the Pfam database. By using this database we circumvented the need to use orthologous groups or infer homology. For certain families the Pfam domains correspond to full-length genes, whereas for others it is only a domain or even a motif. Although certain domain duplications are not fully independent of each other due to their presence in a single gene upon duplication, it is not unlikely that truly separated genes co-duplicated as well. Ideally, one would want to define the unit, either a domain or full-length gene, that evolved as an individual entity during eukaryogenesis. However, for various domains/genes it would be simply impossible to identify such a single entity, for example for domains that were independent upon acquisition or invention, but fused during eukaryogenesis and were therefore interdependent in LECA. This is especially probable given the abundance of gene fusion events during eukaryogenesis[56].

Sister group identity

7% of the acquisitions had an unclear prokaryotic ancestry. Both bacteria and archaea were present in the sister group with no phylum comprising a majority. A tentative explanation is that the identity of the donor is obscured due to post-acquisition HGT among distantly related prokaryotes. The tendency of these acquisitions to duplicate was similar to the Pfams with an archaeal ancestry (Fig. 2). This suggests that a large fraction of this group reflect genes present in the host lineage. Furthermore, a relatively large fraction of these acquisitions had another eukaryotic clade with LECA families in their sister group (34%, between 3 and 10% for the other groups), indicating that some of these acquisitions are placed in an incorrect, deep phylogenetic position. The stem and duplications lengths of these families with an unclear prokaryotic ancestry, however, were similar to those from families acquired from bacteria. Further research into these families is needed to elucidate their phylogenetic origin.

Branch lengths analysis

The stem lengths of acquisitions that happened simultaneously should approximate the same value, enabling us to assess the effect of duplications on branch lengths. Assuming the deep mitochondrial origin outside the alphaproteobacteria[8], all acquisitions with alphaproteobacteria as sister group should correspond to the same event, namely the divergence of the pre-mitochondrial and alphaproteobacterial lineages. We observed a difference in stem lengths between duplicated and non-duplicated families from alphaproteobacterial origin, with duplicated families corresponding to longer stems (Extended Data Fig. 5a). Even using the shortest branch as stem, which we chose in case of duplications, could not fully account for the difference in stem lengths in these few duplicated families. In contrast, no difference in stem lengths with duplications was seen for acquisitions with an Asgard archaeal sister group (Extended Data Fig. 5b). We also looked at the effect of duplications on the stem lengths for the numerous duplications that occurred in the vertebrate stem. For these more recent duplications we observed a longer vertebrate stem in case of duplications (Extended Data Fig. 5f), in line with the alphaproteobacterial-related duplications. The presence of duplications can result in a subtle yet significant accelerated evolutionary rate in both daughter lineages.

Because we had detected more duplicated families with an Asgard archaeal sister group than an alphaproteobacterial one, we looked more in depth into the first. We could not detect a clear pattern of acceleration after duplications in both daughter lineages for different functional groups (Extended Data Fig. 5c-d). The barely significant difference for duplications related to cellular processes and signalling was dependent on the presence of outliers. Duplications that resulted in the transition from a homomer to a heteromer could have had a different effect on evolutionary rate as the selection pressures on the protein interface has changed. We did not observe a difference between duplications in families that underwent such a transition and other families (Extended Data Fig. 5e). However, the number of the first group was low and involved all duplications in these families, not only those resulting in the homomer-heteromer transition. Further research into these different effects of duplications is warranted. In conclusion, we could not confidently distinguish differences in rates for different groups of proteins upon duplication that could bias our results.

The inferred timing of acquisitions represent the *earliest* possibility of the actual acquisition, because they are the result of taxon sampling (i.e. which of the present-day organisms have been discovered, sequenced and/or included in the analysis) and historical contingency (i.e. which lineages have not gone extinct). Duplication nodes, on the other hand, represent the *latest* possibility of the actual acquisition, and therefore they could be used to attenuate the inferred acquisition time point.

Comparison with Tria **et al.**[20]

Our conclusions are in stark contrast with a recent preprint[20], which reported remarkably fewer gene duplications and relatively many duplications in bacterial-related genes (compared to archaeal-related genes), which they interpret as being derived from the proto-mitochondrion. Based on their findings, the authors concluded that gene duplications support a eukaryogenesis model in which mitochondria entered early in eukaryogenesis, into a relatively simple, prokaryote-like host. We think this conclusion is insufficiently supported by their approach and resulting observations, because these have some clear deficits.

First and foremost, they infer very few eukaryogenesis duplications: 713 compared to 4,564 in our main dataset (see Supplementary Table 1). As an illustration: they did not recover well-documented greatly expanded protein families such as protein kinases and small GTPases[12,14], which we were able to recover (see Supplementary Table 2). The family that according to this preprint was most duplicated during eukaryogenesis was the dynein light chain family with 12 duplications.

Second, because they only inferred gene trees for eukaryotic sequences, they could not distinguish between duplications that happened during eukaryogenesis, those that happened before and pseudoparalogues (e.g., cytosolic and mitochondrial ribosomal proteins). Moreover, their limited usage of gene phylogenies also prohibits them from specifying the potential identity of the prokaryotic donor lineage.

Third, they do not discriminate between genes with alphaproteobacterial and another bacterial origin, but instead label all eukaryotic genes with bacterial affiliations as coming from the mitochondrial endosymbiont. Some, if not most, of these genes might in fact have been acquired through HGT from other bacterial lineages. Potentially, mixing these contributes to the relatively high number of gene duplications that count for endosymbiont-derived genes.

