Supplemental information

Bioinformatic multi-species DNA damage proteome analysis

For identification of the ortholog sequences related to DNA damage presented in this paper, we first started by using the list of human DNA damage proteins catalogued by the group of Rick Wood (http://sciencepark.mdanderson.org/labs/wood/DNA_Repair_Genes.html#Human%20DNA%20Repair%20Genes). We then run BLAST using all the human DNA damage proteins (129 proteins listed in Table 1 to 5 and Table S1 to S3) on the proteome of all the selected species used in this work (2). The version of the build for the proteome of every species is presented in supplementary Table 1. For two proteins to be considered ortholog, we are requiring that they both are in the top 5 reciprocal best BLAST hits of one another. For example, when BLASTing the human protein A on the proteome of *L.major,* if the best hit is protein B then we required that protein A is in the top 5 BLAST hits of protein B in the human proteome. This is what we defined as reciprocal top 5 best BLAST hits (4). We tried different values for the length of the top list to assign the reciprocal best hit and we obtained maximal concordance at top 5 with a maximum concordance of 80% for the know human yeast ortholog proteins. We repeated this entire reciprocal top 5 best BLAST hits using the yeast DNA damage proteome (94 proteins listed in Table 1 to 5 and Table S1 to S3). We performed exactly the same analysis as for the human DNA damage proteome. We had to merge the results from the two analyses on the basis of the list of common ortholog proteins between human and yeast. For the merged analysis we used a color code that is depicted in Fig.S2. A

white square means that neither the human nor the yeast analyses were able to retrieve an ortholog. A green square means that only the human analysis was able to retrieve and ortholog. A red square means that only the yeast analysis was able to retrieve an ortholog. When both the human and yeast analysis identified the same ortholog, this is depicted by a yellow square. If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square. For this study we used two different stringency values for the E-values of BLAST algorithm 0.001 and 0.05. We considered 0.001 as being stringent and 0.05 as being a more permissive approach to identify ortholog sequences. The analyses as well as the figures were generated using custom Python and R scripts.

The approach we used to retrieve ortholog is different from the majority of BLAST analyses performed in the papers that we will further refer to in the remaining of this review. The major difference reside in the reciprocal portion of the BLAST analysis in which we require the putative ortholog protein to also mapped back to the query protein in human (or yeast). This step is critical to insure the putative ortholog correspond really to the best human ortholog protein. It is thus highly plausible that our systematic approach will miss some of the previously ortholog proteins identified by the community. We also mentioned we used two different stringent E-value cutoffs and this could also explain that our systematic analysis will miss previously identified ortholog due to the use of different BLAST E-values.

The phylogenetic analysis of the RAD51 paralogs was done in two steps. The first step was to obtained and aligned the sequences of all the putative paralog sequences for human, *S cerevisiae*, *S.pombe* and the Trypanosamatids. We aligned the sequences using MAFFT and generated a phylogenetic tree from the alignement using Neighbor-joining with the JTT model and 500 bootstrap resampling (3). The database TriTrypDB-4.1 (1) was used for *L.infantum* (8241 proteins), *L.major Friedlin* (8412), *L.braziliensis* (8357), *T.bruceiTreu927* (9826), *T.congolense* (13459), *T.cruzi Esmeraldo-Like* (10342), *T.vivax* (11885) and 03-Feb-2011-downloads.yeastgenome.org/sequence/S288C_reference/orf_protein/ for *S.cerevisiae* (5887), 19-Mar-2012 - ftp.sanger.ac.uk/pub/yeast/pombe/Protein_data/ for *S.pombe* (5143) and 15-Oct-2012 - ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/ for human (34677).

Supplemental References

1. **Aslett, M., C. Aurrecoechea, M. Berriman, J. Brestelli, B. P. Brunk, M. Carrington, D. P. Depledge, S. Fischer, B. Gajria, X. Gao, M. J. Gardner, A. Gingle, G. Grant, O. S. Harb, M. Heiges, C. Hertz-Fowler, R. Houston, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, F. J. Logan, J. A. Miller, S. Mitra, P. J. Myler, V. Nayak, C. Pennington, I. Phan, D. F. Pinney, G. Ramasamy, M. B. Rogers, D. S. Roos, C. Ross, D. Sivam, D. F. Smith, G. Srinivasamoorthy, C. J. Stoeckert, Jr., S. Subramanian, R. Thibodeau, A. Tivey, C. Treatman, G. Velarde, and H. Wang.** 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res **38:**D457-462.
2. **Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421.
3. **Katoh, K., and D. M. Standley.** 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution **30:**772-780.
4. **Stuart, J. M., E. Segal, D. Koller, and S. K. Kim.** 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science **302:**249-255.

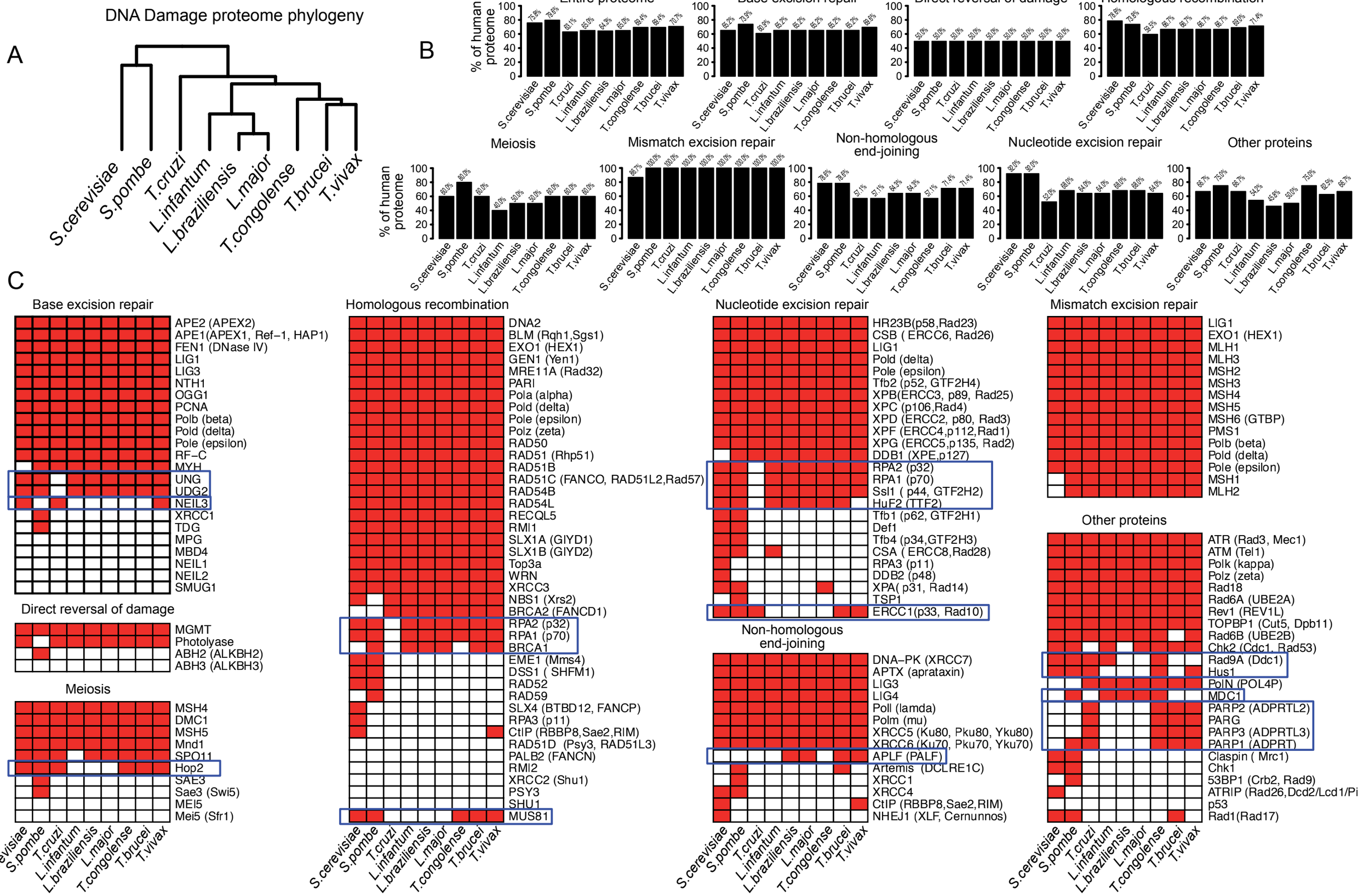

FIG S1 Summary for the bioinformatics analyses using BLAST E-value < 0.05. A) Complete hierarchical clustering using Euclidean distance of the list of retrieved DNA damage proteins in the different species analysed in the bioinformatics analysis. B) Barplots presenting the different percentage of proteins retrieved in the different DNA damage categories for all the species analyzed using E-value < 0.05. C) Heatmaps presenting the retrieved proteins (red square) for all the species in all the different DNA damage categories. Blue rectangles surrounded proteins that are only detected partially in the Trypanosamatids.

