Electronic Supplementary Material (ESI) for Molecular BioSystems

Supplemental Material for the paper: Protein disorder in the centrosome correlates with complexity in cell types number

G. S. Nido, R. Méndez, A. Pascual-García, D. Abia and U. Bastolla

Centro de Biología Molecular "Severo Ochoa", (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain

Figures list

Table list

Fraction of predicted disordered, coiled-coil and globular residues

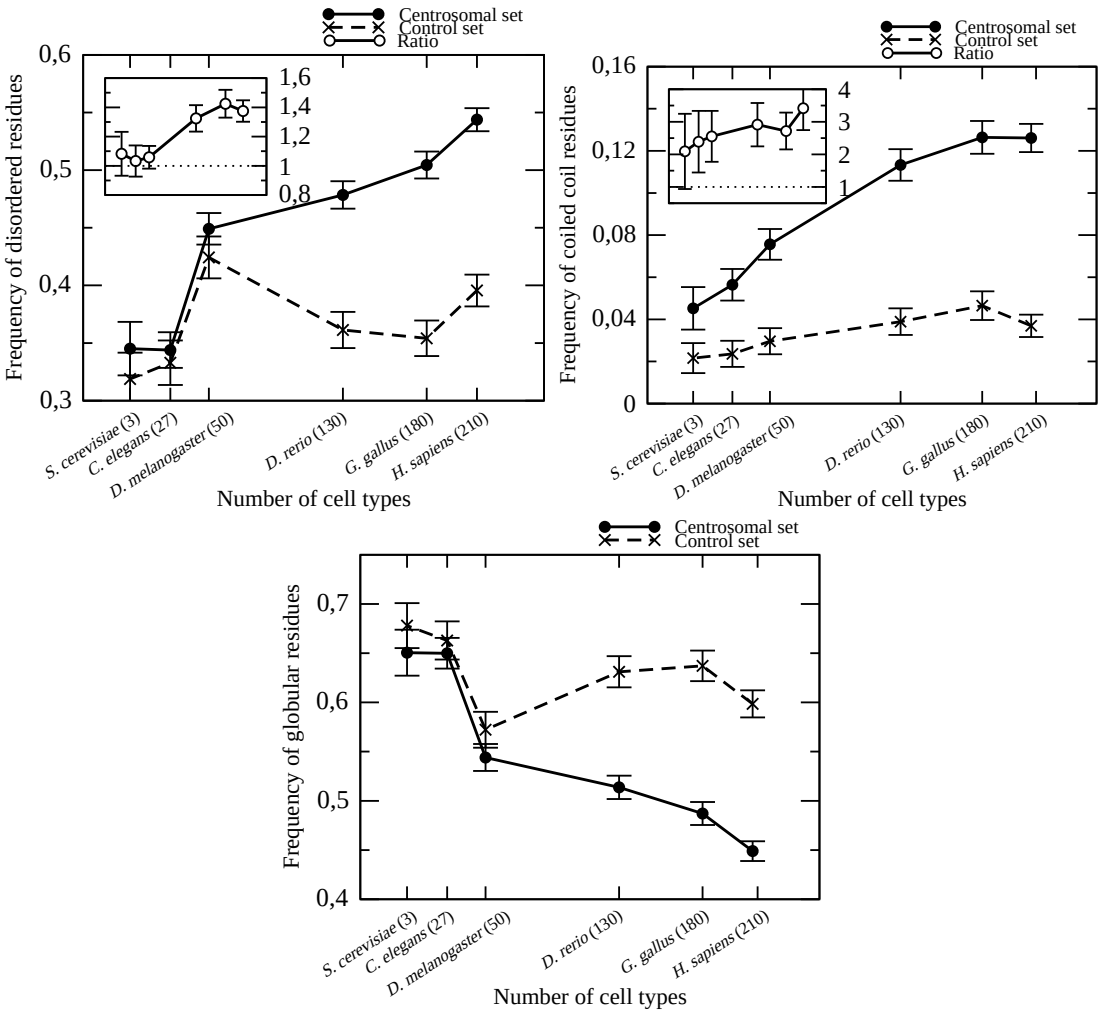

FIG. 1: Fraction of predicted disordered residues (top left) and fraction of predicted coiled-coil residues (top right) versus number of cell types for centrosomal (circles) and control proteins (crosses). Inset: ratio of the fraction of disordered or coiled-coil residues in centrosomal and control proteins. Bottom: fraction of predicted globular (neither disordered nor coiled-coil) residues in centrosomal and control proteins.