Fourth, they did not include the Asgard archaea in their analysis, which are crucial for any inference about eukaryogenesis. This might explain why the duplications in the cytoskeletal and ubiquitin systems were not correctly identified as duplications associated to archaeal acquisitions[5,6] in their analysis. This may have led to an underestimation of the duplications in host-related genes.

Supplementary References

53. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).

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

55. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

56. Méheust, R. *et al.* Formation of chimeric genes with essential functions at the origin of eukaryotes. *BMC Biol.* **16**, 30 (2018).

Supplementary Tables

Supplementary Table 1. Comparison of different datasets.

	Pfam-ScrollSaw trees	Trees from recreated KOG-to-COG clusters	Original KOG-to-COG clusters (no trees)[11]
Acquisitions	4,335	3,460	1,092
Inventions	1,334	883	1,058
Duplications	4,564	4,888	1,987
LECA families	10,233	9,231	4,137
Multiplication factor	1.81	2.12	1.92

Supplementary Table 2. Most expanded acquisitions or inventions during eukaryogenesis.

Pfam	Ancestry	Number of LECA families
Total		
Mitochondrial carrier*	Invention	123
Protein kinase	Planctomycetes	106
RING-finger/U-box	Actinobacteria	92
PH domain	*Haloplasma*	82
Ubiquitin	Asgard archaea	76
C2 domain	Prokaryotes	72
RNA recognition motif	Aβγ-proteobacteria	71
Tetratricopeptide repeat	Firmicutes	66
POZ domain	Chlamydiae	50
FYVE/PHD zinc finger	Invention	46
Asgard archaea		
Ubiquitin	Asgard archaea	76
Vps51 domain superfamily	Asgard archaea	19
Cyclin	Asgard archaea	19
Helix-turn-helix	Asgard archaea	16
Thioredoxin	Asgard archaea	15
Helix-turn-helix	Asgard archaea	11
Golgi-transport	Asgard archaea	10
Helix-turn-helix	Asgard archaea	10
Gelsolin repeat	Asgard archaea	10
Gelsolin repeat	Asgard archaea	10
Alphaproteobacteria		
Sterile alpha motif	Alphaproteobacteria	10
Galactosyltransferase	Alphaproteobacteria	9
EF-hand 8	Alphaproteobacteria	8
Iron/zinc purple acid phosphatase-like protein C	Alphaproteobacteria	5
DDE superfamily endonuclease	Alphaproteobacteria	5
ABC transporter	Alphaproteobacteria	5
Alpha/beta hydrolase fold	Alphaproteobacteria	5
Ferric reductase	Alphaproteobacteria	4

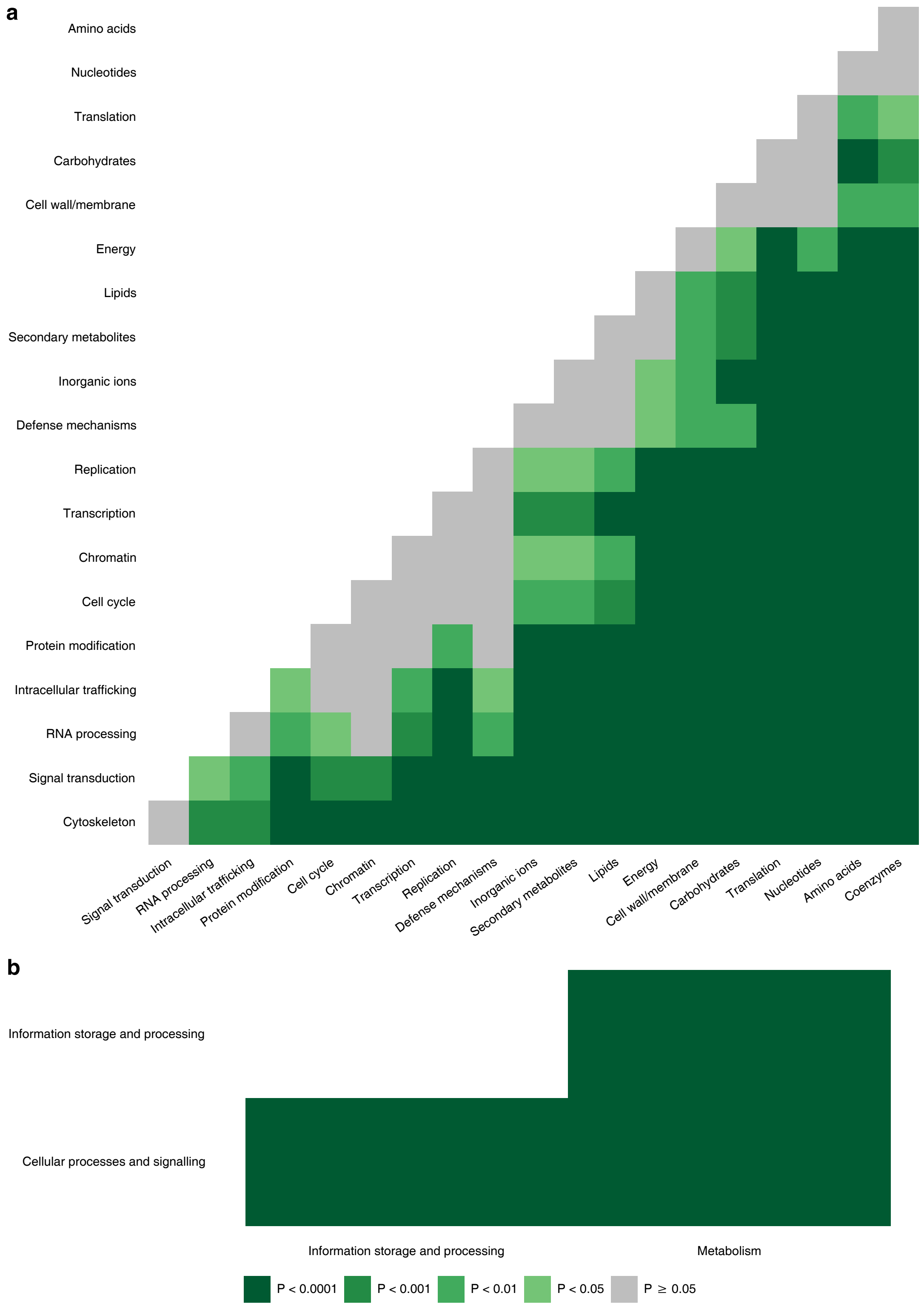
*A mitochondrial carrier protein typically contains three of these domains.