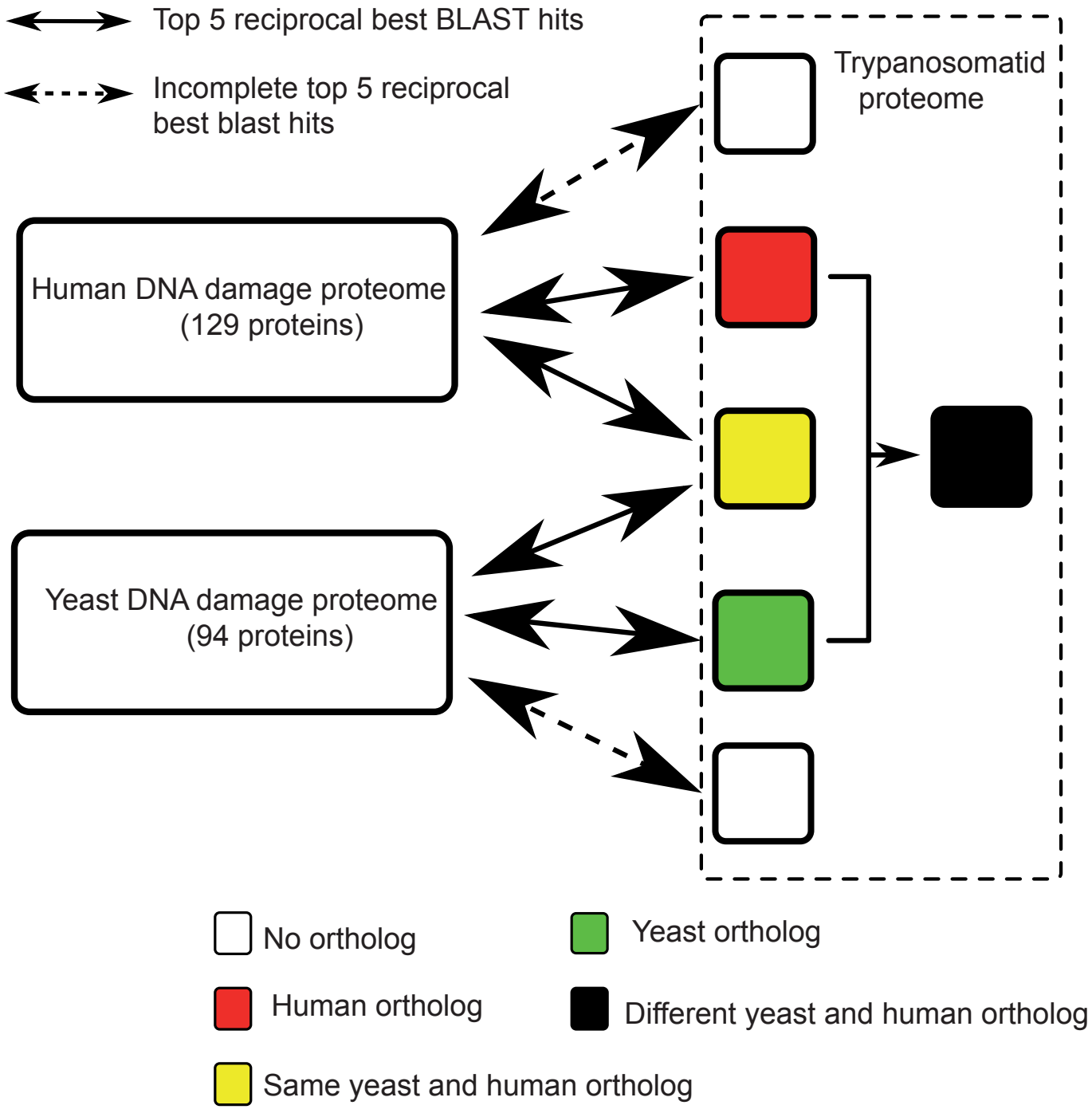

FIG S2 The top 5 reciprocal best BLAST hits approach used in this paper to retrieve bioinformatically the ortholog proteins in the other species. We used both the human and the yeast DNA damage proteome containing respectively 129 and 94 proteins. The arrows represent the reciprocal best hits approach and the different squares represent the different outcomes following the yeast and human analysis.

