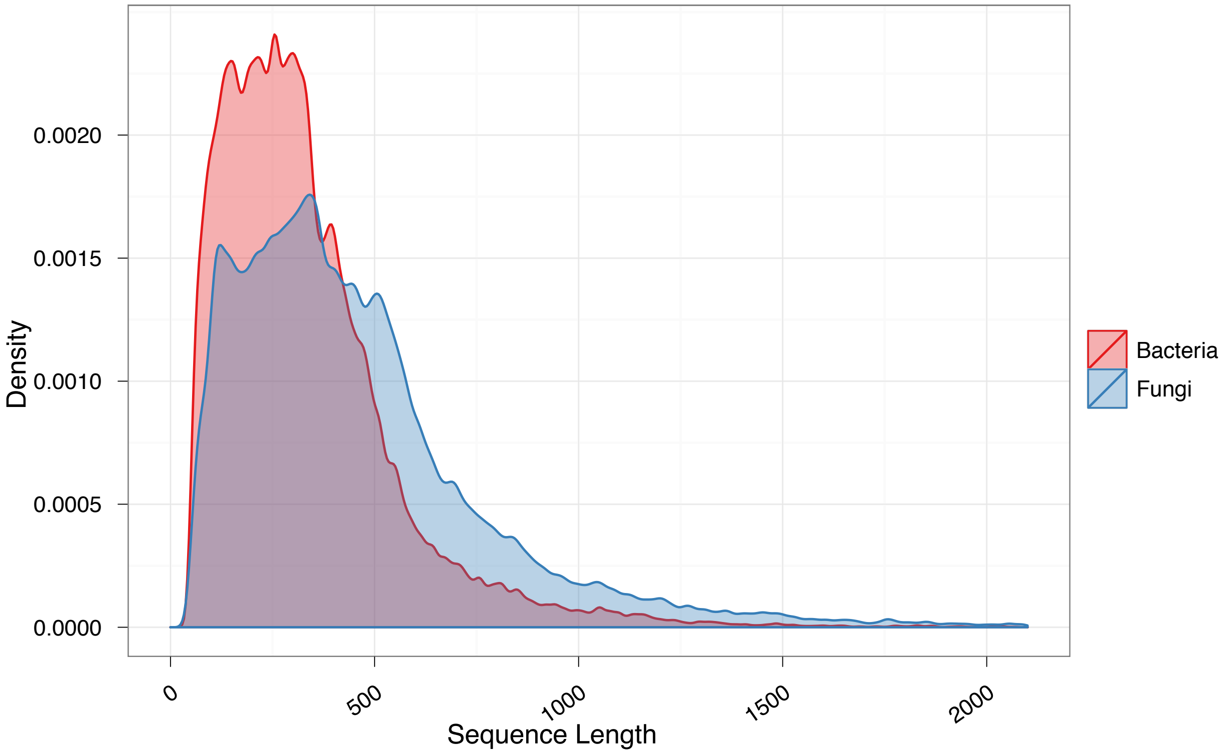
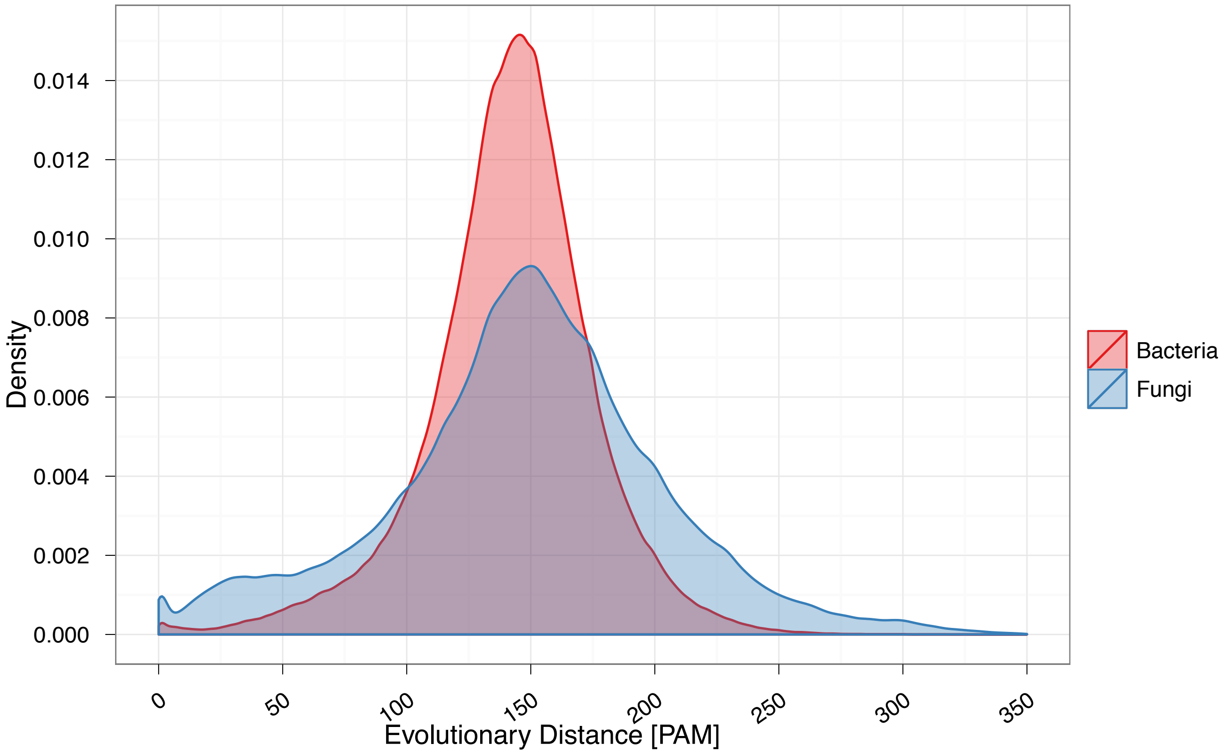
Supplementary materials for

Speeding up all-against-all protein comparisons while maintaining sensitivity by considering subsequence-level homology

Lucas D. Wittwer, Ivana Piližota et al.



Supplementary figure 1: **Distribution of sequence length in bacteria and fungi datasets (in number of amino acids).**

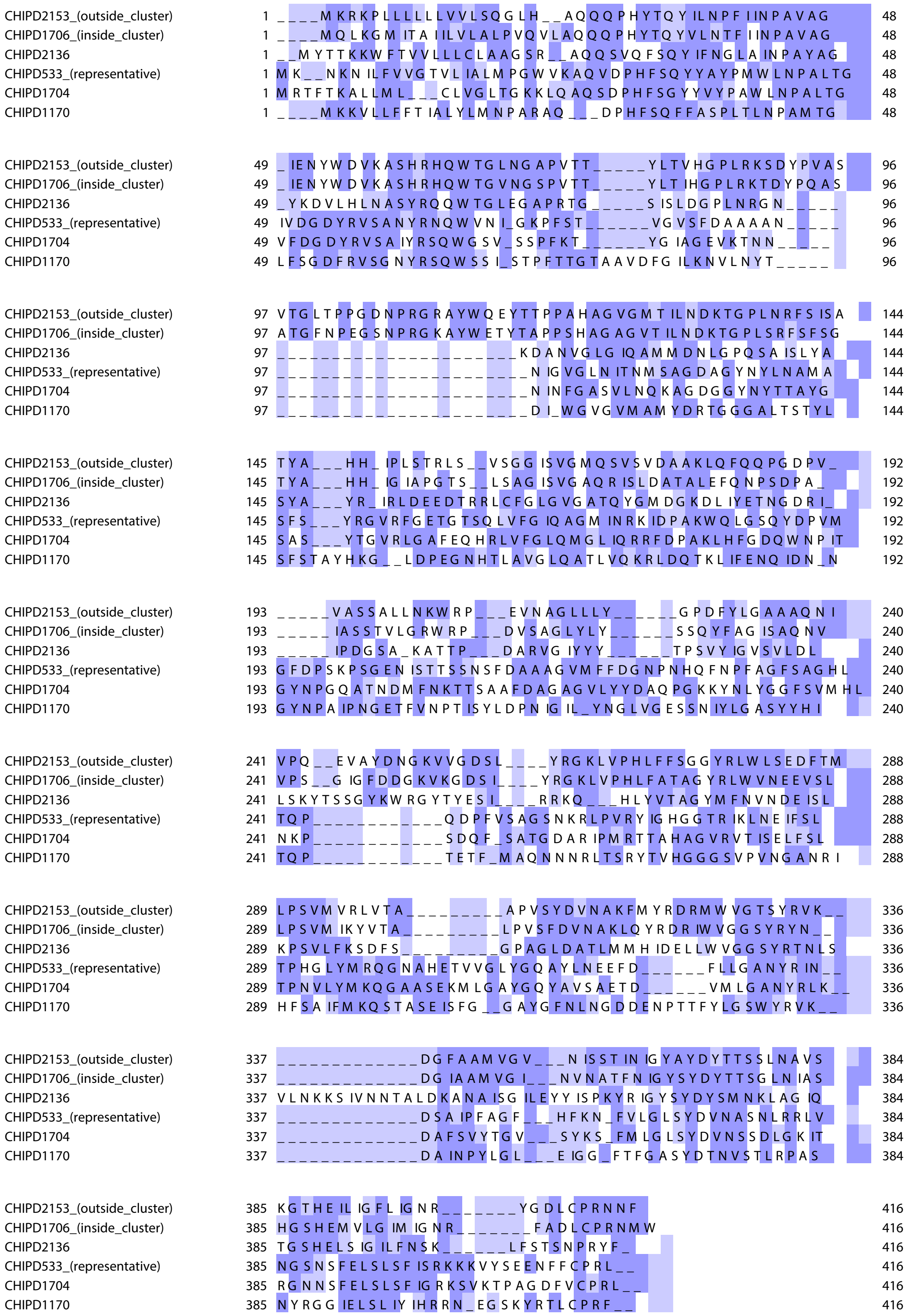


Supplementary figure 2: **Distribution of evolutionary distances among inferred homologous pairs (based on full all-against-all in bacteria and fungi datasets).**