Supplementary Table 3. Effect of different duplication consistency and LECA coverage thresholds.

Duplication consistency score	LECA coverage score	Number of LECA families	Number of unclassified nodes	Number of eukaryotic clades without LECA families	Fraction well-supported* LECA nodes	Fraction well-supported* duplication nodes
0	**0**	23,567	5,304	19,661	0.47	0.26
	5	19,724	4,801	21,556	0.43	0.26
	10	15,671	4,013	23,095	0.41	0.27
	15	12,531	3,205	24,314	0.42	0.28
	20	10,248	2,591	25,145	0.43	0.29
	25	8,648	2,000	25,731	0.45	0.30
10	**0**	18,588	2,928	19,661	0.53	0.24
	5	16,028	3,221	21,556	0.51	0.24
	10	13,317	2,522	23,095	0.49	0.26
	15	11,048	2,137	24,314	0.50	0.26
	20	9,339	1,916	25,145	0.51	0.28
	25	8,083	1,651	25,731	0.52	0.28
20	**0**	16,547	2,354	19,661	0.55	0.24
	5	14,335	2,514	21,556	0.53	0.24
	10	12,092	2,029	23,095	0.52	0.25
	15	10,233	1,772	24,314	0.52	0.26
	20	8,821	1,586	25,145	0.53	0.27
	25	7,764	1,397	25,731	0.54	0.28
30	**0**	15,241	1,976	19,661	0.56	0.25
	5	13,161	1,924	21,556	0.54	0.25
	10	11,147	1,673	23,095	0.54	0.26
	15	9,523	1,490	24,314	0.54	0.27
	20	8,306	1,360	25,145	0.55	0.28
	25	7,420	1,235	25,731	0.55	0.29

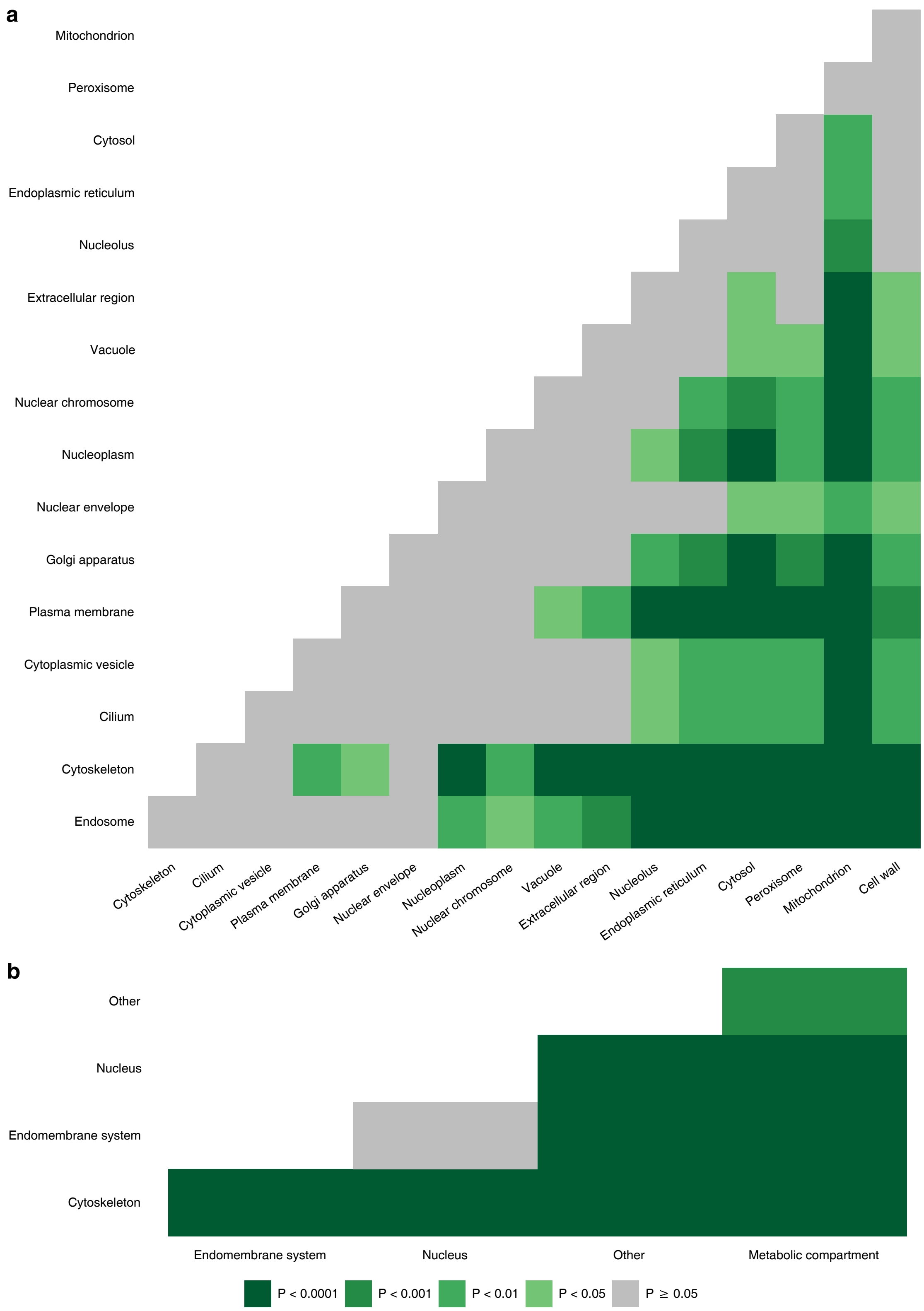
*Ultrafast bootstrap support value 95 or higher.