Robustness of results with respect to disorder prediction

We predicted disordered residues using four different algorithms: DISOPRED2 [1], FoldIndex [2], IUPred [3] and DisEMBL [4]. In Fig.3 we show that predictions based on FoldIndex and IUPred are qualitatively very similar to the predictions based on DISOPRED2 reported in the main text. In contrast, DisEMBL does not show significant differences between centrosomal and control proteins for vertebrates (bottom left). However differences become significant when the control data set is enlarged (bottom right). This is due to the fact that, as it was reported in a recent study [5], DisEMBL tends to miss the predictions of long disordered regions. We confirmed this trend in Supplemental Table I.

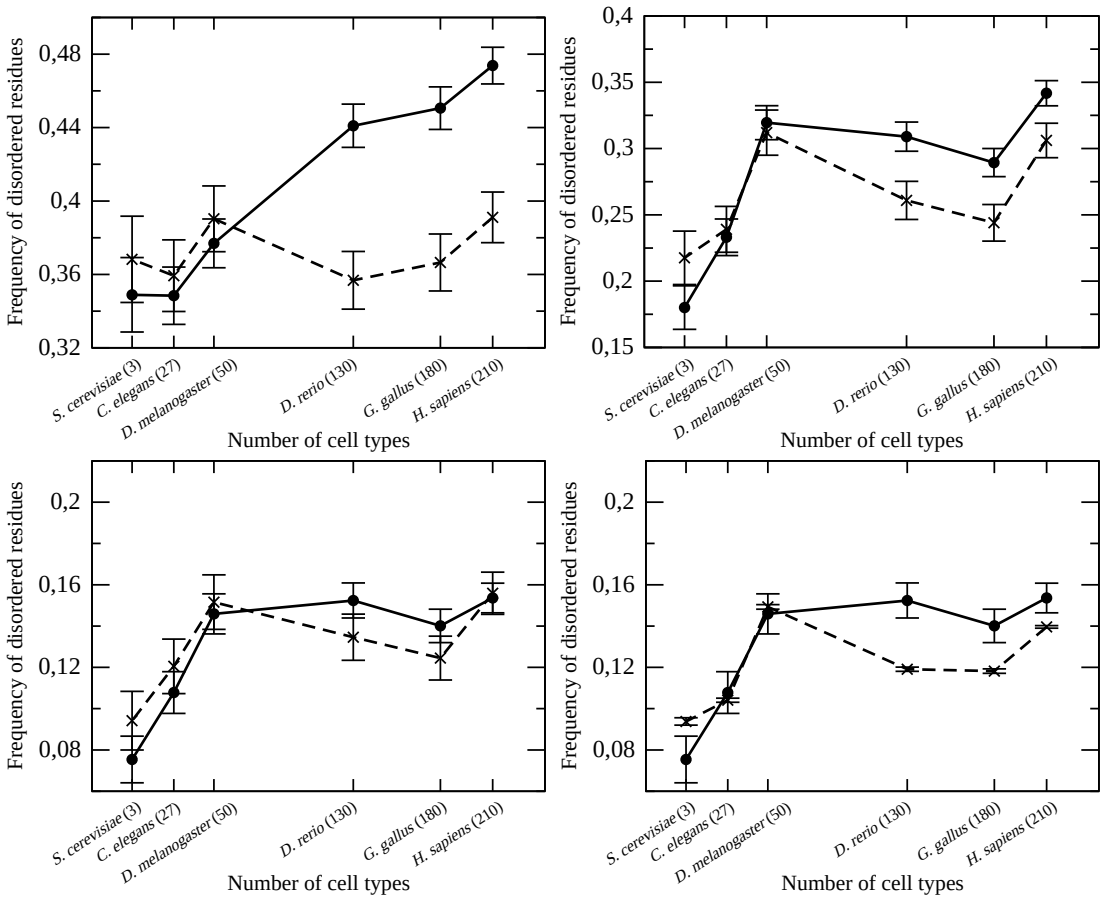

FIG. 2: Frequency of disordered residues predicted by different algorithms. Top left, FoldIndex [2], top right IUPred [3], bottom left DisEMBL [4] applied to the datasets used in this work. Bottom right DisEMBL applied to control datasets containing all of the available proteins in Ensembl for each genome.

Bias of DisEMBL in missing long disordered stretches

	DISOPRED2		**DisEMBL**	
	Centrosome	Control	Centrosome	Control
H. sapiens	0.86	0.70	0.24	0.28
G. gallus	0.86	0.73	0.27	0.29
D. rerio	0.84	0.74	0.25	0.27
D. melanogaster	0.80	0.77	0.23	0.32
C. elegans	0.74	0.66	0.23	0.23
S. cerevisiae	0.78	0.70	0.14	0.19

TABLE I: Fraction of residues contained in stretches of at least 40 consecutive predicted disordered residue (number of disordered residues in long stretches divided by the total number of predicted disordered residues) for various predictors and data sets. Different from other predictors, DisEMBL has a bias to predict only short stretches of disordered residues.

