Supplementary material

Supplementary Text

Pfam domain content is not the sole determinant of protein traceability.

Individual proteins have traceabilities close to one across the entire tree of life despite the absence of Pfam domains (see supplementary fig. S6). In turn, examples abound where the traceability is low although at least one Pfam domain could be annotated in the sequence. This highlights that other factors such as protein specific substitution rates reflected in the scaling factor κ, and also protein specific indels rates influence protein traceability. We therefore explored the dependency of protein traceability on protein-specific evolutionary rate captured in κ, and on the indels rates, respectively. We grouped the yeast proteins into four bins depending on their traceability in *E. coli*: (i) <0.25, (ii) 0.25 – 0.5, (iii) 0.5 – 0.75, and (iv) >0.75. We then selected from each bin randomly 25 yeast proteins. For each of these 100 proteins, we subsequently doubled and halved its κ, respectively, and assessed the effect on the protein's traceability. Likewise, we changed the indels rates by a factor of 10 and 0.1. The results are shown in supplementary fig. S5. Note, that a change of the indels rates by an order of magnitude was necessary to observe a noticeable effect in the mean traceabilities. Supplementary fig. S5 shows that the traceability is negatively correlated with both rates, however the change of κ has a substantially stronger effect. The figure, however, also suggest that only slight changes of the evolutionary parameters, as they may be caused by the variance of the evolutionary parameter estimates should not have a severe effect on the traceability estimates.

Sensitivity and specificity of the ortholog search tool

Spurious ortholog assignments can be a further reason for incongruences between traceability of a protein and of its phyletic distribution. A recent benchmark has again revealed that so far, no ortholog assignment tool is error free, and individual approaches differ in both sensitivity and specificity (Altenhoff, et al. 2016). Obviously, both will have an effect on whether or not an ortholog is detected for a seed protein with a given traceability. For example, our results slightly change, when we switch the ortholog search procedure for the 6,352 yeast proteins. Using the OMA-based (Roth, et al. 2008) ortholog search, we detect in about 5% of the cases a eukaryotic ortholog despite a predicted traceability of below 0.75. If we repeat the same analysis, this time determining the phylogenetic profiles across eukaryotes with OrthoDb (Zdobnov, et al. 2017), for which the authors claim a higher sensitivity, the fraction of identified eukaryotic orthologs with traceability below 0.75 increases slightly to 7%. In such instances, only a case-by-case assessment of whether or not the additionally identified candidates indeed represent genuine orthologs can resolve the issue.

Supplementary Tables

Supplementary table S1 | List of 232 representative species from the three domains of life

Supplementary table S2 | Traceabilities of 6352 S. cerevisiae proteins in 232 representative species

Supplementary table S3 | Traceabilities of yeast proteins in E. coli and classification into essential genes and the LUCA genes

Supplementary table S4 | Traceability analysis of the Mycoplasma mycoides genes representing the minimal gene set for a self-replicating cell (Syn3.0)

Supplementary table S5 | Phylogenetic profile and traceaebility of yeast proteins involved into core metabolic pathways in microsporidia

Supplementary table S6 | Phylogenetic profile and traceaebility of yeast meiotic proteins in microsporidia