Supplementary Figures



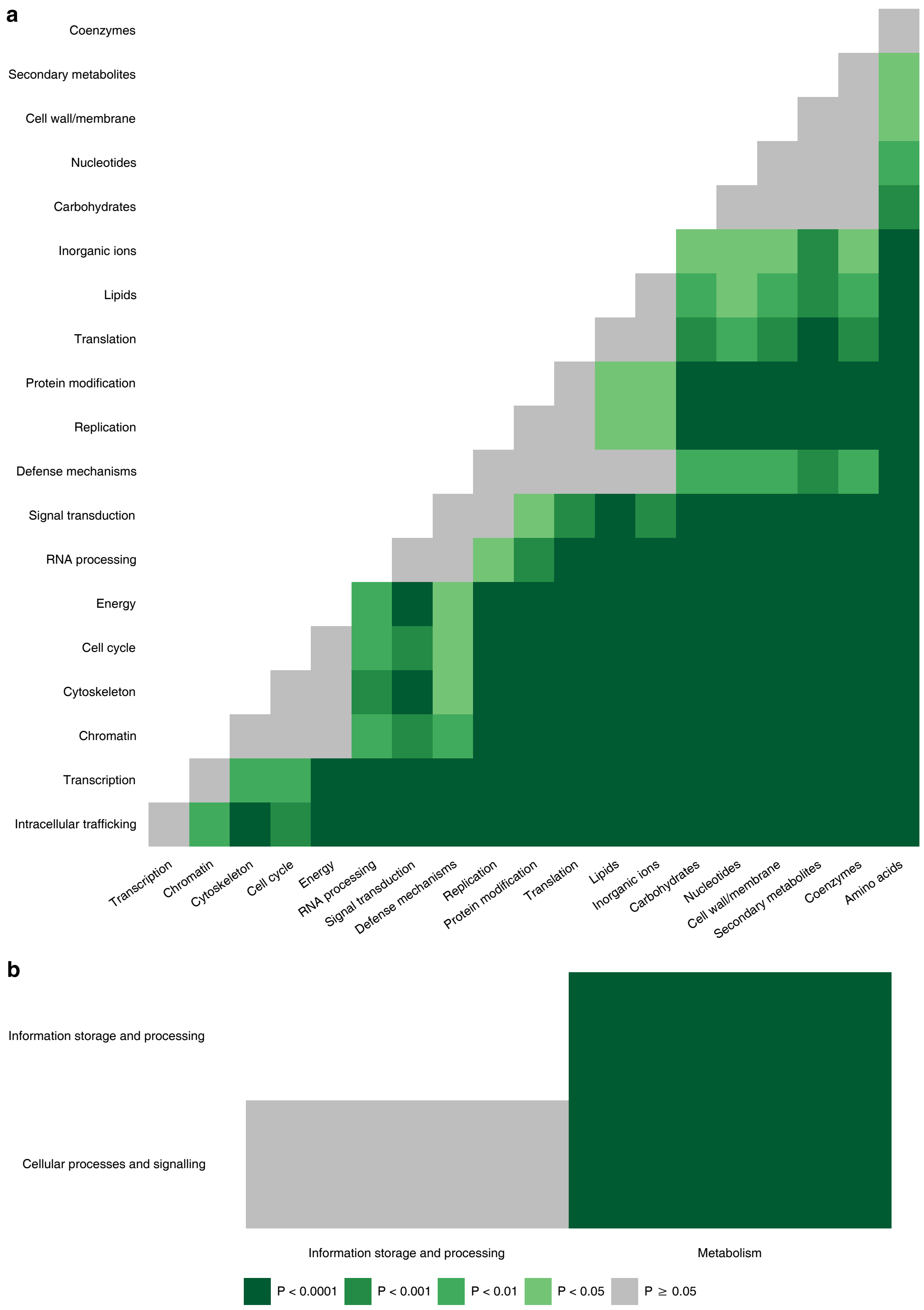
Supplementary Fig. 1 | Contribution of duplications to families with a particular function.

Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from duplications for different functional categories (**a**) and the corresponding broad categories (**b**). The values for each functional category are shown in Fig. 1c. The axis labels are ordered based on the odds of duplication.