Case studies of two missing homologs:

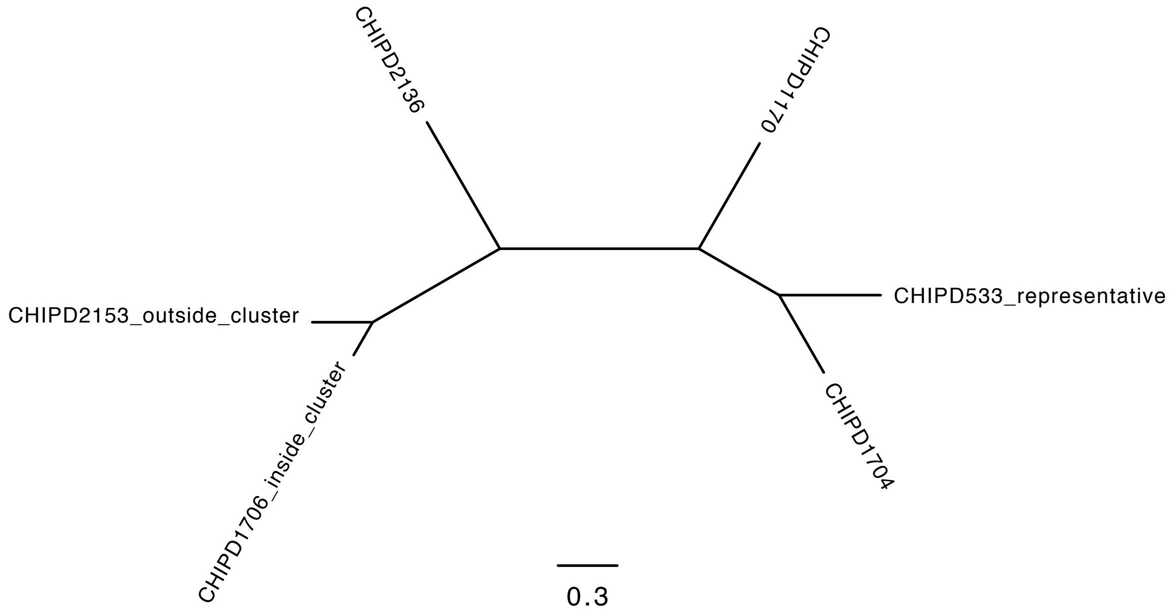
To illustrate the nature of missing homologs, we provide two detailed descriptions of high-scoring homologous pairs that are missed by the new approach (1 representative, subsequence homology). In both cases, one sequence is added to a cluster with an alignment score slightly above the threshold while the other is not added due to a score just below the threshold:

1) Bacteria: Sequences CHIPD1706 and CHIPD2153 (we use OMA IDs unless stated otherwise) have an alignment score of 2238.183 (estimated PAM distance: 40.3). CHIPD1706 is member of a cluster with CHIPD533 as representative because the score 141.248 is above the threshold (137.75). CHIPD1706 however is not part of the cluster because the alignment score with the representative is 117.317 only and thus below the threshold. Supplementary figure 3 depicts a multiple sequence alignment of the these three sequences, and additionally three other cluster members. Supplementary figure 4 depicts a phylogenetic tree of the sequences and confirms that the terminal branch of CHIPD2153 (outside) is slightly longer than that of CHIPD1706 (inside).