Supplementary Figures

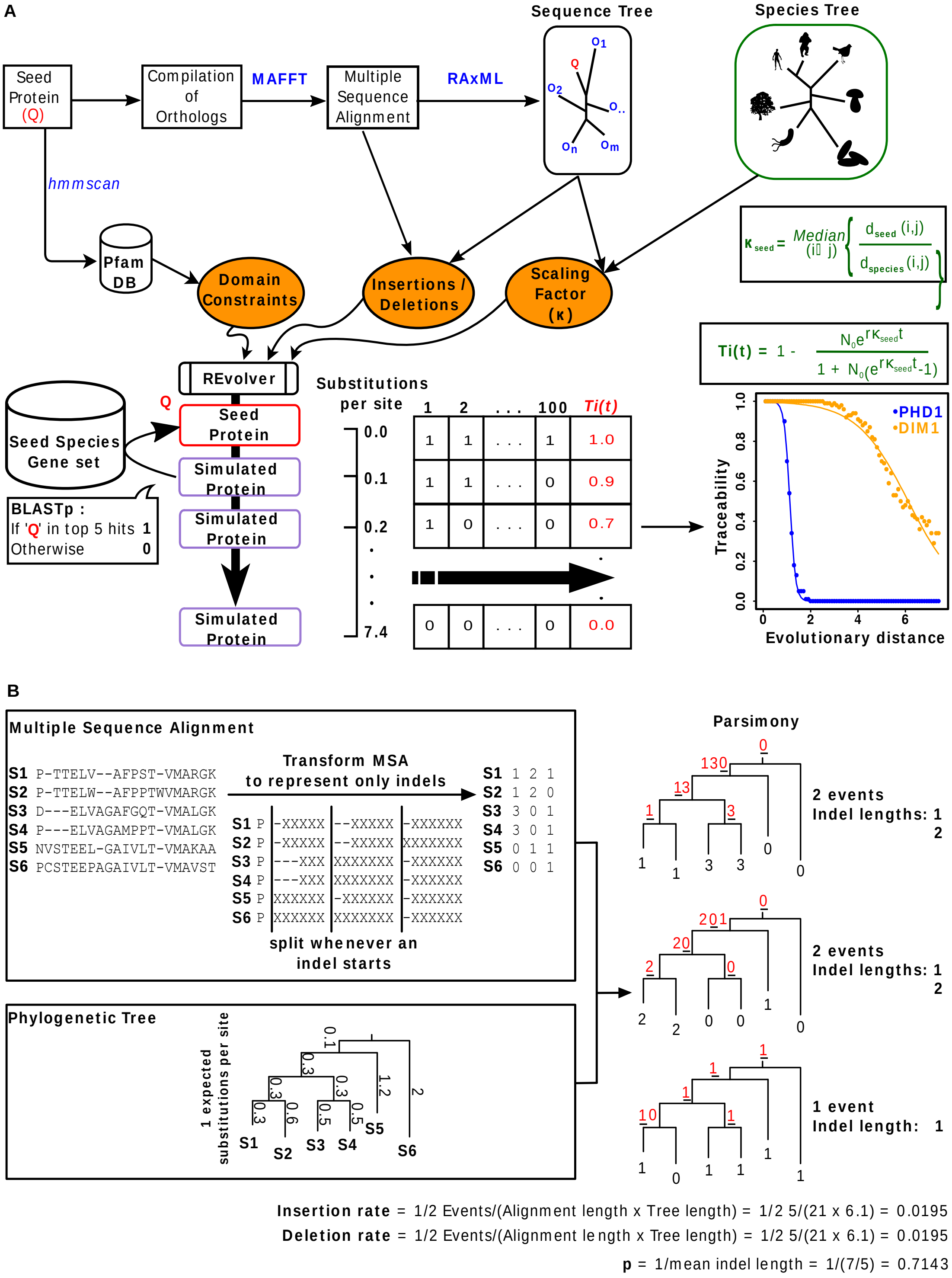

Figure S1 | The workflow of protTrace. **A**, Overview of the individual steps to assess the evolutionary traceability, *Ti(t)*, of a protein. The procedure is described in full detail in the Results section of the main text. **B**, Maximum parsimony based approach to estimate insertions / deletions (indels) rates and length distribution parameters. We split the MSA whenever a gap starts. Subsequently, we construct a transformed alignment by counting the gaps (if any) for every sequence in each split alignment part. We then calculate the maximum parsimony score for each column of the transformed alignment given the tree inferred earlier from the original alignment. Here, the maximum parsimony score is the number of insertions and deletions required to obtain the transformed alignment. Insertion and deletion rates per position, respectively, are then obtained by dividing the half of the number of events by the product of the tree length and the alignment length. The insertion and deletion lengths of one most parsimonious solution are used to infer p, the parameter for the geometric length distribution.

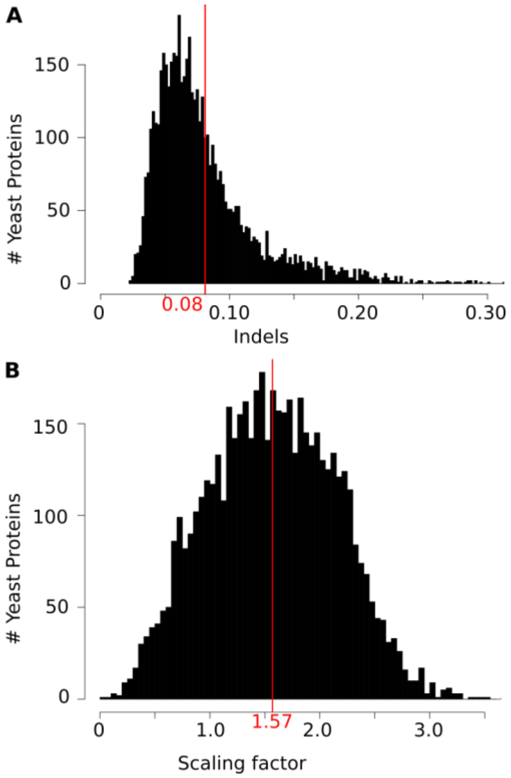

Figure S2 | Distribution of evolutionary parameter estimates across the yeast gene set. **A**, The histogram shows the distribution of the insertion/deletion (indel) rates estimated for all yeast proteins having at least three orthologs. The mean value is indicated in red. **B**, The histogram shows the distribution of the scaling factor κ_{seed} for all yeast proteins. The mean is indicated in red.