Robustness of results with respect to coiled-coil prediction

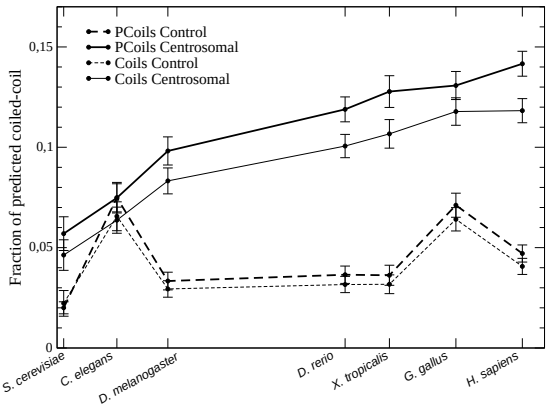

FIG. 3: Frequency of coiled-coil residues predicted by the algorithms PCOILS (thick lines) and ncoil (thin lines). Note that for this figure an independent control data-set has been used for all organisms.

Propensity of disorder and coiled-coil predictions to co-occur

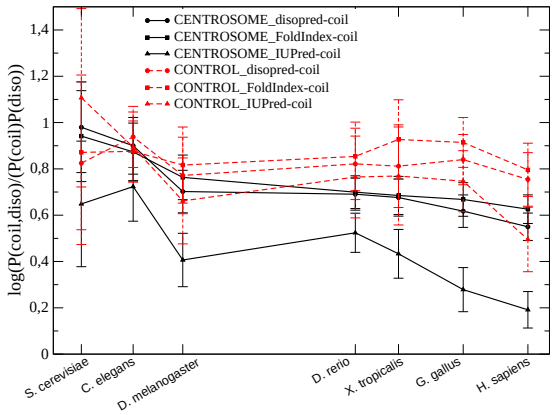

FIG. 4: Propensities $p(x, y) = \ln\left(\mathrm{P}(x, y)/\mathrm{P}(x, y)\mathrm{P}(x, y)\right)$, where x and y represent the event that a given site is predicted as disordered and coiled-coil, for various pairs of disorder and coiled-coil predictions. Positive propensity means that x and y tend to co-occur more than at random (here this refers to disorder and coiled-coil predictions). Propensities were significantly positive for all data-sets and all pairs of disordered and coiled-coil predictors, except for a few datasets using the DisEMBL predictor (not shown). Using *ncoil* or *Pcoils* for coiled-coil predictions yields the same propensities within the statistical error (not shown). Therefore, the correlation between disorder and coiled-coil does not depend on the predictor used. Note that propensities are slightly but systematically larger for control than for centrosomal proteins and, for the latter, they tend to decrease with organism complexity, consistent with the fact that in centrosomal proteins of more complex organisms there is a larger fraction of residues predicted to be disordered but not coiled-coil.

Robustness of results with respect to the protein length distribution

We predicted disordered and coiled-coil residues for an additional control dataset consisting of 500 random proteins for each of the model species, with the same length distribution as the centrosomal set, using bins of 50 amino acids of length. For disorder predictions, we used the DISOPRED2 algorithm. In Fig.5 we show that we obtain the same results with this control set, both for disorder and coiled-coil frequency.

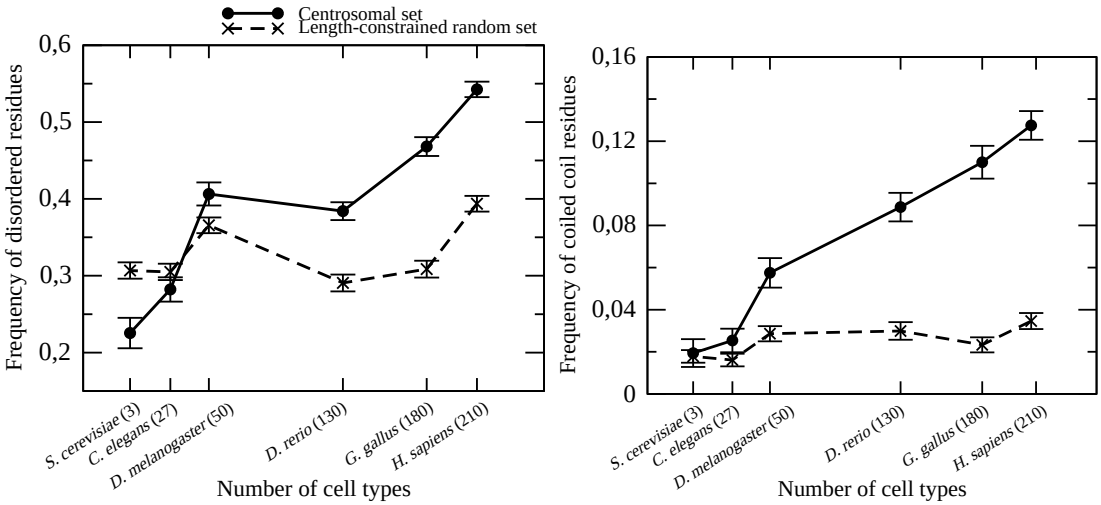

FIG. 5: Frequency of disordered (left) and coiled-coil residues (right) for the centrosomal set and the length-constrained random set.