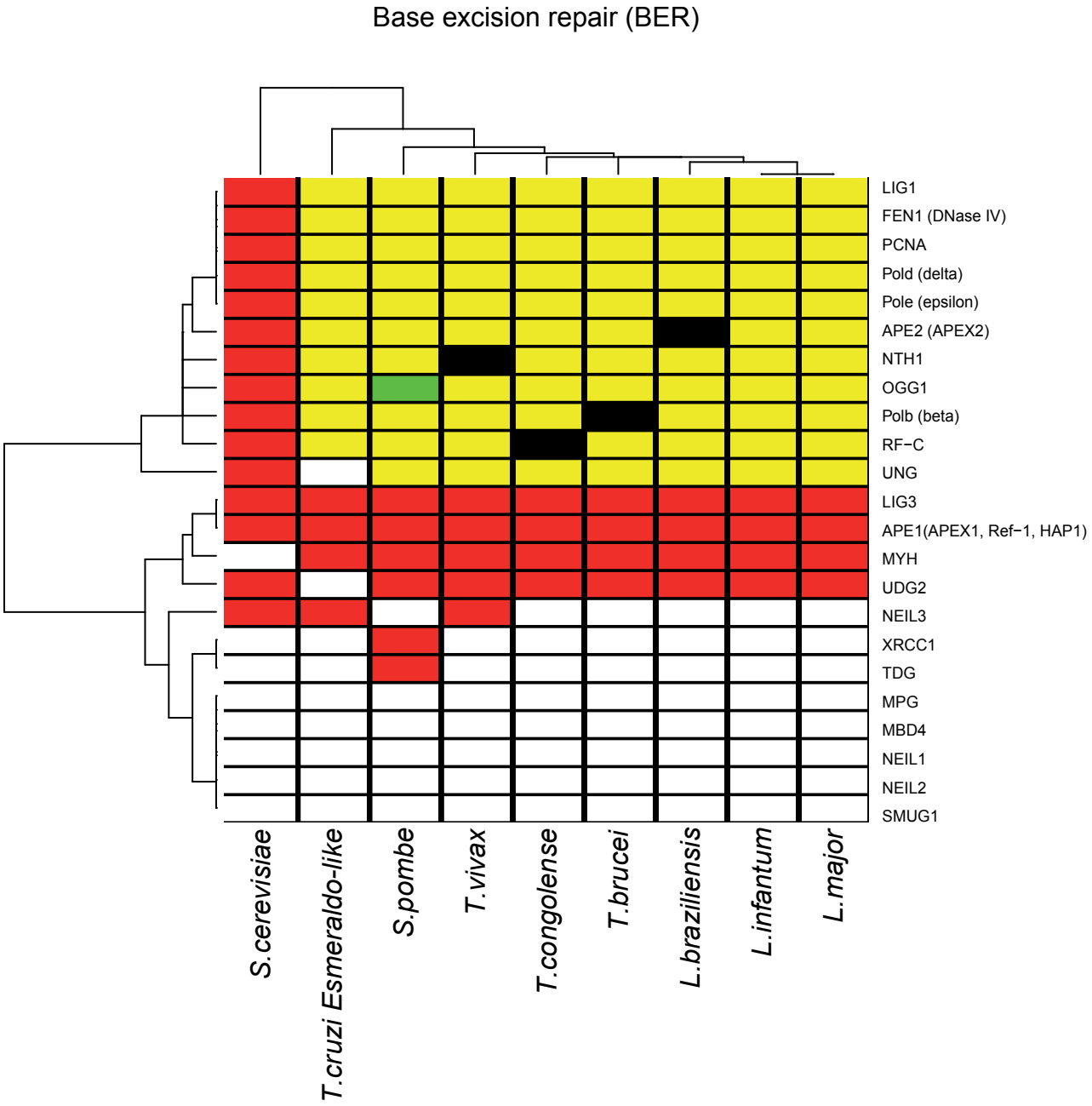

FIG S3 Base excision repair homologs in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

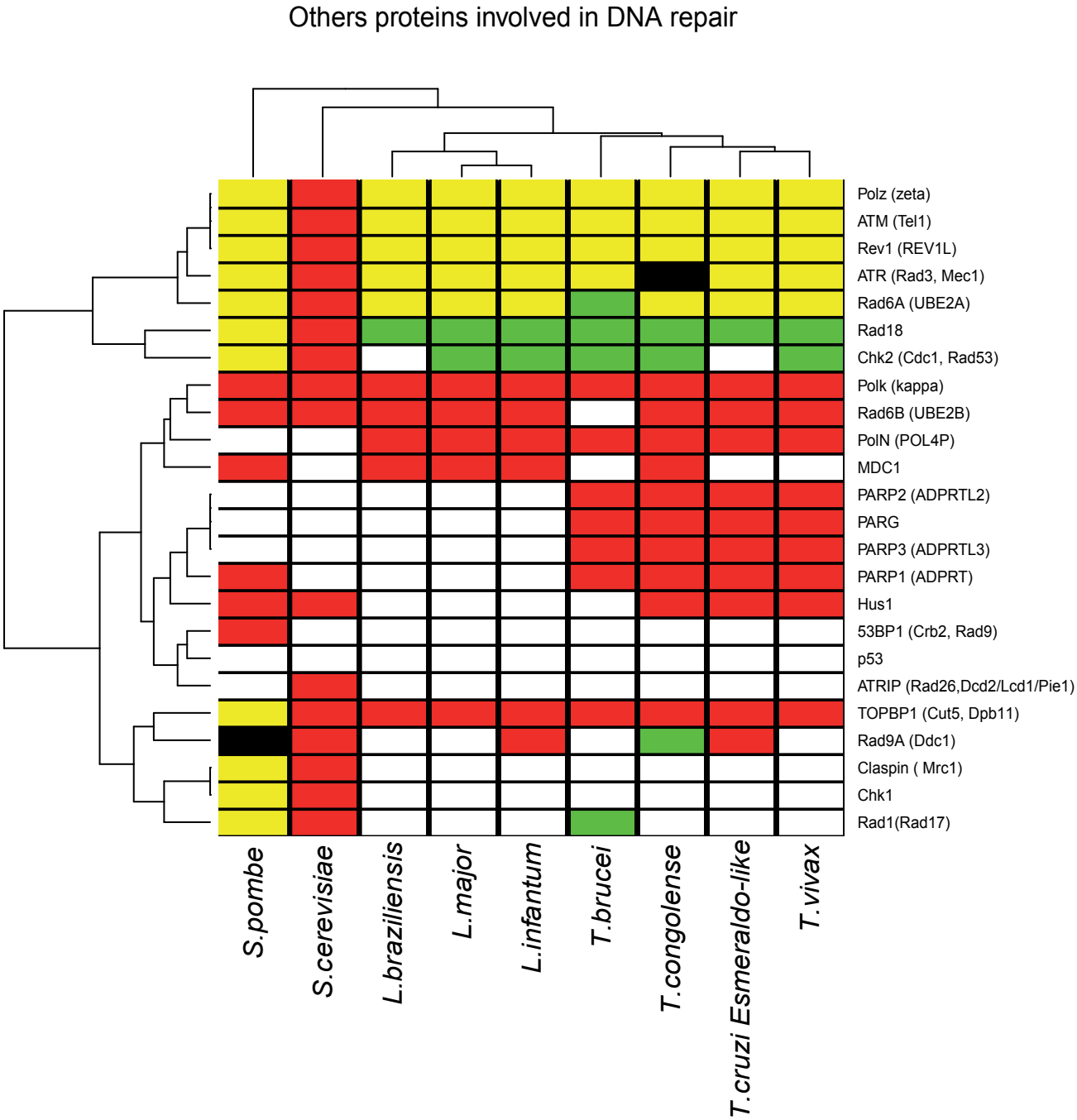

FIG S4 Other proteins involved in DNA repair in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