```
CHIPD2153_(outside_cluster)    1 ____MKRKPLLLLLLVVLSQGLH__AQQQPHYTQYILNPFIINPAVAG  48
CHIPD1706_(inside_cluster)     1 ____MQLKGMITAIILVLALPVQVLAQQQPHYTQYVLNTFIINPAVAG  48
CHIPD2136                      1 __MYTTKKWFTVVLLLCLAAGSR__AQQSVQFSQYIFNGLAINPAYAG  48
CHIPD533_(representative)      1 MK__NKNILFVVGTVLIALMPGWVKAQVDPHFSQYYAYPMWLNPALTG  48
CHIPD1704                      1 MRTFTKALLML___CLVGLTGKKLQAQSDPHFSGYYVYPAWLNPALTG  48
CHIPD1170                      1 ____MKKVLLFFTIALYLMNPARAQ___DPHFSQFFASPLTLNPAMTG  48

CHIPD2153_(outside_cluster)   49 _IENYWDVKASHRHQWTGLNGAPVTT_____YLTVHGPLRKSDYPVAS  96
CHIPD1706_(inside_cluster)    49 _IENYWDVKASHRHQWTGVNGSPVTT_____YLTIHGPLRKTDYPQAS  96
CHIPD2136                     49 _YKDVLHLNASYRQQWTGLEGAPRTG_____SISLDGPLNRGN_____  96
CHIPD533_(representative)     49 IVDGDYRVSANYRNQWVNI_GKPFST______VGVSFDAAAAN_____  96
CHIPD1704                     49 VFDGDYRVSAIYRSQWGSV_SSPFKT______YGIAGEVKTNN_____  96
CHIPD1170                     49 LFSGDFRVSGNYRSQWSSI_STPFTTGTAAVDFGILKNVLNYT_____  96

CHIPD2153_(outside_cluster)   97 VTGLTPPGDNPRGRAYWQEYTTPPAHAGVGMTILNDKTGPLNRFSISA 144
CHIPD1706_(inside_cluster)    97 ATGFNPEGSNPRGKAYWETYTAPPSHAGAGVTILNDKTGPLSRFSFSG 144
CHIPD2136                     97 ______________________KDANVGLGIQAMMDNLGPQSAISLYA 144
CHIPD533_(representative)     97 _______________________NIGVGLNITNMSAGDAGYNYLNAMA 144
CHIPD1704                     97 _______________________NINFGASVLNQKAGDGGYNYTTAYG 144
CHIPD1170                     97 _______________________DI_WGVGVMAMYDRTGGGALTSTYL 144

CHIPD2153_(outside_cluster)  145 TYA___HH_IPLSTRLS__VSGGISVGMQSVSVDAAKLQFQQPGDPV_ 192
CHIPD1706_(inside_cluster)   145 TYA___HH_IGIAPGTS__LSAGISVGAQRISLDATALEFQNPSDPA_ 192