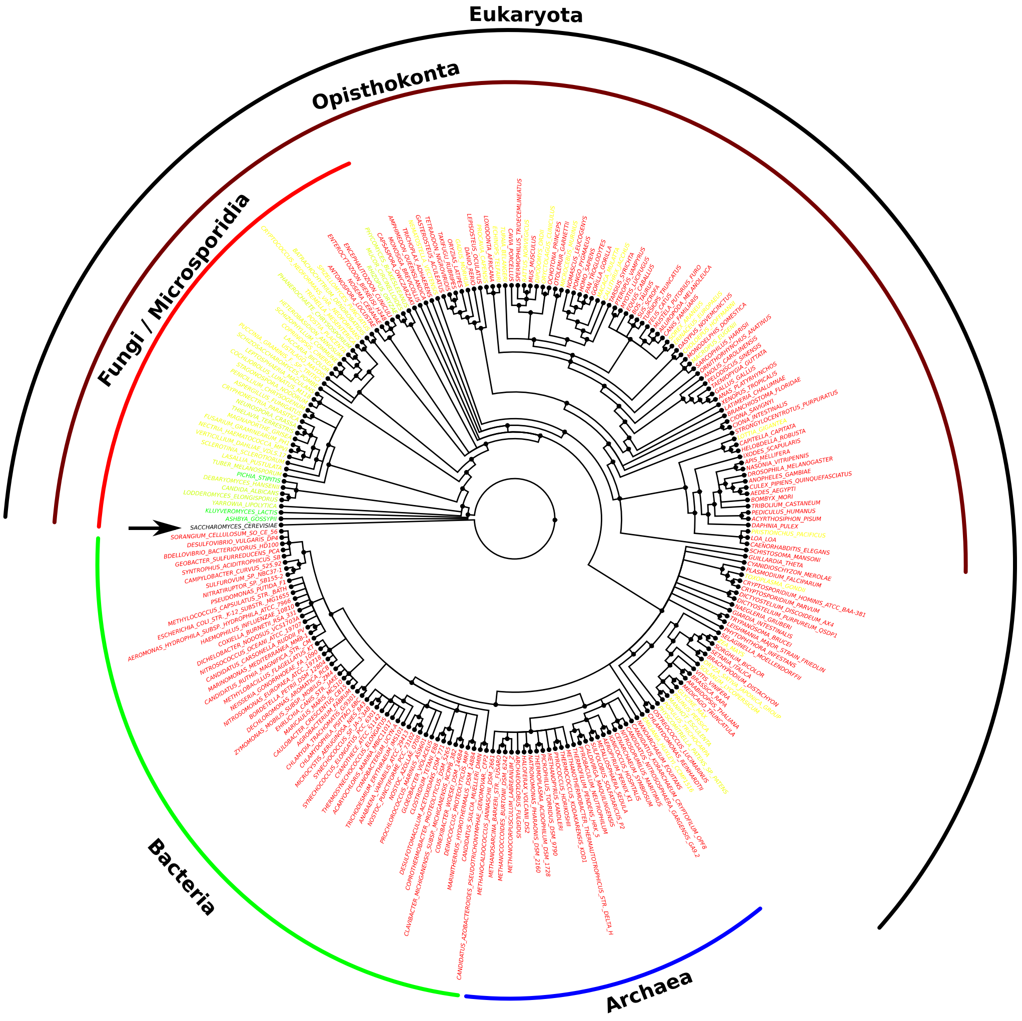

Figure S3 | Tree view of the traceability of yeast MSR2 across the 232 target taxa. The black arrow indicates the position of *Saccharomyces cerevisiae,* the species the seed-protein was derived from. Green taxon labels indicate a high, yellow an intermediate, and red a low traceability of yeast MSR2 in the respective species. The cladogram was rooted with S. cerevisiae.

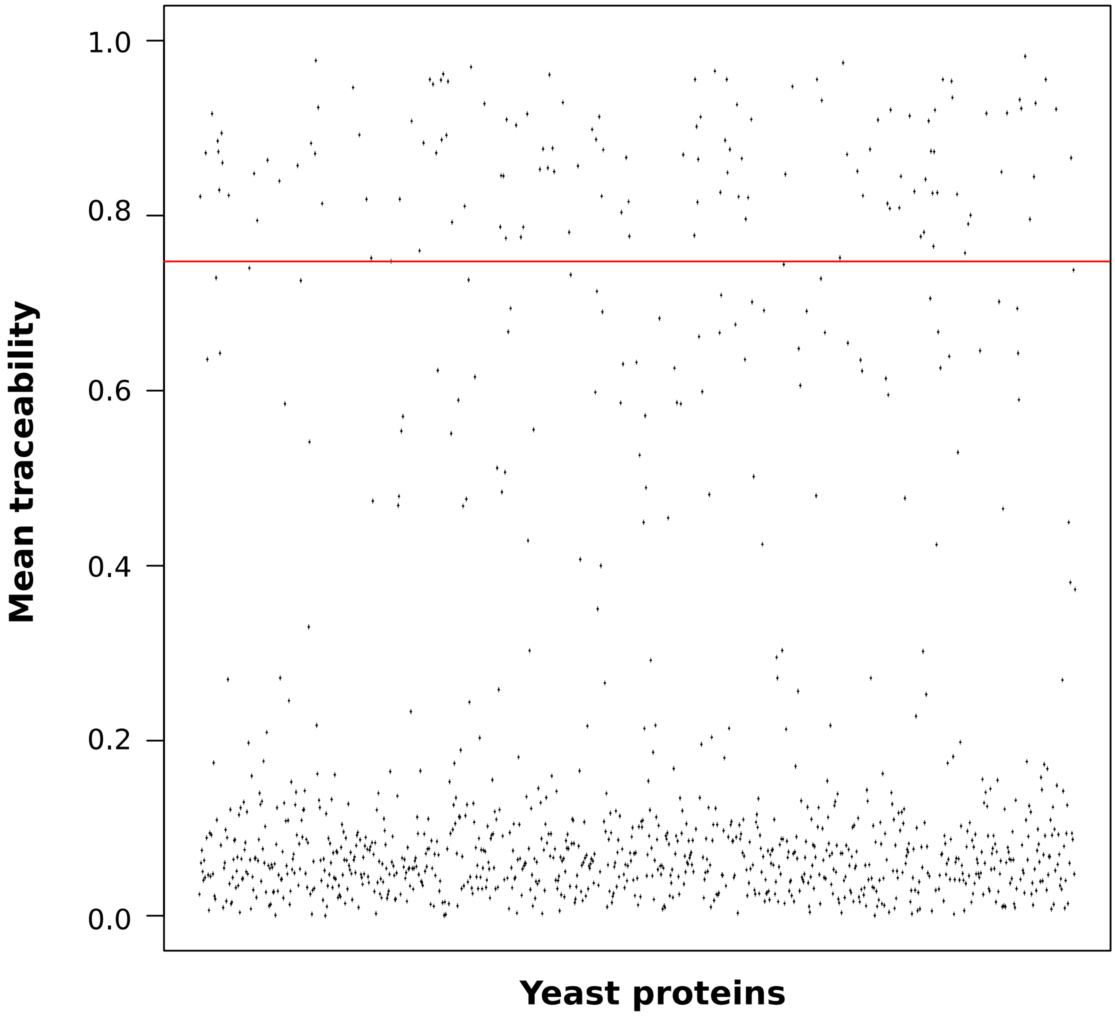

Figure S4 | Mean traceabilities for the proteins with default scaling factor and default indel rate. The figure shows the distribution of mean traceabilities for the yeast proteins without orthologs. In these cases, we could not empirically assess the protein-specific scaling factor κ_{seed} and the parameters for modelling the indel process. Instead, we used the default values of κ_{seed} = 1.57 and an indel rate of 0.8 (*see* supplementary figure S2). While most proteins have an overall low traceability, there is a considerable fraction with mean traceabilities of 0.75 and above (red line). This indicates that the use of the default values for the evolutionary rate estimates does therefore not determine a low traceability.