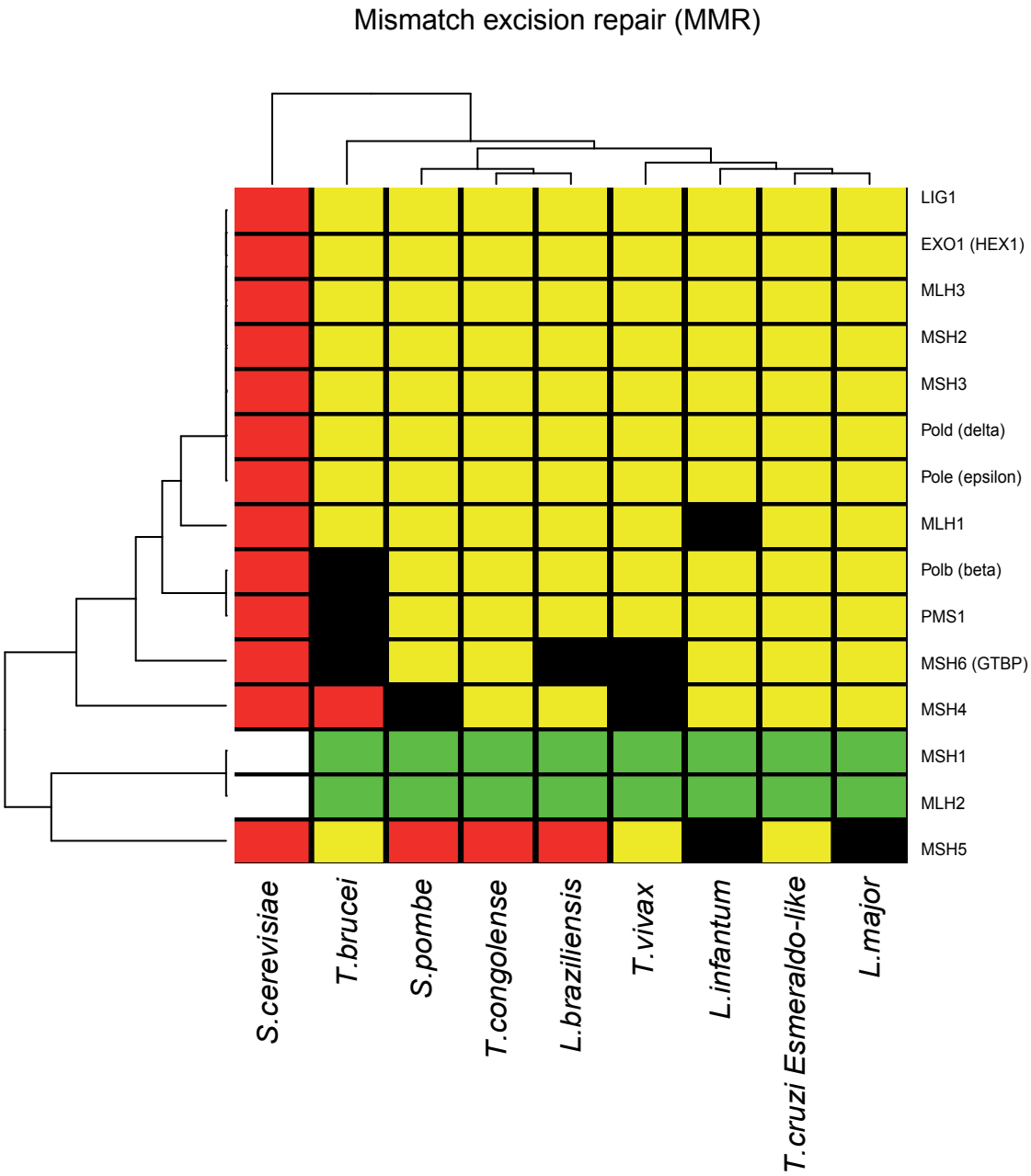

FIG S5 Mismatch repair homologs in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

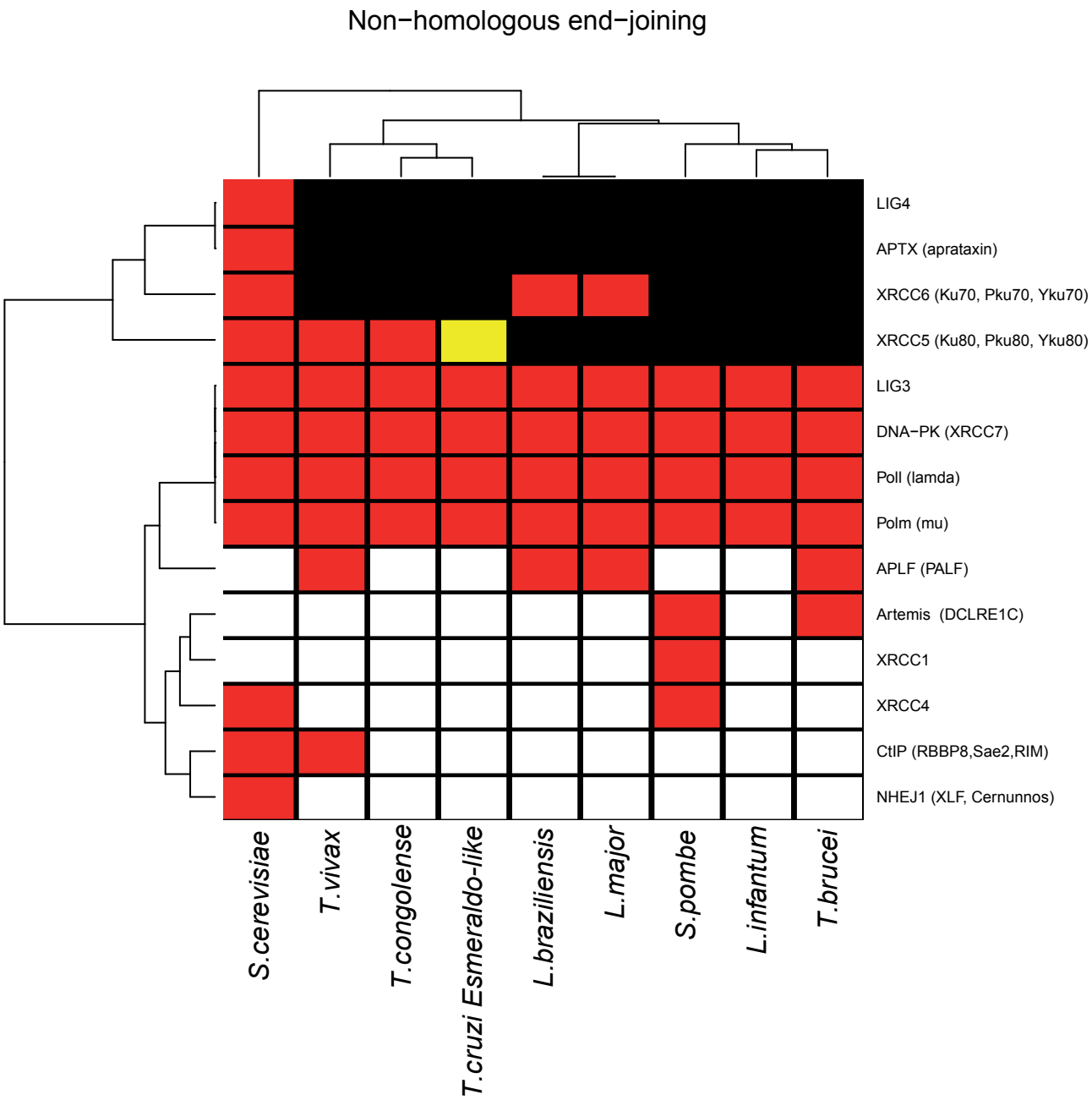

FIG S6 Non-homologous end joining homologs in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