CHIPD2136                    145 SYA___YR_IRLDEEDTRRLCFGLGVGATQYGMDGKDLIYETNGDRI_ 192
CHIPD533_(representative)    145 SFS___YRGVRFGETGTSQLVFGIQAGMINRKIDPAKWQLGSQYDPVM 192
CHIPD1704                    145 SAS___YTGVRLGAFEQHRLVFGLQMGLIQRRFDPAKLHFGDQWNPIT 192
CHIPD1170                    145 SFSTAYHKG__LDPEGNHTLAVGLQATLVQKRLDQTKLIFENQIDN_N 192

CHIPD2153_(outside_cluster)  193 _____VASSALLNKWRP___EVNAGLLLY______GPDFYLGAAAQNI 240
CHIPD1706_(inside_cluster)   193 _____IASSTVLGRWRP___DVSAGLYLY______SSQYFAGISAQNV 240
CHIPD2136                    193 _____IPDGSA_KATTP___DARVGIYYY______TPSVYIGVSVLDL 240
CHIPD533_(representative)    193 GFDPSKPSGENISTTSSNSFDAAAGVMFFDGNPNHQFNPFAGFSAGHL 240
CHIPD1704                    193 GYNPGQATNDMFNKTTSAAFDAGAGVLYYDAQPGKKYNLYGGFSVMHL 240
CHIPD1170                    193 GYNPAIPNGETFVNPTISYLDPNIGIL_YNGLVGESSNIYLGASYYHI 240

CHIPD2153_(outside_cluster)  241 VPQ__EVAYDNGKVVGDSL____YRGKLVPHLFFSGGYRLWLSEDFTM 288
CHIPD1706_(inside_cluster)   241 VPS__GIGFDDGKVKGDSI____YRGKLVPHLFATAGYRLWVNEEVSL 288
CHIPD2136                    241 LSKYTSSGYKWRGYTYESI____RRKQ___HLYVTAGYMFNVNDEISL 288
CHIPD533_(representative)    241 TQP____________QDPFVSAGSNKRLPVRYIGHGGTRIKLNEIFSL 288
CHIPD1704                    241 NKP____________SDQF_SATGDARIPMRTTAHAGVRVTISELFSL 288
CHIPD1170                    241 TQP____________TETF_MAQNNNRLTSRYTVHGGGSVPVNGANRI 288

CHIPD2153_(outside_cluster)  289 LPSVMVRLVTA_________APVSYDVNAKFMYRDRMWVGTSYRVK__ 336
CHIPD1706_(inside_cluster)   289 LPSVMIKYVTA_________LPVSFDVNAKLQYRDRIWVGGSYRYN__ 336
CHIPD2136                    289 KPSVLFKSDFS_________GPAGLDATLMMHIDELLWVGGSYRTNLS 336
CHIPD533_(representative)    289 TPHGLYMRQGNAHETVVGLYGQAYLNEEFD______FLLGANYRIN__ 336