Pairwise flux of disordered and coiled-coil residues

We consider here the disorder flux from species a to species b, which we defined as the number of changes from residues that are ordered in species a and disordered in species b, minus the number of changes from residues that are ordered in species b and disordered in species a for each kind of mechanism and each pair of species. Specifically, , we computed the difference between the number of residues that are disordered in b and not disordered in a and those that are disordered in a and not disordered in b, normalized by the total number of changes produced by that mechanism. In this way, we assess the presence and strength of a bias in disorder transitions along the evolutionary paths that join species a and b to their common ancestor. In all cases, results are presented in Fig. 6 with species a being more complex than species b (larger number of cell types). A positive value therefore means that there is a net flux from order to disorder along the evolutionary branch leading to the more complex species.

The first row of Fig. 6 presents changes due to new proteins, i.e. disordered residues present in a protein that has no ortholog in the other species. The disorder balance is always positive for the more complex species except for the comparison between frog and zebra-fish where it is approximately zero. This can in part be explained by the fact that the data sets represent orthologs of human proteins. However, the bias is much stronger for centrosomal than for control proteins. For instance around 60% of the residues in human centrosomal proteins having no orthologs in the other species are disordered, whereas this percentage reduces to 40% for control proteins. In the case of *H. sapiens*, this percentage is almost independent of the species compared, which may be due to the fact that genes of other species always have an ortholog in *H. sapiens* by construction of the data sets. For other species a, the more distant the compared species b, the larger the bias, as expected. Also for other species the bias is much larger for centrosomal than for control proteins. For instance between yeast and chicken the bias is around 50% for centrosomal proteins and 35% for control proteins (in both cases, the balance is positive).

The flux due to large indels is reported in the second row of Fig. 6. The disorder flux in the comparison between human proteins and proteins of non-mammals vertebrates is much larger for the centrosome, where it reaches 89%, than for the control (40%). The flux is always positive for comparisons between vertebrates, except for the comparison between frog and zebra-fish, where it is almost zero, both for centrosomal and for control proteins. The disorder flux due to large indels always goes towards the fly for all pairwise comparisons except with human centrosomal proteins. For worm, the disorder flux is significantly higher for centrosome than for control proteins, being

always positive for centrosomal proteins except for the comparison with the frog, whereas for control proteins it is negative except in the comparisons with human and fly. Finally, the disorder flux with respect to yeast proteins is negative for centrosomal proteins, except in the comparison with the fly, while for control proteins also human and worm present a positive flux, i.e. yeast centrosomal proteins tend to be more disordered than their orthologs in more complex species. Comparing the first and second row in Fig. 6, we can see that Drosophila proteins have a strong tendency to gain disorder through large indels rather than through new proteins, which is consistent with its proteins being the longest among our model organisms except human, (see main text), despite its genes contain significantly fewer exons than vertebrate genes. *D. rerio* on the other hand has the shortest proteins in both sets, and its protein lost disorder through long indels in all comparisons except with the frog. Note however that *D. rerio* also has the largest number of centrosomal genes (paralogs) among our model organisms. Yeast is characterized by a very small number of centrosomal proteins, and it often gains disorder through large indels. These observations suggest that there is a trade-off between the tendency to gain disorder through new proteins and through large indels.

Small indels are intermediate between large indels and substitutions and they produce almost zero flux, so we do not discuss them. Substitutions are represented in Fig. 6 third row. They produce very small flux (note the scale). There are some interesting trends: yeast proteins gained disorder through substitutions both for centrosomal and, more strongly, control proteins; *C. elegans* control proteins (but not centrosomal ones) lost disorder through substitutions, *D. melanogaster* control proteins tend to gain disorder through substitutions; higher vertebrates centrosomal proteins tend to gain disorder through substitutions with respect to the frog proteins.

In Fig. 6, the disorder flux due to a given type of change is normalized by dividing it by the total number of changes of that type. This normalization quantifies the strength of the bias.