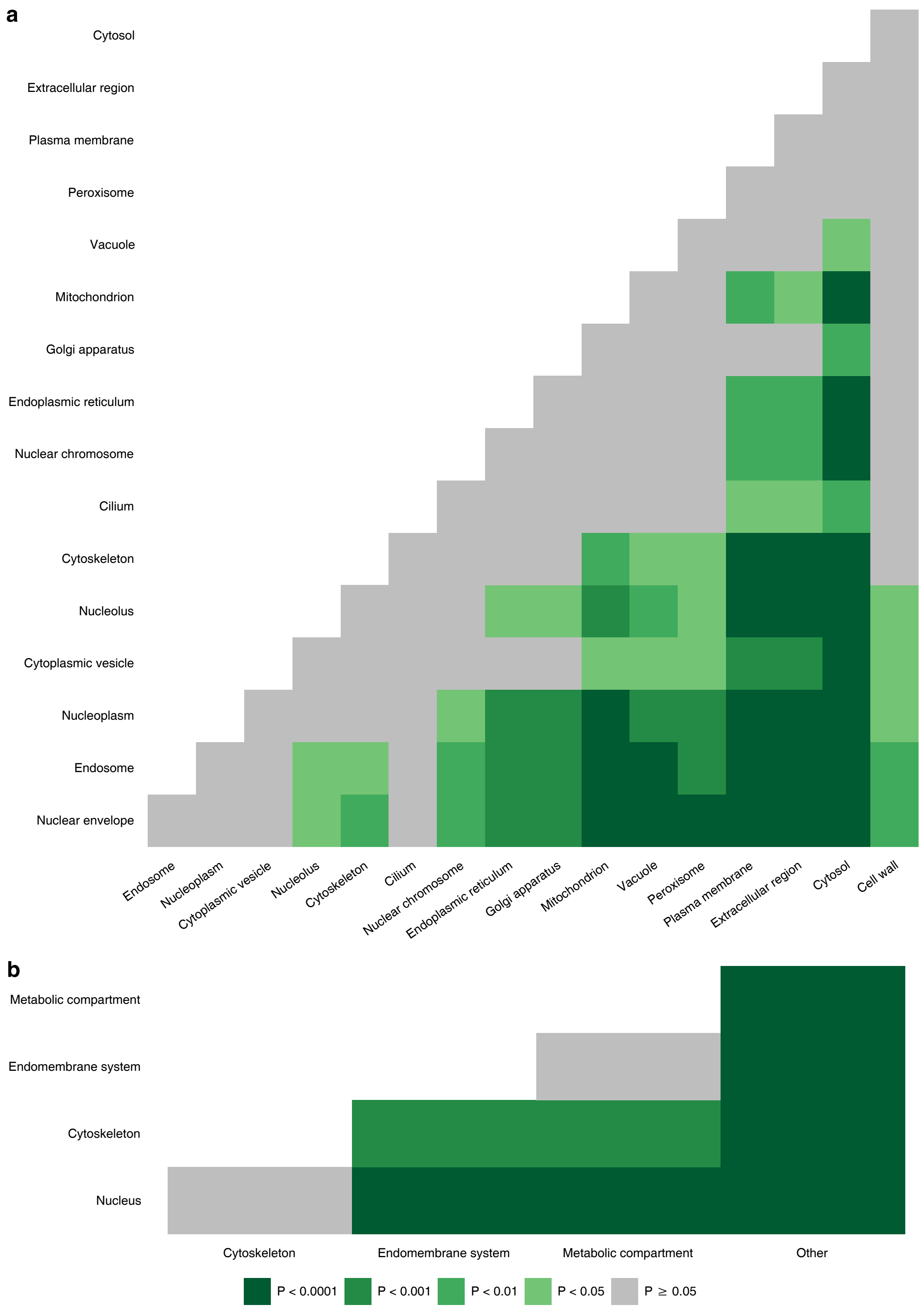


Supplementary Fig. 2 | Contribution of duplications to families with a particular cellular localisation.

Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from duplications for different localisations (**a**) and the corresponding broad categories (**b**). The values for each localisation are shown in Fig. 1d. The axis labels are ordered based on the odds of duplication.