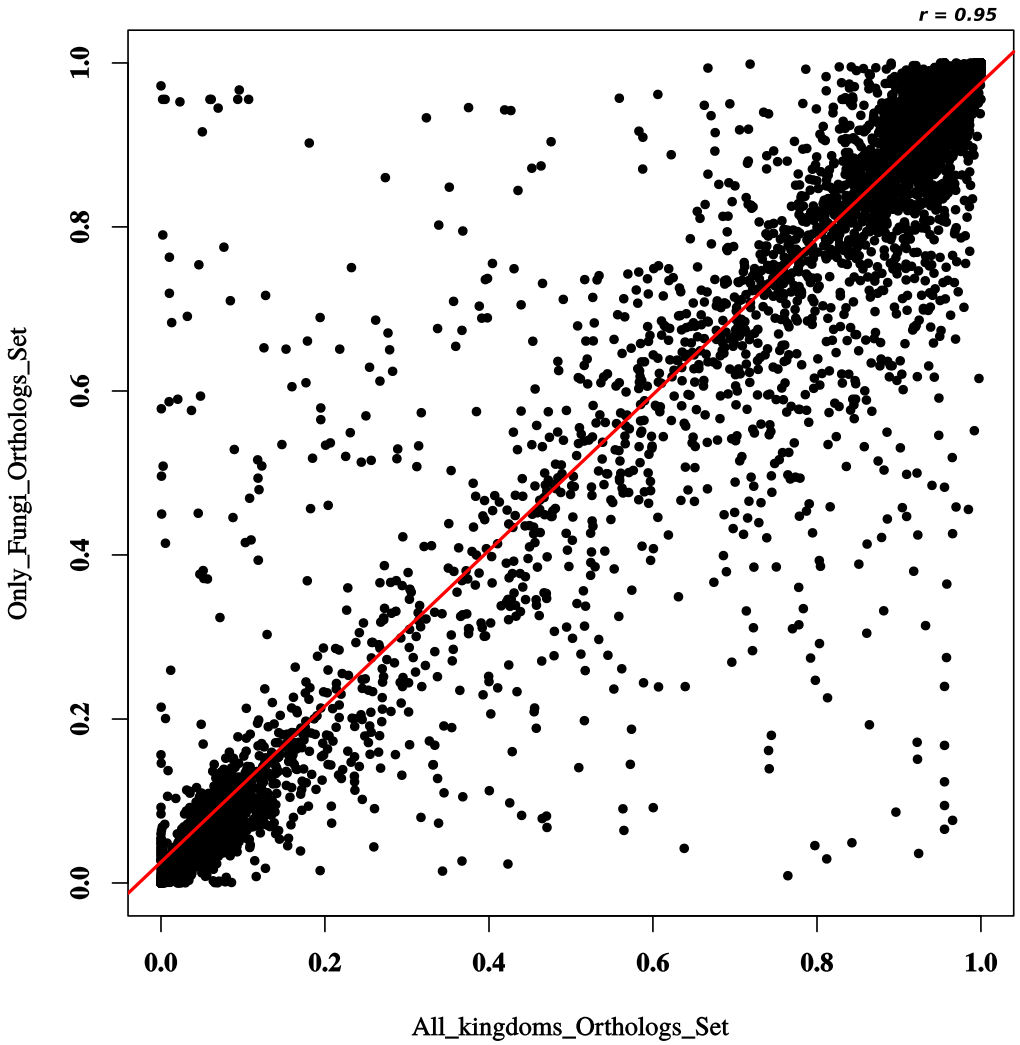

Figure S5 | Mean traceabilities of the yeast protein set based on different training data. We computed the protein-specific evolutionary parameters for the yeast proteins using orthologs from the full set of 232 species (x axis), and only from fungal species (y axis). The resulting mean traceability estimates are largely unaffected by the difference in diversity of the underlying training data (r = 0.95). This indicates that the phylogenetic diversity of the training data has almost no impact on the traceability estimates for the yeast proteins.