The analogous picture for the flux of coiled-coil residues through new proteins, large insertions and substitutions is presented in Fig.7. It is qualitatively similar to the flux of disordered residues, except that the flux of disorder due to new proteins and large insertions is much stronger than the analogous flux of coiled-coil residues, and the flux of coiled-coil residues due to substitutions is much stronger than the corresponding flux of disordered residues. The difference with respect to the control is typically larger for the coiled-coil flux than for the flux of disordered residues.

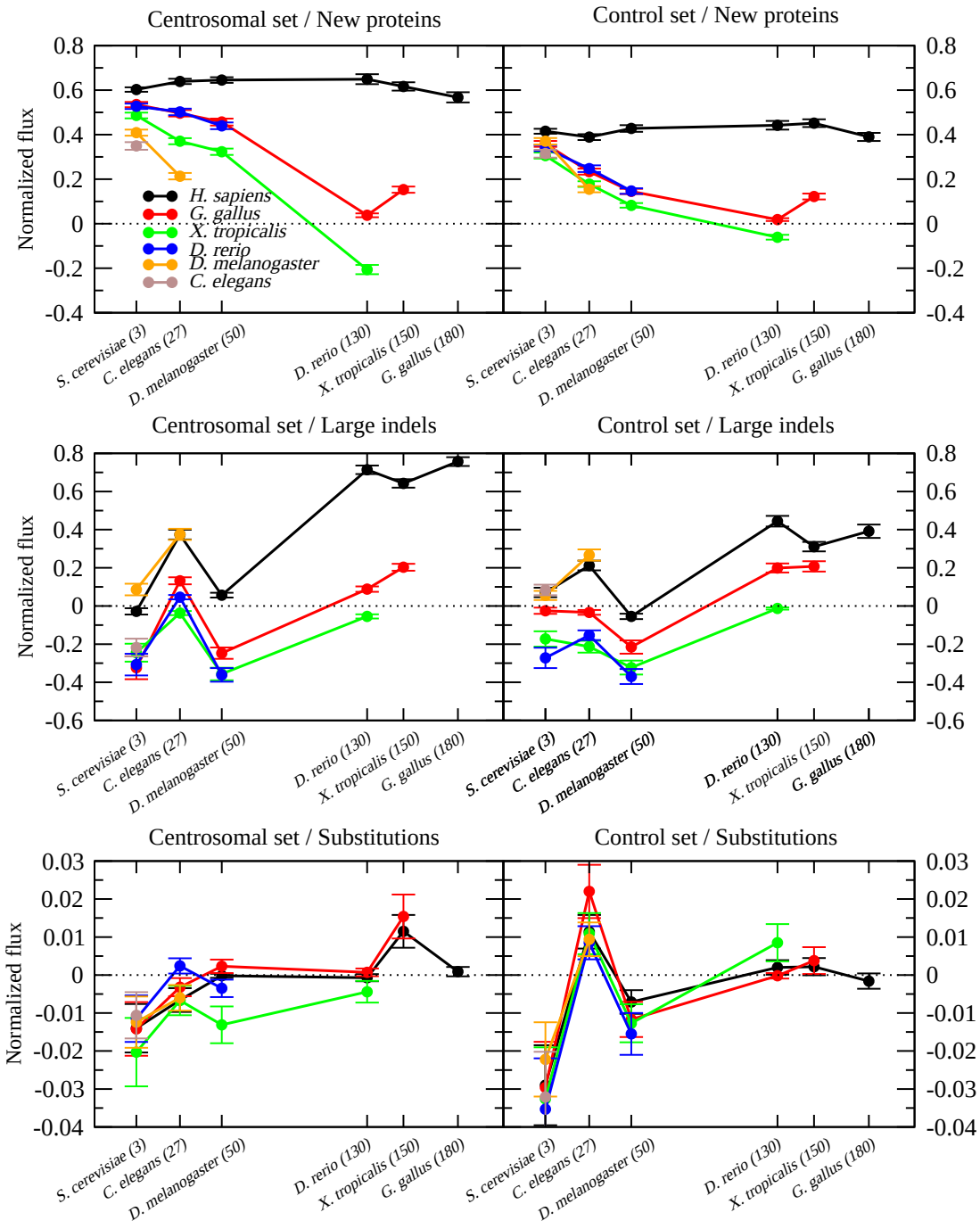

FIG. 6: Disorder flux. For each type of genetic change and each pair of species, we compute the difference between the number of residues that are disordered in the more complex species and ordered or absent in the other species minus those that are ordered or absent in the more complex species and disordered in the other, normalized by the total number of changes. Changes correspond to proteins that do not have orthologs in the other species (top row), large insertions (middle row, more than 20 a.a.) and substitutions.