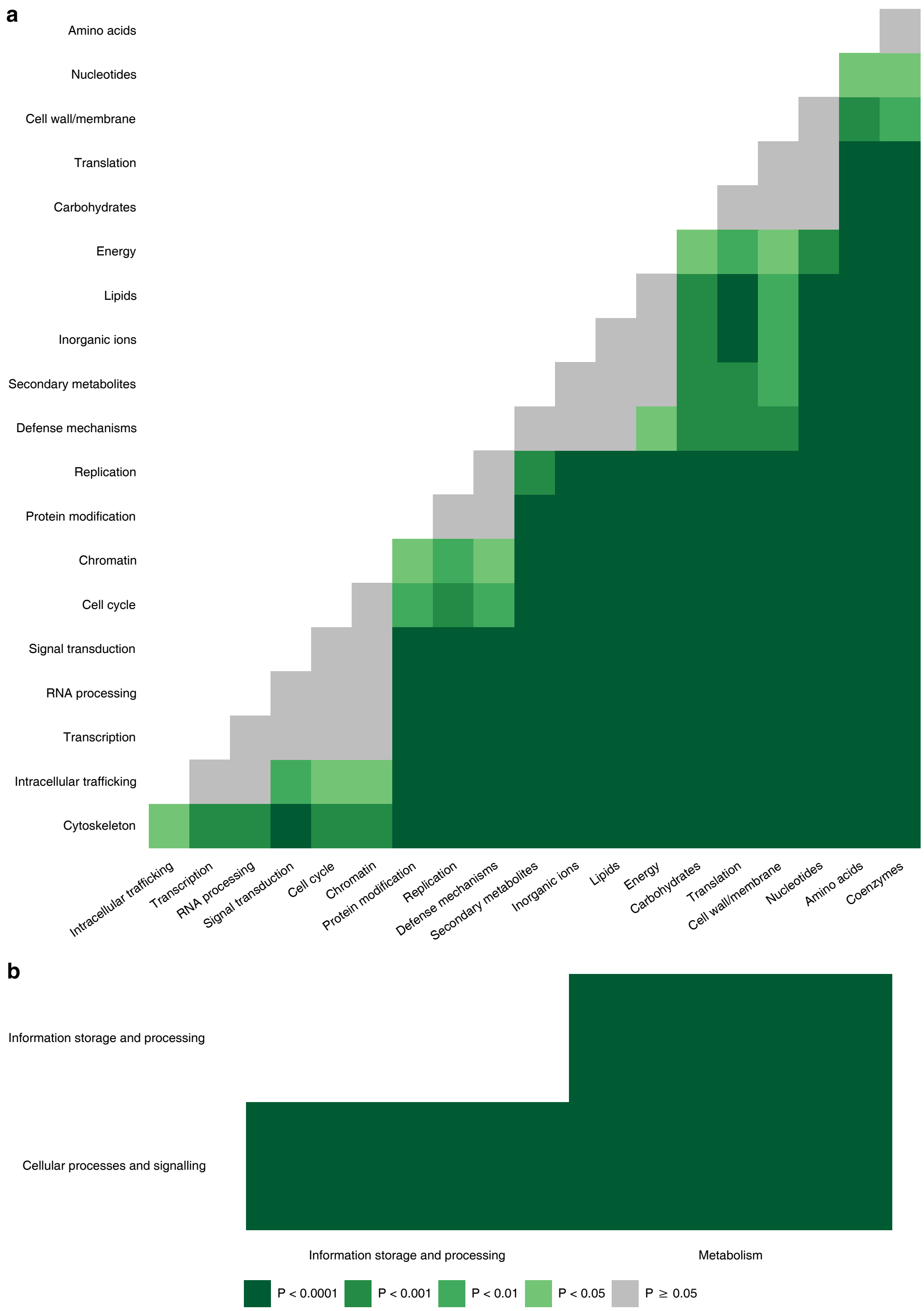


Supplementary Fig. 3 | Contribution of inventions to families with a particular function. Statistical significance of pairwise comparisons (Fisher's exact tests) between the proportions of LECA families being derived from inventions for different functional categories (**a**) and the corresponding broad categories (**b**). The values for each functional category are shown in Extended Data Fig. 3a. The axis labels are ordered based on the invented fraction.



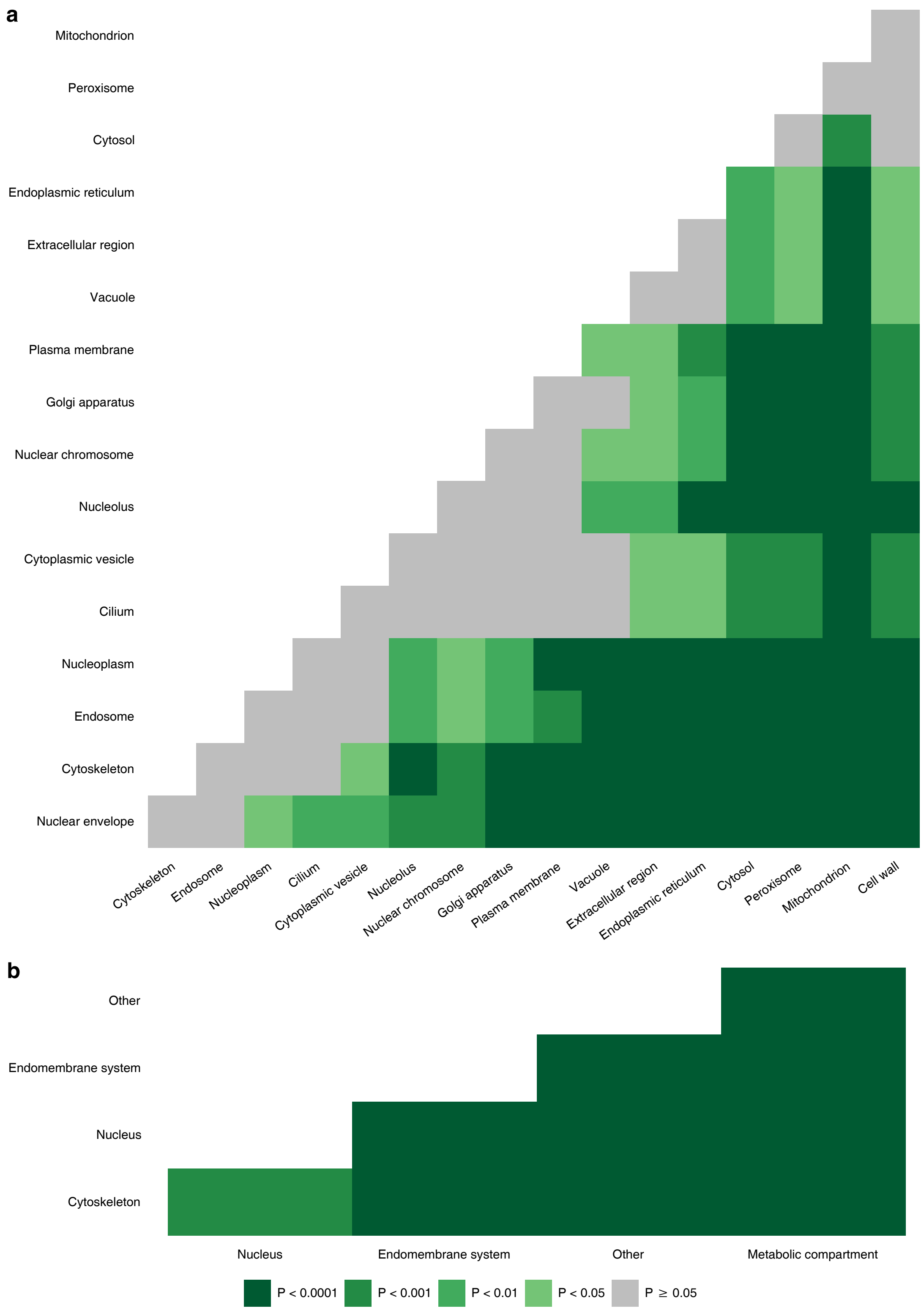
Supplementary Fig. 4 | Contribution of inventions to families with a particular cellular localisation.

Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from inventions for different localisations (**a**) and the corresponding broad categories (**b**). The values for each localisation are shown in Extended Data Fig. 3c. The axis labels are ordered based on the invented fraction.



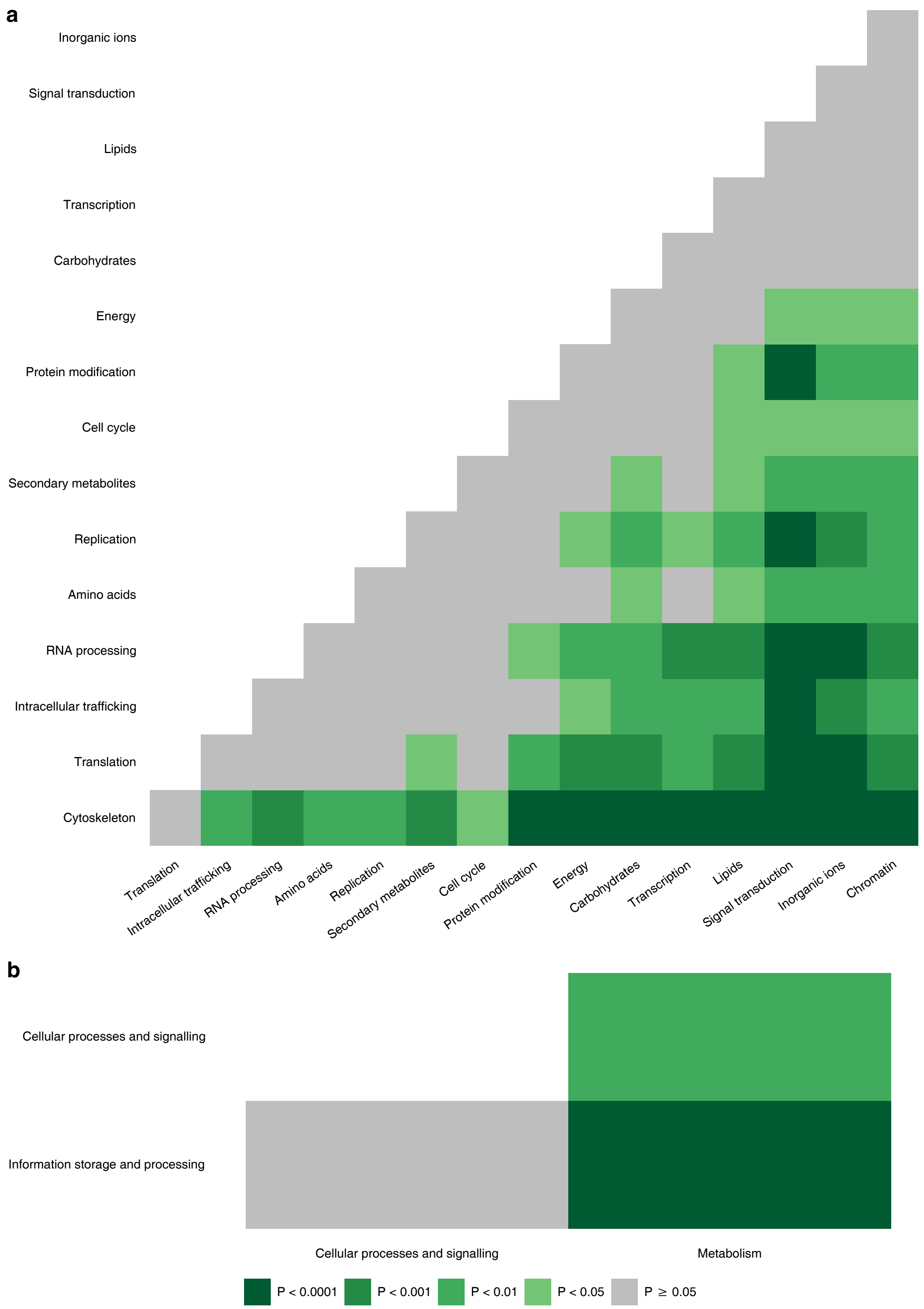
Supplementary Fig. 5 | Contribution of innovations to families with a particular function.