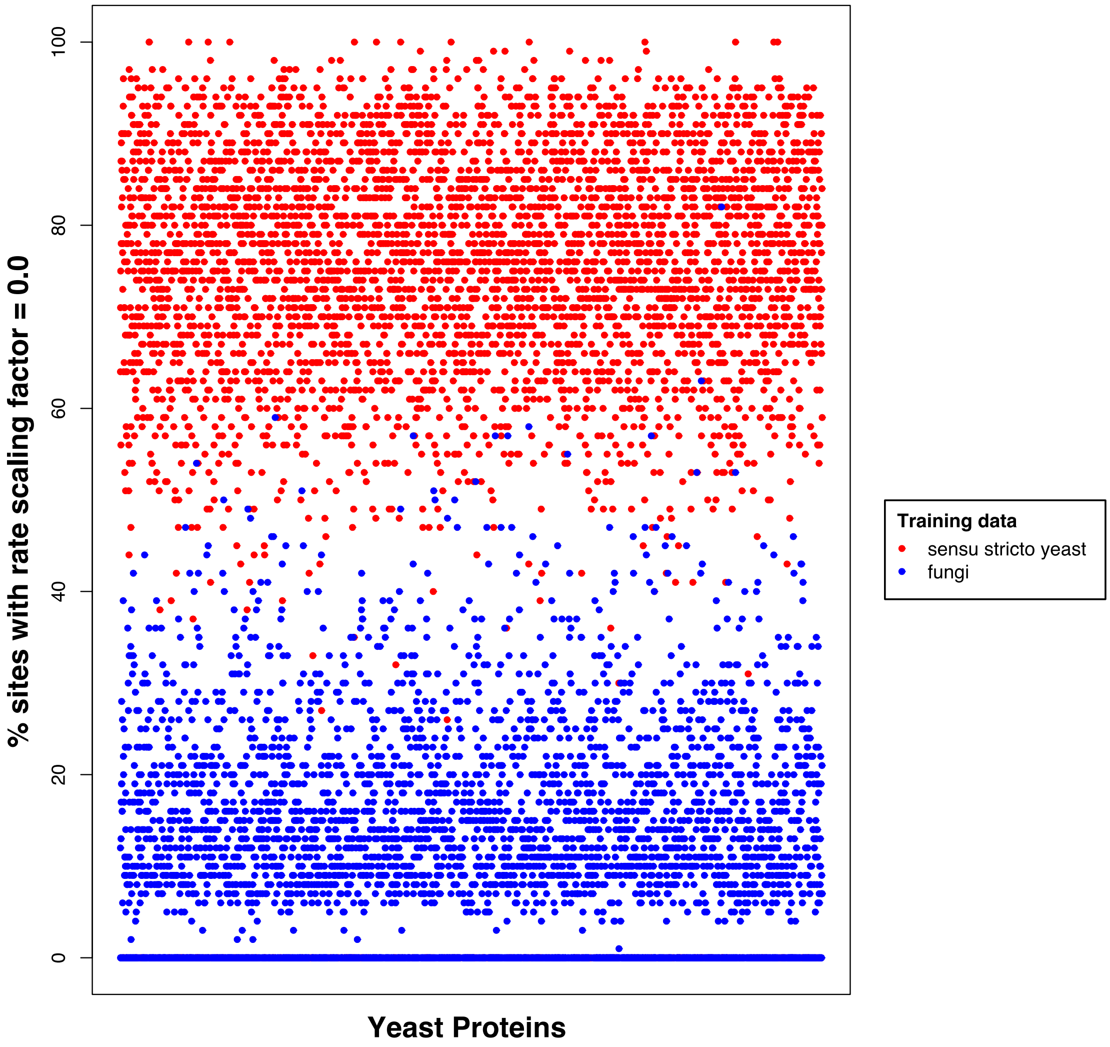

Figure S6 | Influence of the training data on the site-specific rate scaling factor estimation. We compiled for the 5,259 yeast proteins analyzed by Moyers and Zhang (2016) the training data as described in the original publication. We then estimated the relative rates per site with TreePuzzle using a discrete Γ distribution with 16 rate categories, again in analogy to Moyers and Zhang (2016). The plot shows for each alignment the fraction of sites with a relative rate of 0 (red dots). We then repeated the analysis for the same yeast proteins, this time using an alignment of a phylogenetically diverse set of fungal orthologs to infer the site specific rates (blue dots). The analysis reveals a substantially influence of the composition of the training data on the estimation of the site rates. The use of the evolutionary closely related set of sensu stricto yeast orthologs for inferring the constraints results in a substantial fraction of positions with relative rates of 0. Such positions will remain constant in the course of simulated evolution, and as a consequence result in a high traceability of the respective protein. If, however, the phylogenetically diverse set of orthologs is used for inferring the relative rates for the same set of sequences, the fraction of constant sites decreases substantially. As a consequence, the sequences are now more free to change in the course of simulated evolution, and their traceability will decrease.