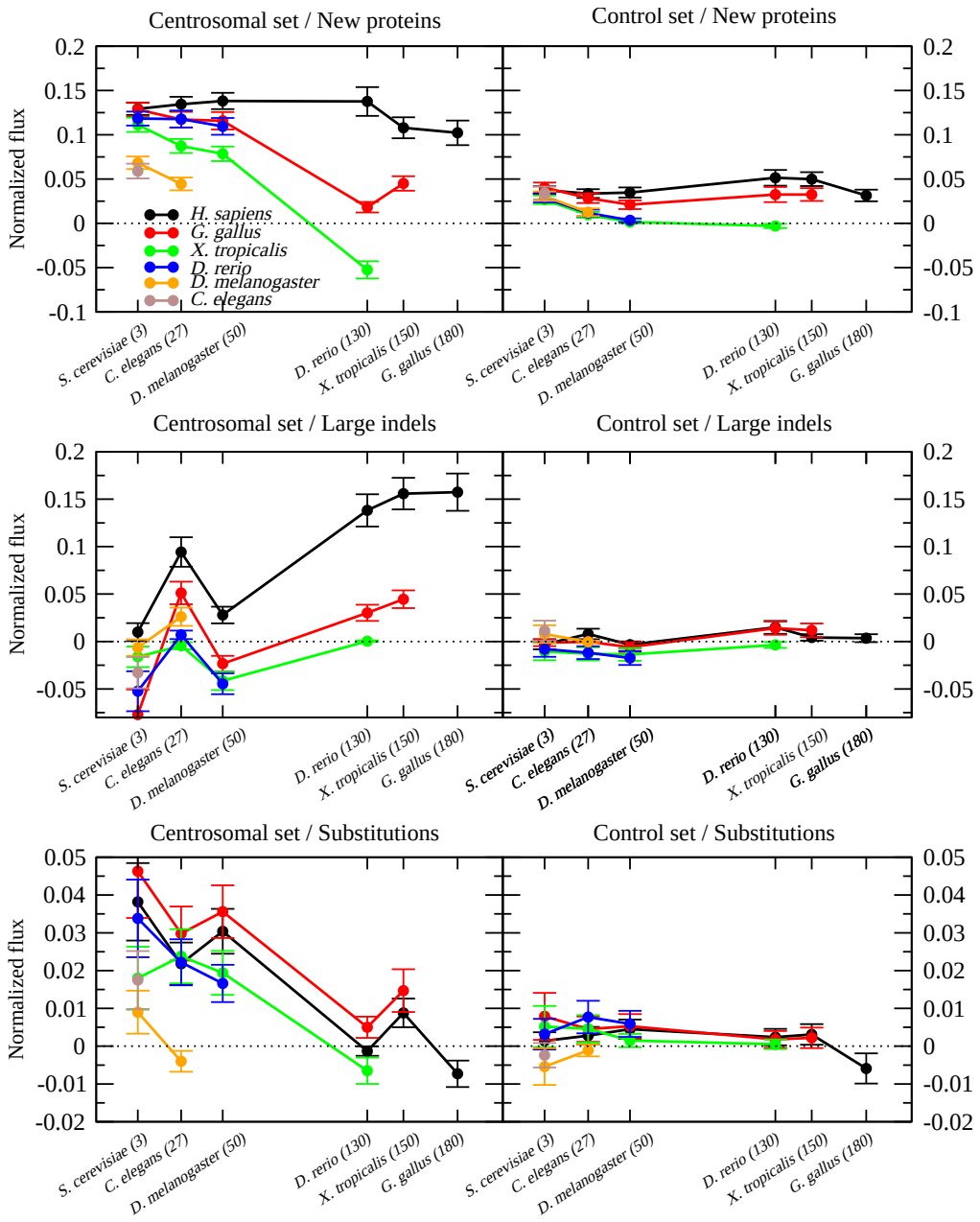

FIG. 7: Coiled-coil flux. For each type of genetic change and each pair of species, we compute the difference between the number of residues that are coiled-coil in the more complex species and not coiled-coil or absent in the other species minus those that are not coiled-coil or absent in the more complex species and coiled-coil in the other, normalized by the total number of changes. Changes correspond to proteins without ortholog in the other species (top row), large insertions (middle row, more than 20 a.a.) and substitutions. Note that the flux of coiled-coil residues due to substitutions is large and positive on the branch leading to the more complex species (in the comparisons between vertebrates and invertebrates) or essentially zero in the comparisons between vertebrates, being much stronger for centrosomal than for control proteins.