Statistical significance of pairwise comparisons (χ^2 contingency table tests) between the proportions of LECA families being derived from a eukaryotic innovation (invention or duplication) for different functions (**a**) and the corresponding broad categories (**b**). The values for each functional category are shown in Extended Data Fig. 3b. The axis labels are ordered based on the innovated fraction.

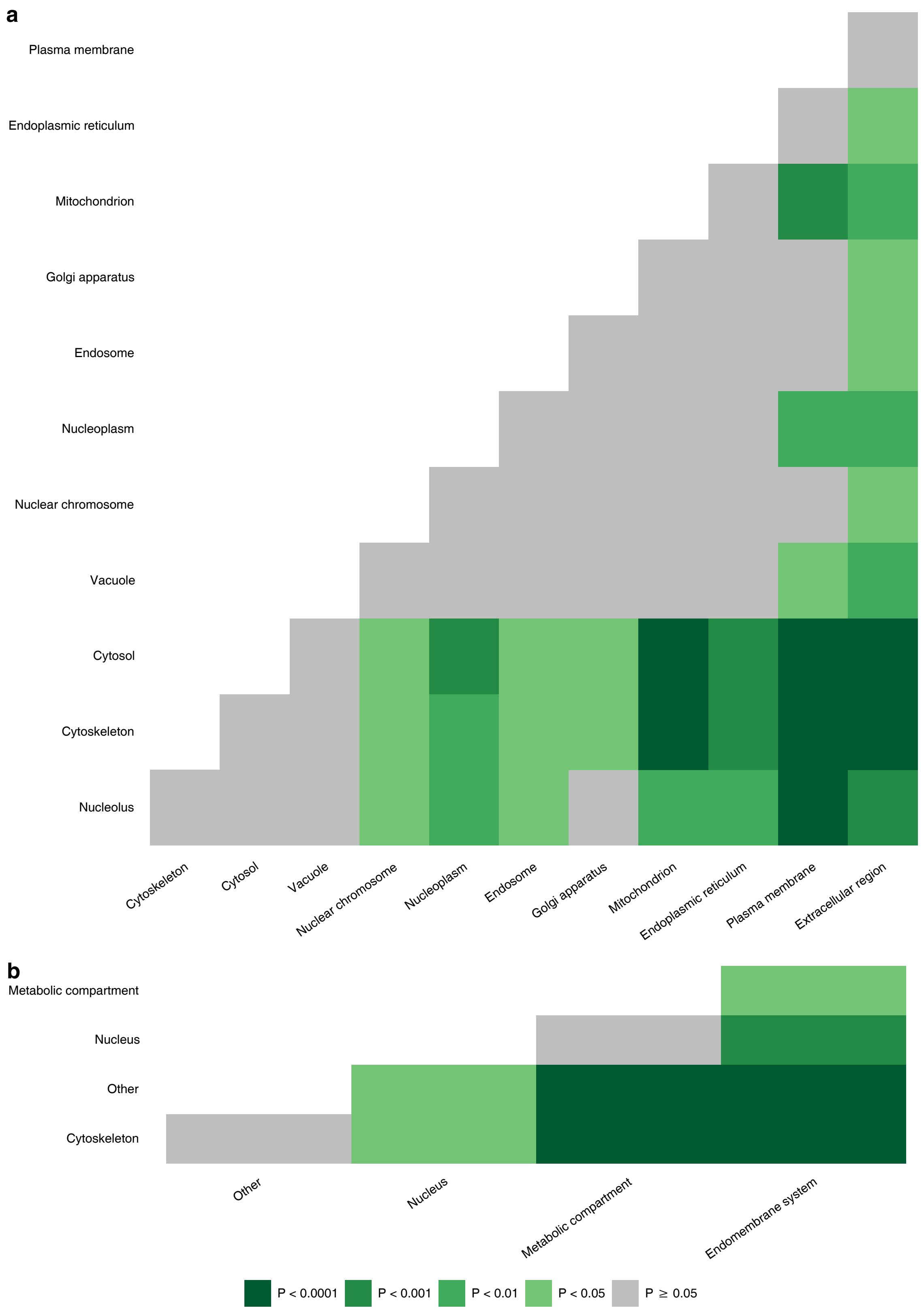


Supplementary Fig. 6 | Contribution of innovations to families with a particular cellular localisation.

Statistical significance of pairwise comparisons (Fisher's exact tests) between the proportions of LECA families being derived from a eukaryotic innovation (invention or duplication) for different localisations (**a**) and the corresponding broad categories (**b**). The values for each localisation are shown in Extended Data Fig. 3d. The axis labels are ordered based on the innovated fraction.



Supplementary Fig. 7 | Comparison of duplication lengths between different functions. Statistical significance of pairwise comparisons (Mann-Whitney *U* tests) between the duplication lengths for different functions (see Fig. 4a) (**a**) and the corresponding broad categories (**b**). The axis labels are ordered based on the median of duplication lengths.



Supplementary Fig. 8 | Comparison of duplication lengths between different cellular localisations.

Statistical significance of pairwise comparisons (Mann-Whitney *U* tests) between duplication lengths for different localisations (see Fig. 4b) (**a**) and the corresponding broad categories (**b**). The axis labels are ordered based on the median of duplication lengths.