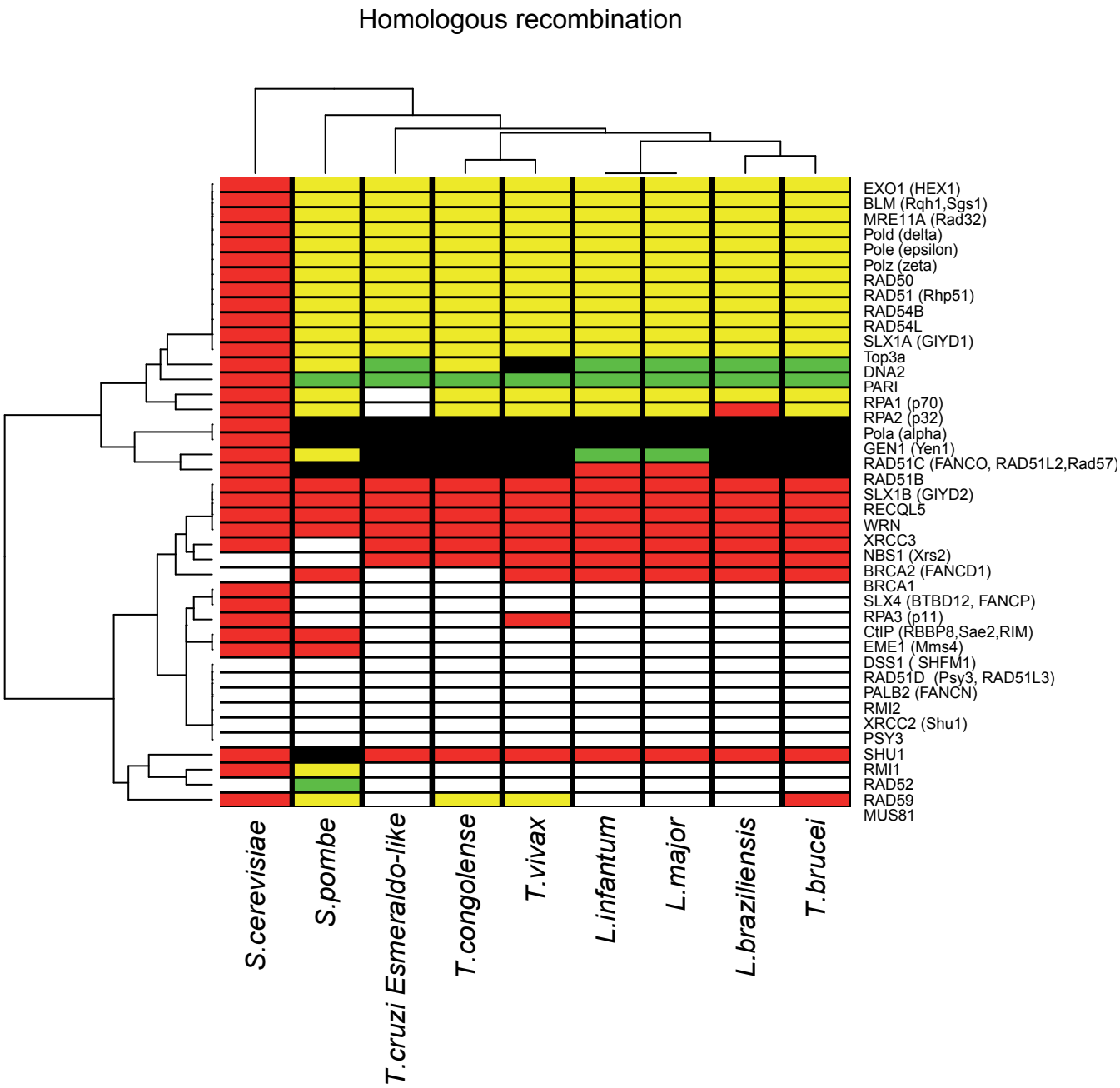

FIG S7A Homologous recombination proteins in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

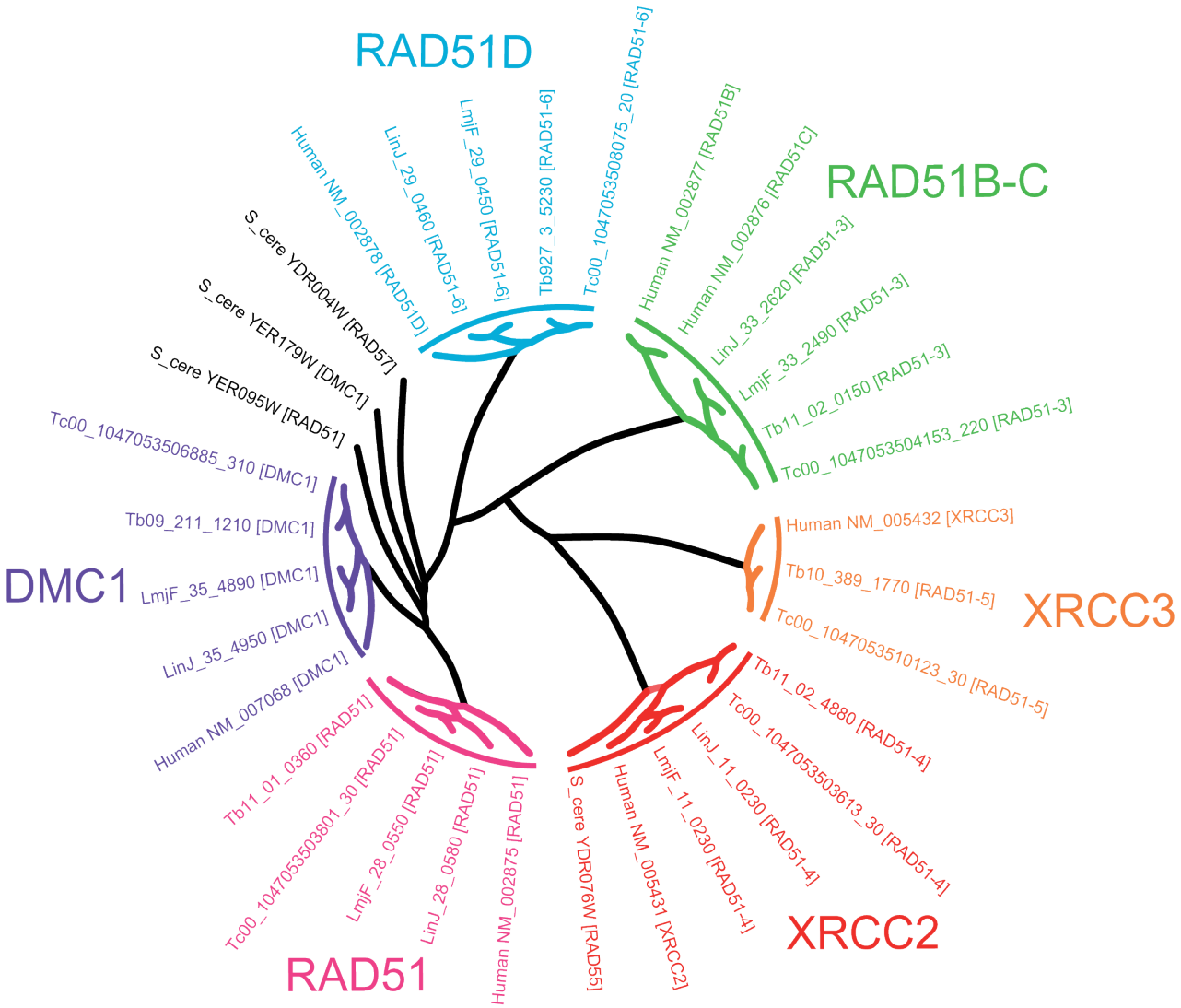

FIG S7B The phylogenetic analysis of the RAD51 family proteins in human, *S.cerevisiae, Leishmania infantum, Leishmania major, Trypanosoma brucei,* and *Trypanosoma cruzi.*