Robustness of results with respect to the threshold on sequence identity

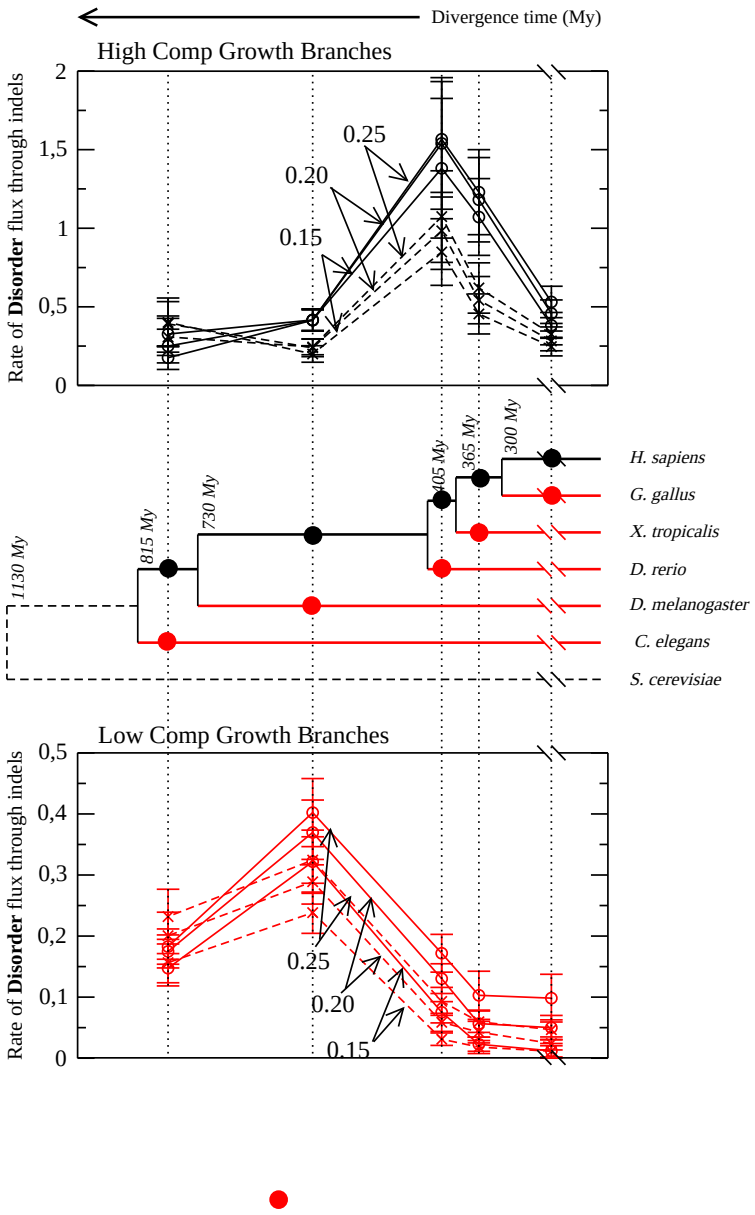

FIG. 8: In our evolutionary reconstruction of indel events, we clustered together long insertions with sequence identity above a length dependent threshold $t = s + (1-s)4/L$ where L is he insertion length. We show here that changing the sequence identity parameter s in the range from 0.15 to 0.25 does not change the results qualitatively. Namely, significant differences between the centrsome (solid lines) and the control (dashed line) are conserved for all s. The difference between High Complexity Growth and Low Complexity Growth branches decreases with s but, if it is significant at $s = 0.15$, it remains significant at least up to $s = 0.25$.