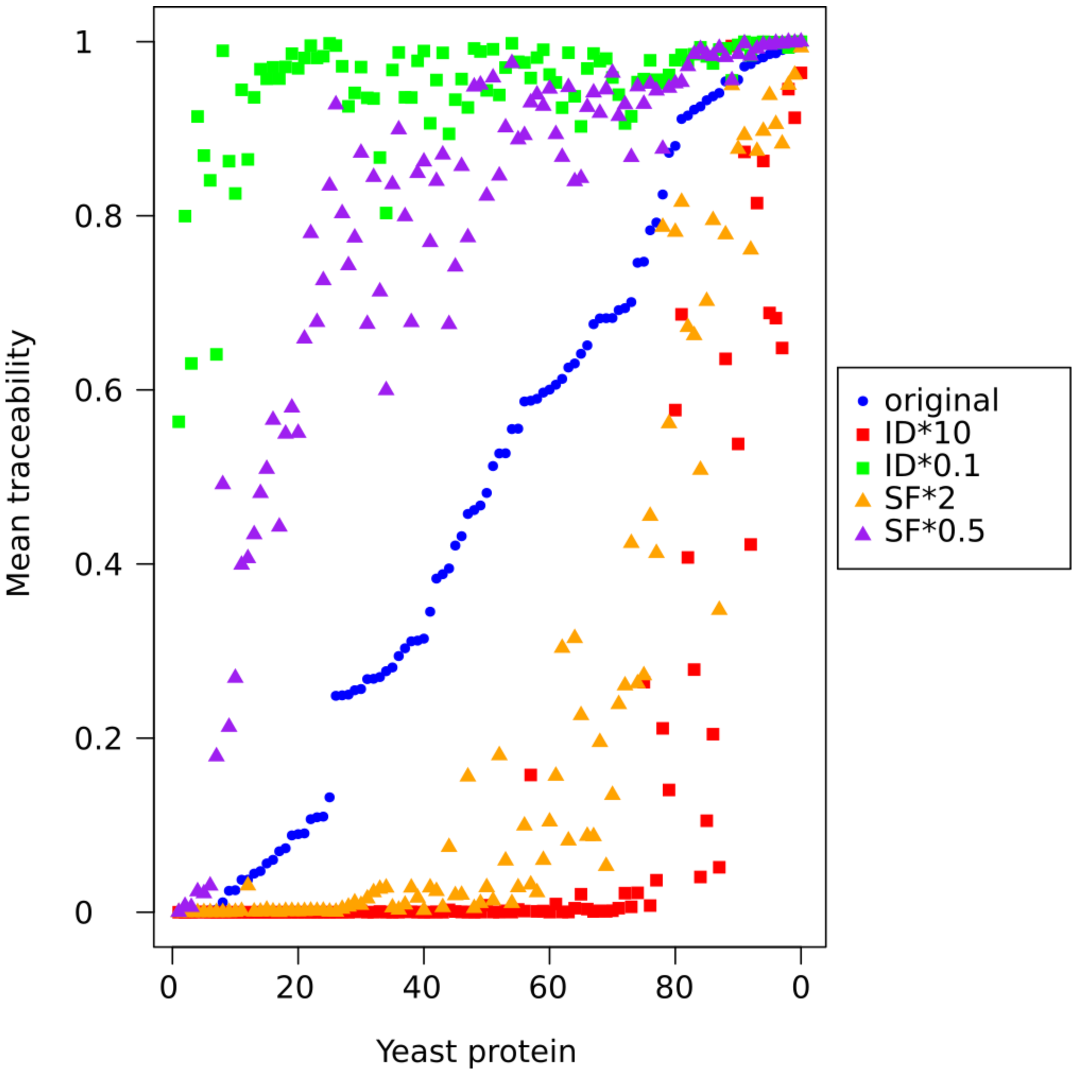

Figure S7 | Effect of scaling factor and insertion/deletion rate variation on the traceability estimates. We plotted the mean traceability estimates across 232 species for 100 yeast proteins using the scaling factor (SF) and the indel rates (ID) as inferred from the training data (blue dots). We then assessed the effect on the traceability estimates when doubling or halving the scaling factor, and when increasing or decreasing the indels rates by a factor of 10, respectively. Note, that doubling or halving the indels rate had only very minor effect on protein traceabilities (not shown).

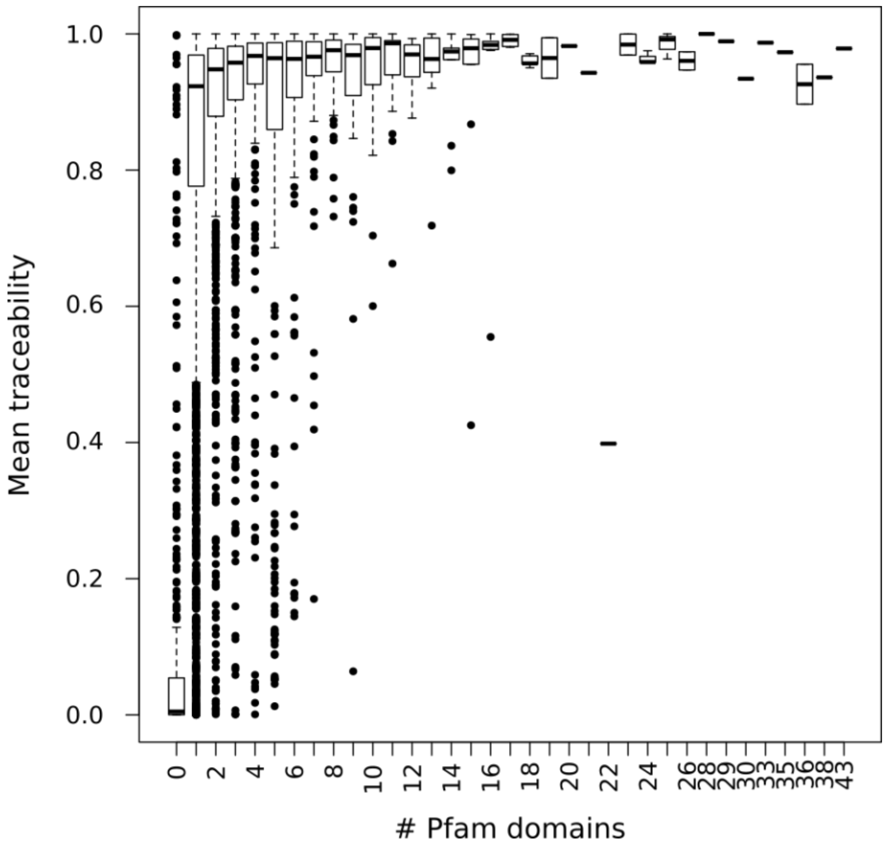

Figure S8 | Pfam domain content influences protein traceability. The box plot shows the distribution of mean traceabilities across 232 taxa for yeast protein harboring 0 up to 43 Pfam domains. The plot shows that Pfam domain content, in general are tightly correlated. However, individual proteins can have high traceabilities even without harboring any Pfam domain. In these cases, low rates for substitutions and indels drive the traceability. In turn, there is a considerable set of proteins with low mean traceabilities despite the presence of Pfam domains. In these cases, the constraints imposed by the pHMM representing the domain are not sufficient to drive local sequence conservation to an extent that it suffices for an ortholog detection over larger evolutionary distances.