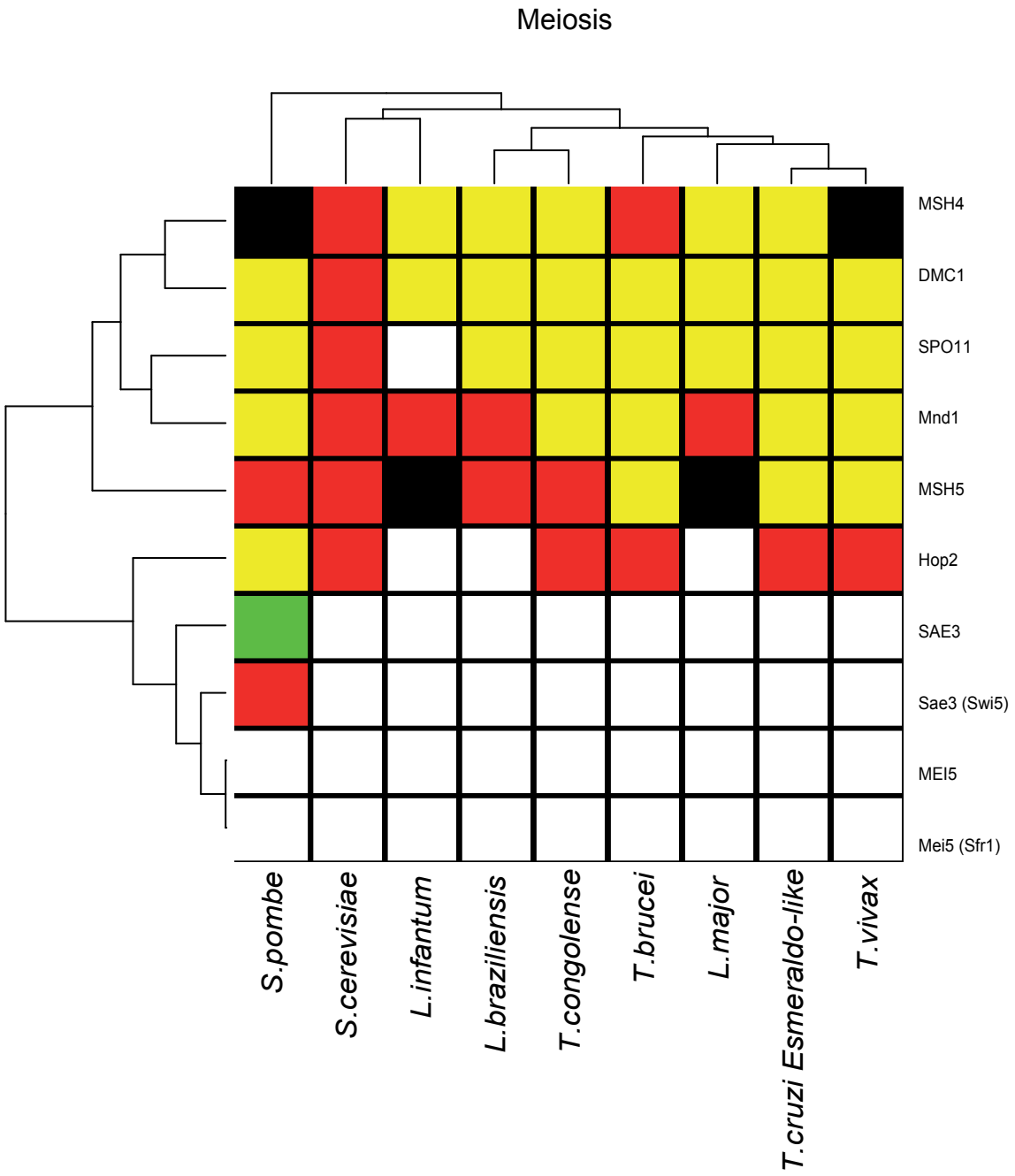

FIG S8 Putative meiotic proteins in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

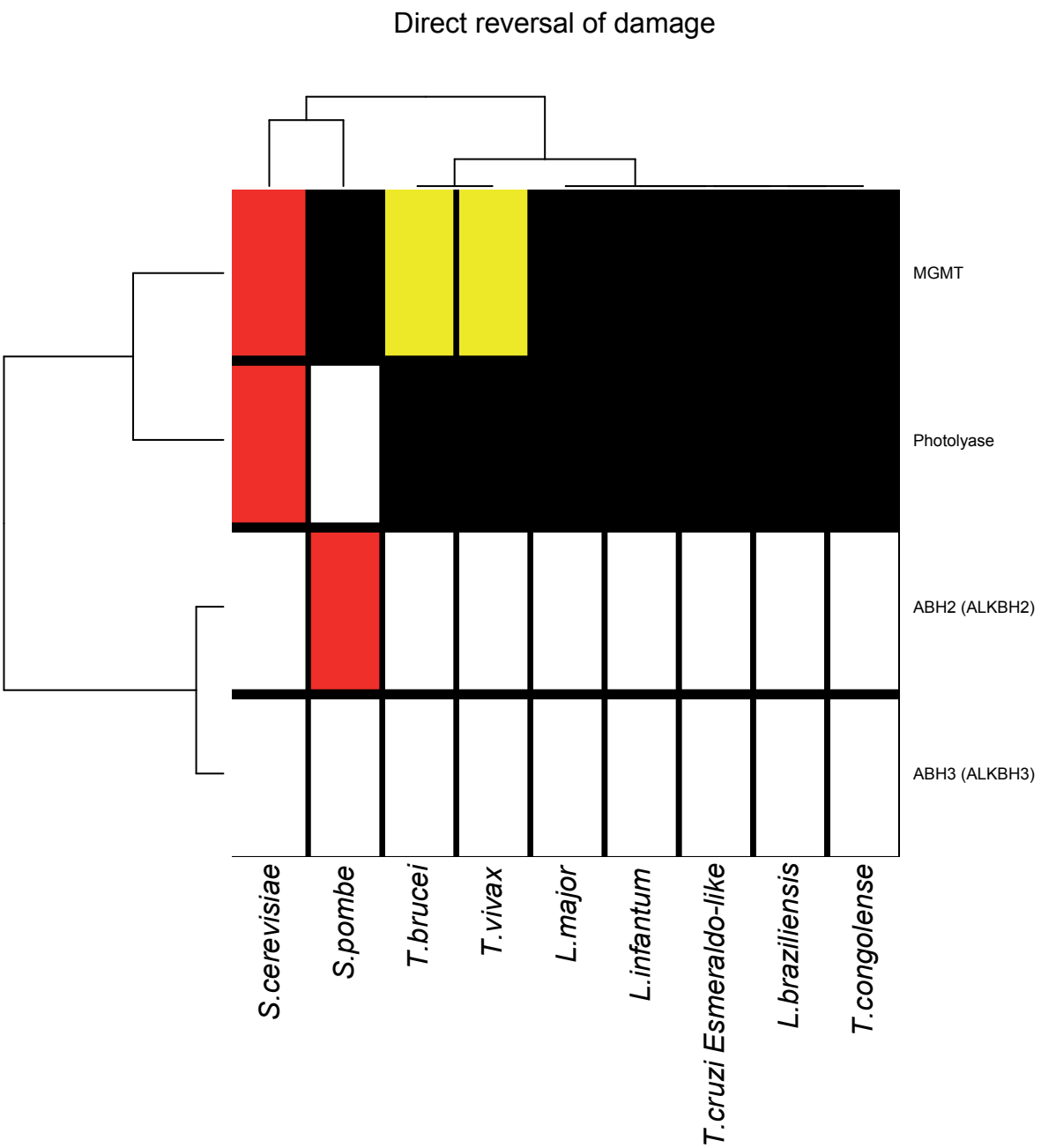

FIG S9 Direct reversal of damage proteins in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