CHIPD1704                    289 TPNVLYMKQGAASEKMLGAYGQYAVSAETD______VMLGANYRLK__ 336
CHIPD1170                    289 HFSAIFMKQSTASEISFG__GAYGFNLNGDDENPTTFYLGSWYRVK__ 336

CHIPD2153_(outside_cluster)  337 _____________DGFAAMVGV___NISSTINIGYAYDYTTSSLNAVS 384
CHIPD1706_(inside_cluster)   337 _____________DGIAAMVGI___NVNATFNIGYSYDYTTSGLNIAS 384
CHIPD2136                    337 VLNKKSIVNNTALDKANAISGILEYYISPKYRIGYSYDYSMNKLAGIQ_ 384
CHIPD533_(representative)    337 _____________DSAIPFAGF___HFKN_FVLGLSYDVNASNLRRLV 384
CHIPD1704                    337 _____________DAFSVYTGV___SYKS_FMLGLSYDVNSSDLGKIT 384
CHIPD1170                    337 _____________DAINPYLGL___EIGG_FTFGASYDTNVSTLRPAS 384

CHIPD2153_(outside_cluster)  385 KGTHEILIGFLIGNR_______YGDLCPRNNF                 416
CHIPD1706_(inside_cluster)   385 HGSHEMVLGIMIGNR_______FADLCPRNMW                 416
CHIPD2136                    385 TGSHELSIGILFNSK______LFSTSNPRYF_                 416
CHIPD533_(representative)    385 NGSNSFELSLSFISRKKKVYSEENFFCPRL__                 416
CHIPD1704                    385 RGNNSFELSLSFIGRKSVKTPAGDFVCPRL__                 416
CHIPD1170                    385 NYRGGIELSLIYIHRRN_EGSKYRTLCPRF__                 416
```

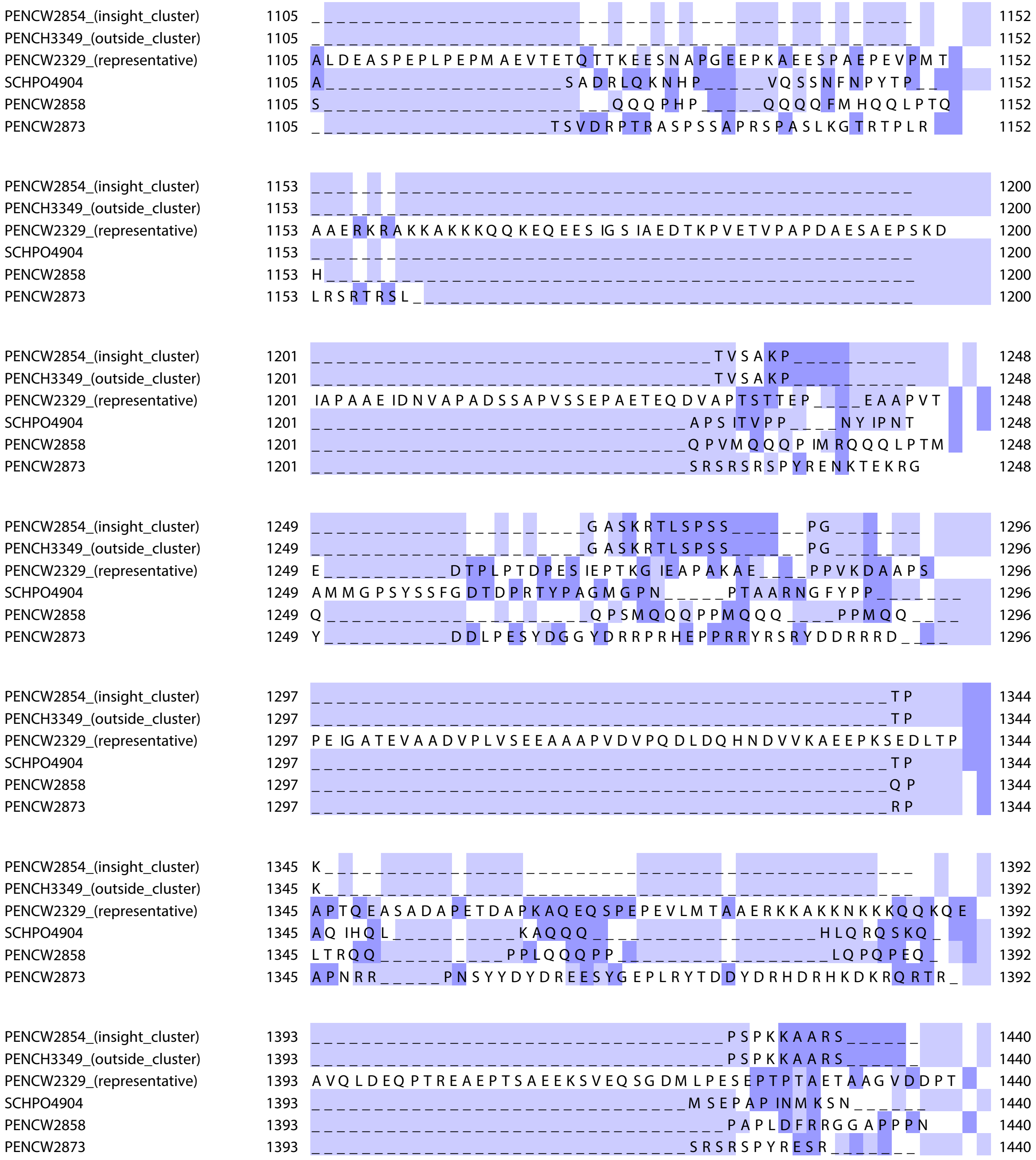
Supplementary figure 3: **Case study #1: Multiple sequence alignment of the cluster to which CHIPD2153 should be included to recover the missing pair CHIPD1706-CHIPD2153.** The corresponding tree is provided in Supplementary figure 4. Alignments drawn with JalView (Waterhouse et al. 2009 DOI:10.1093/bioinformatics/btp033)



Supplementary figure 4: **Case study #1: Distance tree of the cluster and the missing sequence (CHIPD2153).** The corresponding MSA is provided in Supplementary figure 3.