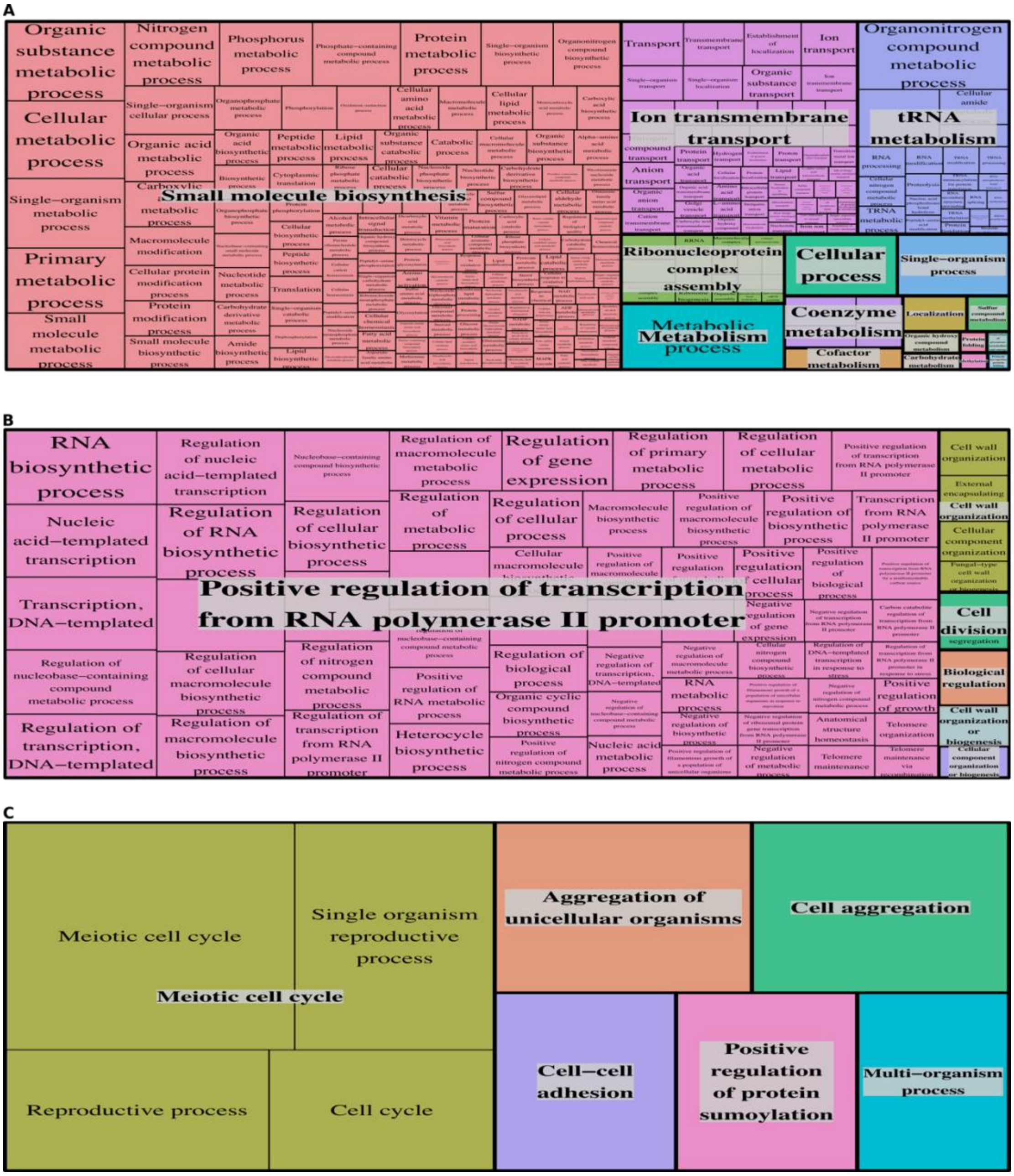

Figure S9 | Gene Ontology term enrichment (Biological Process) in protein sets with different traceabilities. A, GO enrichment in the high traceability bin (P_{det}(Ecoli) ≥ 0.75). B, GO enrichment in the intermediate traceability bin (0.25 ≤ P_{det}(Ecoli) < 0.75). C, GO enrichment in the low traceability bin (P_{det}(Ecoli) < 0.25). The tree maps were generated with REVIGO3. The underlying data is available from https://www.dropbox.com/sh/wdlvabtmvxpl1xp/AADHwIkAu3S1t0pOD3RMCX9aa?dl=0.

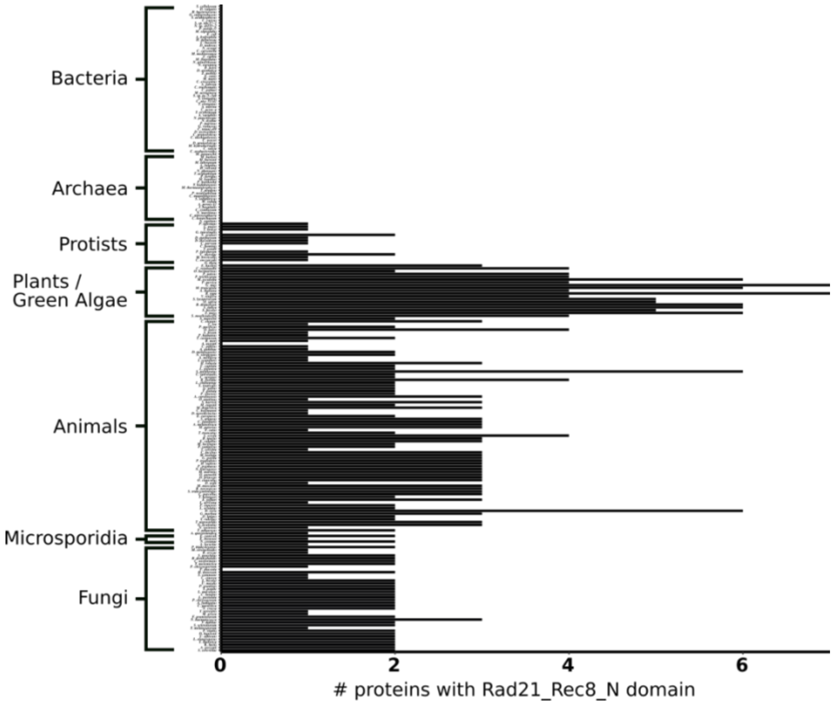

Figure S10 | Number of protein sequences harboring a Rad21_Rec8_N domain. Fungi, microsporidia and animals mostly possess two proteins with this domain. In Plants, four or more proteins are common, which may be a result of whole genome duplications that occurred on the plant lineage. The Rad21_Rec8 domain appears to be absent in prokaryotes.