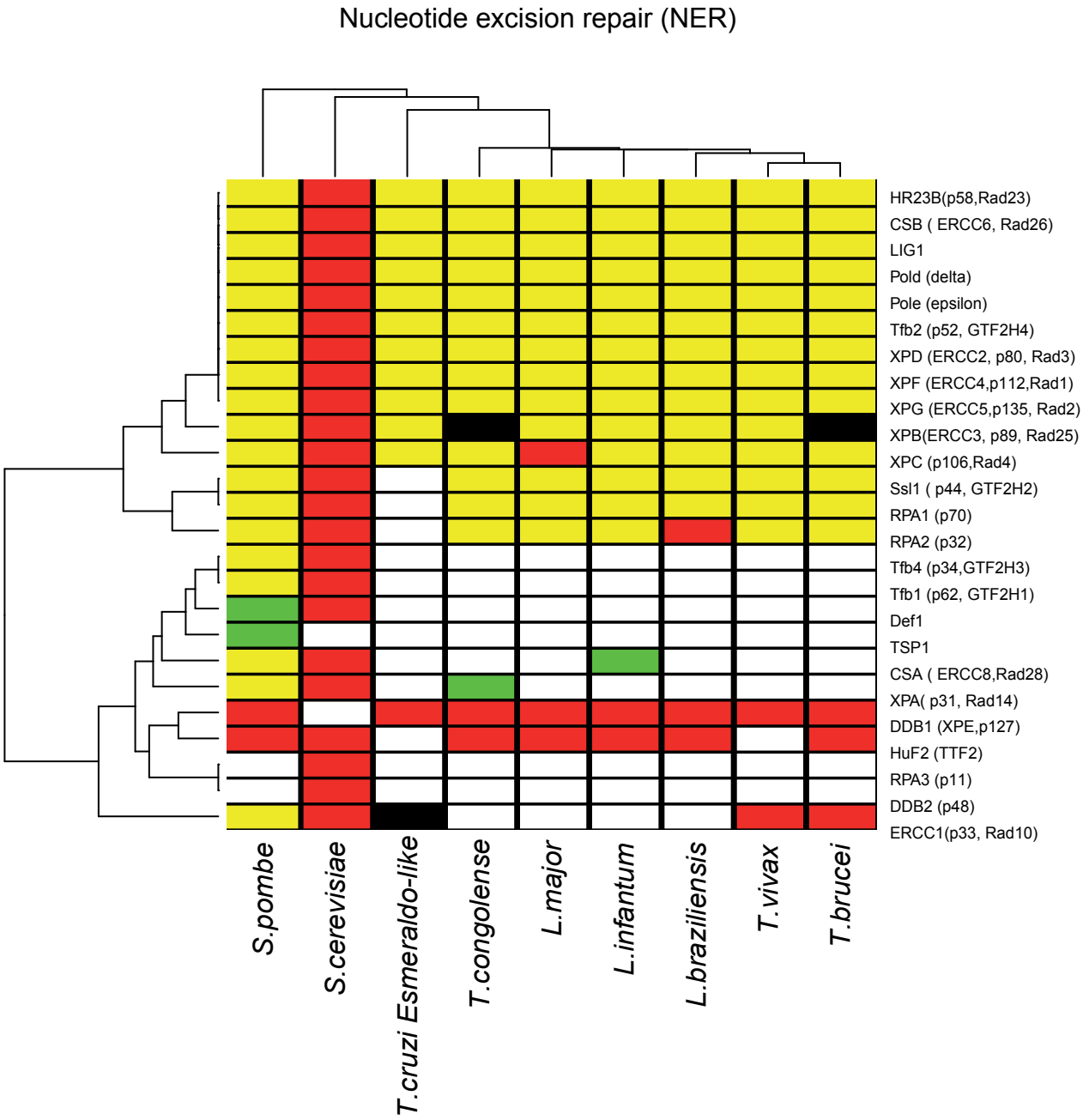

FIG S10 Nucleotide excision repair proteins in Trypanosomatids. White boxes represent no mapping; red represents human mapping only; green represent budding yeast mapping only; yellow represent both human and yeast mapping; If the yeast and human analysis did not retrieve the same ortholog proteins this is depicted by a black square.

Table S1 List of Trypanosomatids genes involved in direct repair included functions and ID.

Name (other name)	Fonction	Gene ID						
		Human	S.cerevisiae	S.pombe	L.infantum	L.major	T.brucei	T.cruzi
MGMT	O6-methylguanine-methyltransferase	NM_002412	YDL200C	SPAC1250.04c(B)	LinJ.26.1800(B)	LmjF.26.1800 (B)	Tb09.160.1490 (Y)	Tc00.1047053508347.20 (B)
ABH2 (ALKBH2)	Dioxygenase able to revert N-alkyl adducts	NM_001001655	X	SPAP8A3.02C (H)	X	X	X	X
ABH3 (ALKBH3)	Dioxygenase able to revert N-alkyl adducts	NM_139178	X	X	X	X	X	X
Photolyase	Photoinduce repair of cyclobutane pyrimidine dimers	NM_004075	YOR386W	X	LinJ.33.0480 (B)	LmjF.33.0470 (B)	Tb927.10.11100 (B)	Tc00.1047053511127.260 (B)

(H) represents found from human homolog only
(Y) represents found from yeast homolog only
(B) means both found from human and yeast homologs
(X) means no homolog

Table S2 List of Trypanosomatids genes involved in nucleotide excision repair (NER) included functions and ID.

Name (other name)	Fonction	Gene ID						
		Human	S.cerevisiae	S.pombe	L.infantum	L.major	T.brucei	T.cruzi
XPA (p31, Rad14)	Xeroderma pigmentosum A: damage recognition	NM_000380	YMR201C	SPBC649.03(B)	X	X	X	X
XPC (p106, Rad4)	Xeroderma pigmentosum C: recruitment of TFIIH	NM_004628	YER162C	SPCC4G3.10c(B)	LinJ.35.3500(B)	LmjF.35.3450(H)	Tb09.211.3040(B)	Tc00.1047053507011.140(B)
HR23B (p58, Rad23)	In complex with XPC, recruitment of TFIIH	NM_005053	YEL037C	SPBC2D10.12(B)	LinJ.30.3350(B)	LmjF.30.3300(B)	Tb927.6.4650(B)	Tc00.1047053511731.10(B)
TFIIH-XPB (ERCC3/p89/Rad25/Ssl2)	Transcription factor : DNA unwinding	NM_000122	YIL143C	SPAC17A5.06(B)	LinJ.32.4070(B)	LmjF.32.3920(B)	Tb927.3.5100(H) Tb11.01.7950(Y)	Tc00.1047053511527.20(B)
TFIIH-XPD (ERCC2/p80/Rad3)	Transcription factor : DNA unwinding	NM_000400	YER171W	SPAC1D4.12(B)	LinJ.24.2370(B)	LmjF.24.2280(B)	Tb927.8.5980(B)	Tc00.1047053511075.30(B)
TFIIH-Tfb1 (p62, GTF2H1)	Transcription factor : DNA unwinding	NM_005316	YDR311W	SPAC16E8.11c(B)	X	X	Tb11.01.1200	X
TFIIH-Tfb2 (p52,GTF2H4)	Transcription factor : DNA unwinding	NM_001517	YPL122C	SPBC13G1.13(B)	LinJ.36.0860(B)	LmjF.36.0800(B)	Tb927.10.5210(B)	Tc00.1047053510297.80(B)
TFIIH-Ssl1 (p44, GTF2H2)	Transcription factor : DNA unwinding	NM_001515	YLR005W	SPCC1682.07(B)	LinJ.24.1750(B)	LmjF.24.1680(B)	Tb927.8.6540(B)	X
TFIIH-TSP1	Trypanosomatid-specific compenent of TFIIH	X	X	X	LinJ.20.0470	LmjF.20.0400	Tb927.1.1080	Tc00.1047053511423.40
TFIIH-TSP2	Trypanosomatid-specific compenent of TFIIH	X	X	X	LinJ.32.0910	LmjF.32.0860	Tb11.01.5700	TcIL3000.11.14400
TFIIH-Tfb4 (p34, GTF2H3)	Transcription factor : DNA unwinding	NM_001516	YPR056W	SPBC30B4.07c(B)	X	X	Tb11.01.7730	X
TFIIH-Tfb5	Trypanosomatid-specific compenent of TFIIH	X	X	X	X	X	Tb10.61.2600	X
XPG (ERCC5, p135, Rad2)	Xeroderma pigmentosum G : 3'-incision of the lesion	NM_000123	YGR258C	SPBC3E7.08c(B)	LinJ.35.3640(B)	LmjF.35.3590(B)	Tb09.211.2870(B)	Tc00.1047053507009.120(B)
ERCC1 (p33, Rad10)	Structure-specific DNA repair endonuclease for the 5'-incision	NM_001983	YML095C	SPBC4F6.15c (B)	X	X	Tb927.7.2060 (H)	Tc00.1047053510165.20 (H)
XPF (ERCC4,p112,Rad1)	Xeroderma pigmentosum F : 5'-incision of the lesion	NM_005236	YPL022W	SPCC970.01(B)	LinJ.08.0150(B)	LmjF.08.0140(B)	Tb927.5.3670(B)	Tc00.1047053509779.10(B)
RPA1 (p70)	Replication protein A 1 : ssDNA binding heterotrimer	NM_002945	YAR007C	SPBC660.13c(B)	LinJ.28.1940(B)	LmjF.28.1820(B)	Tb11.01.0870(B)	X
RPA2 (p32)	Replication protein A 2 : ssDNA binding heterotrimer	NM_002946	YNL312W	SPCC1753.01c (B)	LinJ.15.0310 (H)	LmjF15.0270(H)	Tb927.5.1700 (B)	X
RPA3 (p11)	Replication protein A 3 : ssDNA binding heterotrimer	NM_002947	YJL173C	X	X	X	X	X
Polε	DNA polymerase filling the gap	NM_006231	YNL262W	SPBC25H2.13c(B)	LinJ.35.4430(B)	LmjF.35.4360(B)	Tb09.211.1820(B)	Tc00.1047053506147.180(B)
Polδ	DNA polymerase filling the gap	NM_002691	YDL102W	SPBC336.04(B)	LinJ.33.1790(B)	LmjF.33.1690(B)	Tb927.2.1800(B)	Tc00.1047053510259.6(B)
LIG I (CDC9)	DNA ligase I : ligation	NM_000234	YDL164C	SPAC20G8.01(B)	LinJ.30.3490(B)	LmjF.30.3440(B)	Tb927.6.4780(B)	Tc00.1047053506945.80(B)
DDB1 (XPE, p127)	Damaged DNA binding 1 : heterodimer involved in damage recognition and stimulate excision	NM_001923	X	SPAC17H9.10c(H)	LinJ.30.3770(H)	LmjF.30.3710(H)	Tb927.6.5110(H)	Tc00.1047053509165.49(H)
DDB2 (p48)	Damaged DNA binding 2 : heterodimer involved in damage recognition and stimulate excision	NM_000107	YDL156W	X	X	X	X	X
CSA (ERCC8, Rad28)	Cockayne's syndrome A:E3 ubiquitin-protein ligase complex	NM_000082	YDR030C	SPBC577.09(B)	LinJ.23.1400(Y)	X	X	X
CSB (ERCC6, Rad26)	Cockayne's syndrome B : Responsible for ubiquitination and proteolysis of RNAPII	NM_000124	YJR035W	SPCP25A2.02c(B)	LinJ.14.0900(B)	LmjF.14.0840(B)	Tb927.7.4080(B)	Tc00.1047053508675.20(B)
Def1	Responsible for ubiquitination and proteolysis of RNAPII	NM_004084	YKL054C	SPBC354.10(Y)	X	X	X	X
HuF2 (TTF2)	Human homolog of factor 2 : release RNAPI and II stalled in TCR	NM_003594	YBR114W	SPBC582.10c(H)	LinJ.28.0810(H)	LmjF.28.0760(H)	Tb11.010530(H)	X