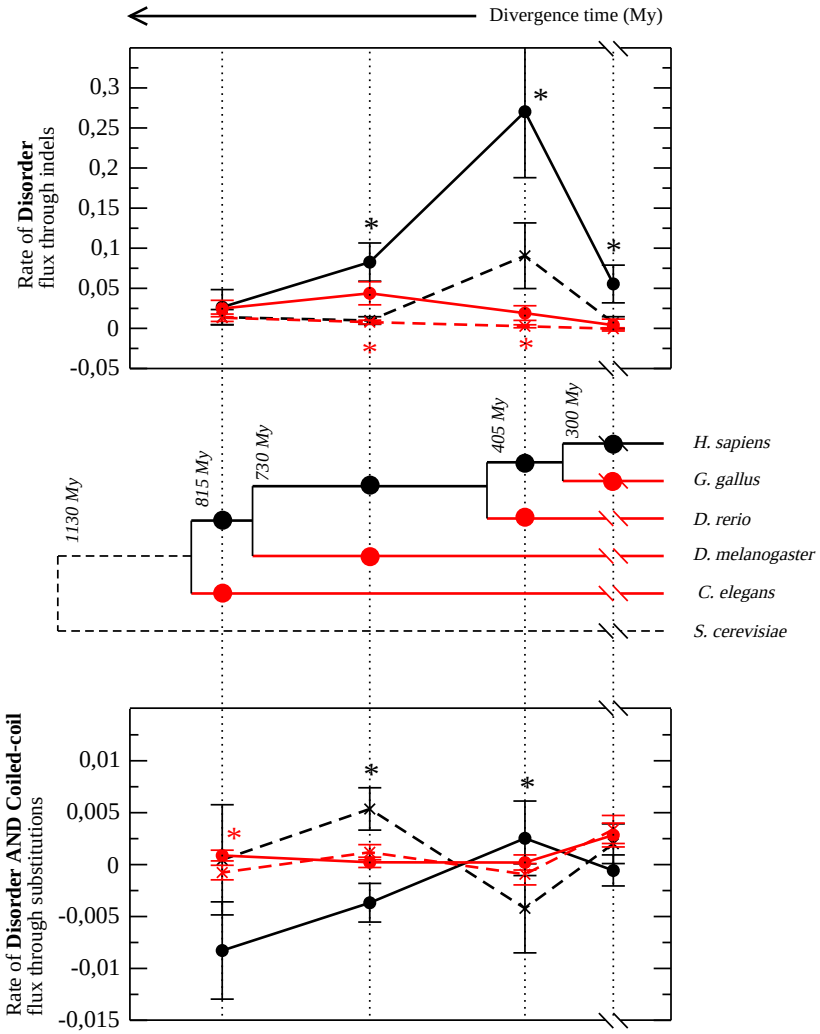

FIG. 9: Rates of coiled coil flux due to large indels (top) and disorder flux due to substitutions (bottom) along the branches of the phylogenetic tree of model species. The abscissa indicates the time at which the branch starts to diverge, for instance $t = 750My$ represents the split between vertebrates and fly. The figure is based on the Coelomata hypothesis. Not assuming this hypothesis gives similar results (see Supplementary Fig. 10. For each node, we distinguish the HCG branch with larger increase in cell types (black) and the LCG branch with smaller increase in cell types (red). The flux in each branch is normalized by the number of proteins at the ancestral node and divided through the length of the branch to obtain the rate.

Evolutionary analysis not relying on the Coelomata hypothesis

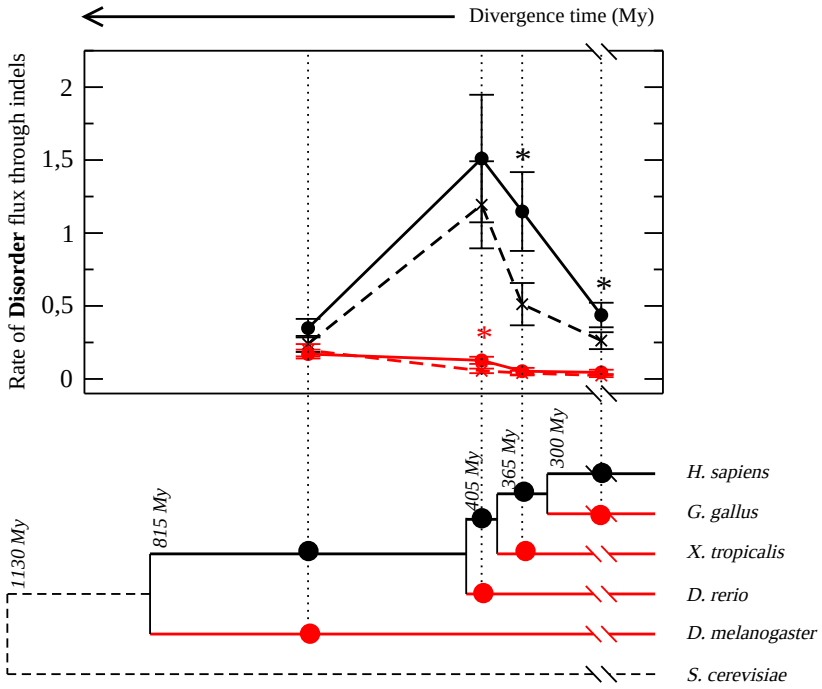

FIG. 10: Rates of disorder increase due to indels per protein and million years reconstructed on trees that do not rely on the Coelomata hypothesis, i.e. omitting *D. melanogaster*. Omitting *C. elegans* yields a qualitatively identical plot. Control lines represent centrosomal proteins, dashed lines represent control proteins.

[1] Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337(3):635-45.

[2] Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex(C): a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21(16):3435-3438.

[3] Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433-3434.

[4] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. 2003. Protein disorder prediction: implications for structural proteomics. Structure 11(11):1453-9.

[5] Sirota FL, Ooi H, Gattermayer T, Schneider G. Eisenhaber F, Maurer-Stroh S. 2010. Parametrization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. BMC Genomics 11(Suppl 1): S15.