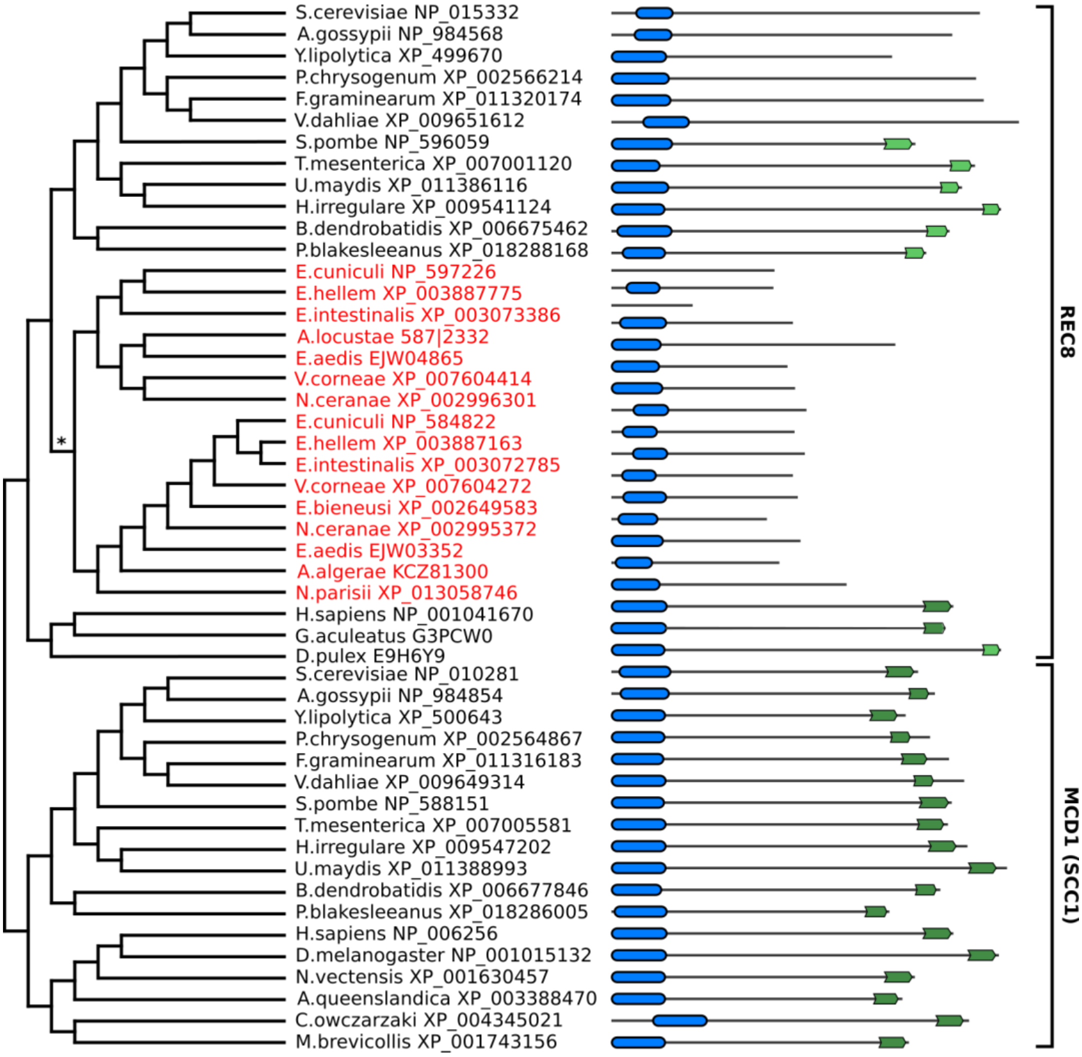

Figure S11 | Pfam Domain architecture evolution in the REC8 and MCD1 gene families. With the exception of two microsporidian proteins, all REC8 and MCD1 proteins share the presence of the Rad21_Rec8_N domain (PF04825) the N-terminus (blue domain). The Rad21_Rec8 domains (PF04824) at the C-terminus of the proteins shows a more diverse presence-absence pattern. All fungal and animal MCD1 (SCC1) proteins share the presence of this domain (shown in green). Within the REC8 clade, the presence of this domain is widespread, however it appears to have been lost twice independently. All microsporidian REC8 proteins (red clade) lack this domain. This indicates a domain loss in the last common ancestor of the microsporidia, and presumably prior to the gene duplication that gave rise to the two paralogous REC8 lineages within the microsporidia (indicated by the asterisk). With that, the microsporidian REC8 proteins resemble the domain architecture of the Sacharomycotina (*S. cerevisiae*, *A. gossypii*, *Y. lipolytica)* and of the Pezizomycotina

(*P. chrysogenum, F. graminearum*, *V. dahliae*), which appear to have lost the C-terminal Rad21_Rec8 domain in their last common ancestor.