(H) represents found from human homolog only
(Y) represents found from yeast homolog only
(B) means both found from human and yeast homologs
(X) means no homolog

Table S3 List of others Trypanosomatids genes involved in DNA repair included functions and ID.

.

Name (other name)	Fonction	Gene ID						
		Human	S.cerevisiae	S.pombe	L.infantum	L.major	T.brucei	T.cruzi
53BP1(Crb2, Rad9)	Tp53 binding protein 1 : DNA damage sensor, limit resection	NM_001141980	X	SPBC342.05 (H)	X	X	X	X
ATM (Tel1)	Ataxia-telangiectesia-mutated kinase: checkpoint-specific damage sensor	NM_000051	YBL088C	SPCC23B6.03c (B)	LinJ.02.0100 (B)	LmjF.02.0120 (B)	Tb927.2.2260 (B)	Tc00.1047053506533.34 (B)
ATR (Rad3, Mec1)	ATM and Rad3-related kinase : checkpoint-specific damage sensor	NM_001184	YBR136W	SPBC216.05 (B)	LinJ.32.1520 (B)	LmjF.32.1460 (B)	Tb11.01.6300 (B)	Tc00.1047053506223.120 (B)
ATRIP (Rad26,Dcd2/Lcd1/Pie1)	ATR interacting protein : recruit ATR	NM_130384	YDR499W	X	X	X	X	X
Chk1	Checkpoint kinase 1 : Effector kinase, an ATR substrat	NM_001274	YBR274W	SPCC1259.13 (B)	X	X	X	X
Chk2 (Cds1, Rad53)	Checkpoint kinase 2 : Effector kinase, an ATM substrat	NM_007194	YPL153C	SPCC18B5.11c (B)	LinJ.17.0070 (Y)	LmjF.17.0060 (Y)	Tb927.7.6220 (Y)	X
Claspin (Mrc1)	Sensor which monitors the integrity of DNA replication fork	NM_022111	YCL061C	SPAC694.06c (B)	X	X	X	X
Hus1 (Mec3)	Hydroxyurea sensitive 1 : heterotrimeric 9-1-1 checkpoint complex, subunits of PCNA-like sensor of damaged DNA	NM_004507	YLR288	SPAC20G4.04c (H) SPBC27B12.05 (Y)	X	X	X	Tc00.1047053466823.10 (H)
MDC1 (NFBD1)	Mediator of DNA-damage checkpoint 1	NM_014641	X	SPBC582.05c (H)	LinJ.34.4070 (H)	LmjF.34.4240 (H)	X	X
p53	Tumor suppressor protein 53: regulation of the cell cycle	NM_000546	X	X	X	X	X	X
PARP-1	Poly (ADP) ribose polymerase-1 : sensor of DNA strand break	NM_001618	X	SPBC2A9.07c (H)	LinJ.25.0770	LmjF.25.0740	Tb927.5.3050 (H)	Tc00.1047053510173.90 (H)
PARP-2	Poly (ADP) ribose polymerase-2 : sensor of DNA strand break	NM_005484	X	X	X	X	X	X
PARP-3	Poly (ADP) ribose polymerase-3 : sensor of DNA strand break	NM_001003931	X	X	X	X	X	X
PARG	Poly(ADP-ribose) glycohydrolase which rapicly degrades PARP	NM_003631	X	X	X	X	Tb09.211.3760 (H)	Tc00.1047053507013.24 (H)
Polk	Translesion synthesis polymerase kappa	NM_016218	YOR346W	SPCC553.07c (H)	LinJ.28.1540 (H)	LmjF.28.1420 (H)	Tb11.01.0040 (H)	Tc00.1047053503755.30 (H)
Polη	Translesion synthesis polymerase	NM_181808	X	X	LinJ.34.1370 (H)	LmjF.34.1260 (H)	Tb927.4.2950 (H)	Tc00.1047053506265.30 (H)
Polζ	Translesion synthesis polymerase zeta consisting of REV3 and REV 7 subunits	NM_002912	YPL167C	SPAC688.10 (B)	LinJ.23.1590 (B)	LmjF.23.1330 (B)	Tb927.8.3290 (B)	Tc00.1047053509769.130 (B)
Rad1(Rad17)	Heterotrimeric 9-1-1 checkpoint complex : subunits of PCNA-like sensor of damaged DNA	NM_002853	YOR368W	SPAC1952.07 (B)	X	X	Tb927.1.1060 (Y)	X
Rad18	PCNA monoubiquination in complex with Rad6 at replication fork stalled	NM_020165	YCR066W	SPBC1734.06 (B)	LinJ.17.0340 (Y)	LmjF.17.0290 (Y)	Tb927.7.6370 (Y)	Tc00.1047053504035.130 (Y)
Rad6A(UBE2A)	PCNA monoubiquination in complex with Rad18 at replication fork stalled	NM_003336	YGL058W	SPAC18B11.07c (B)	LinJ.22.0480 (B)	LmjF.22.0610 (B)	Tb927.8.6090 (Y)	Tc00.1047053506859.10 (B)
Rad6B(UBE2B)	PCNA monoubiquination in complex with Rad18 at replication fork stalled	NM_003337	YGL058W	SPAC18B11.07c (H)	LinJ.22.0480 (H)	LmjF.22.0610 (H)	X	Tc00.1047053506859.10 (H)
Rad9 (Ddc1)	Heterotrimeric 9-1-1 checkpoint complex : subunits of PCNA-like sensor of damaged DNA	NM_004584	YDR217C	SPAC664.07c (H) SPBC342.05 (Y)	LinJ.15.1040 (H)	X	X	Tc00.1047053505843.30 (H)
Rev1	Translesion synthesis polymerase	NM_016316	YOR346W	SPBC1347.01c (B)	LinJ.36.0110 (B)	LmjF.36.0100 (B)	Tb927.10.4480 (B)	Tc00.1047053510963.10 (B)
TopBP1(Cut5, Dpb11)	Topoisomerase-binding protein 1: activation of ATR	NM_007027	YJL090C	SPAC23C4.18c (B)	LinJ.29.1910 (H)	LmjF.29.1790 (H)	Tb927.3.4350 (H)	Tc00.1047053509767.200 (H)

(H) represents found from human homolog only
(Y) represents found from yeast homolog only
(B) means both found from human and yeast homologs
(X) means no homolog