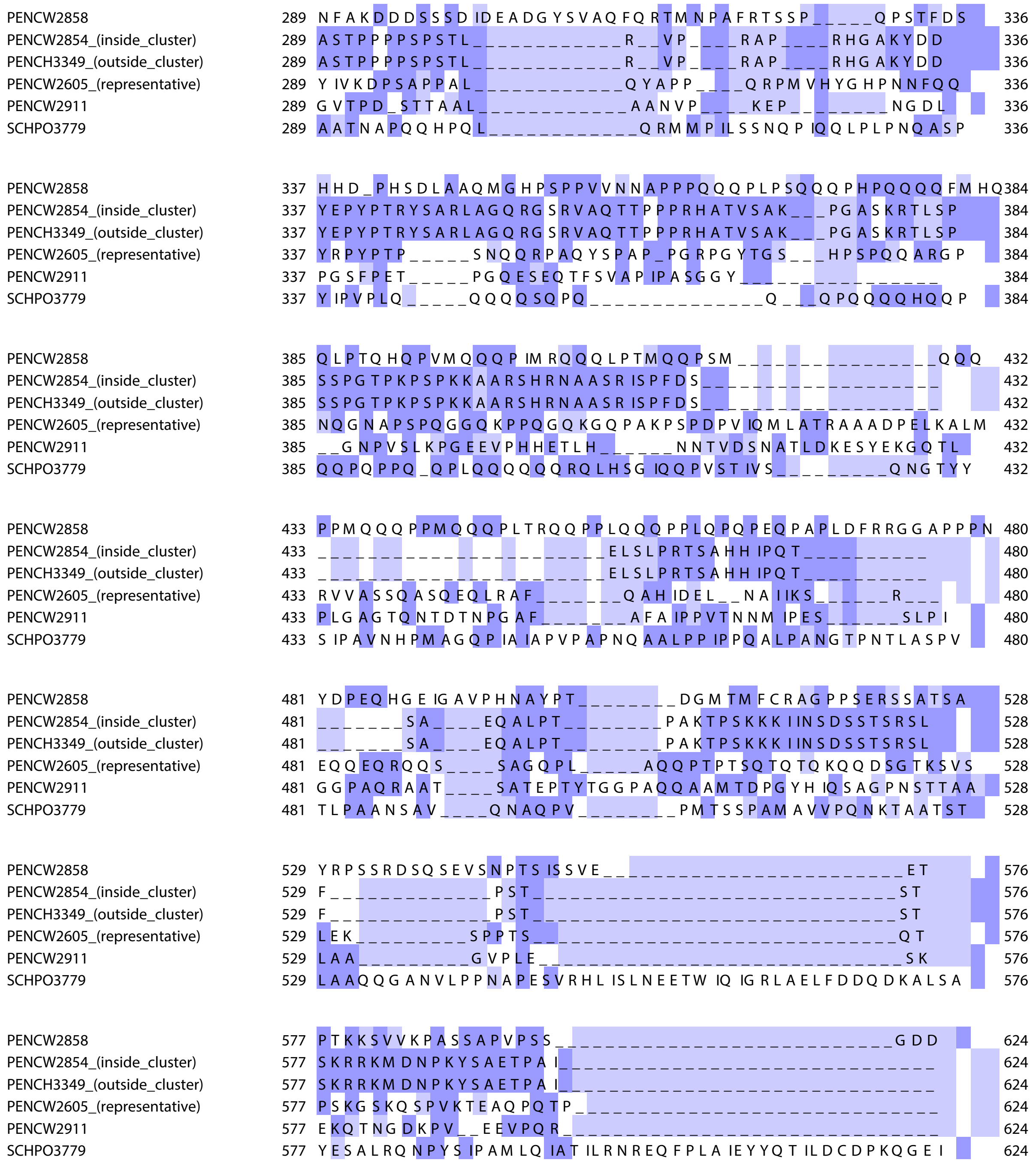
2) Fungi: The homology between sequences PENCW2854 and PENCH3349 is missed despite them being nearly identical (alignment score of 5606.4 and estimated distance of 0.38 PAM units). PENCH2854 is member of two clusters—the first cluster with representative PENCH2329 (score 138.9) and the second cluster with representative PENCW2605 (score 146.6). The alignment scores of PENCH3349 with the two representatives are below the threshold (124.6 and 128.9 respectively). Supplementary figure 5 and 6 provide representative subsets of the multiple sequence alignments for the two clusters.

Furthermore, we note that both clusters are very large: each contain >1000 sequences but only a small fraction of all member pairs are significant (7.8% and 26.96%). As mentioned in the Discussion section, splitting such clusters might improve the sensitivity and runtime of the method.



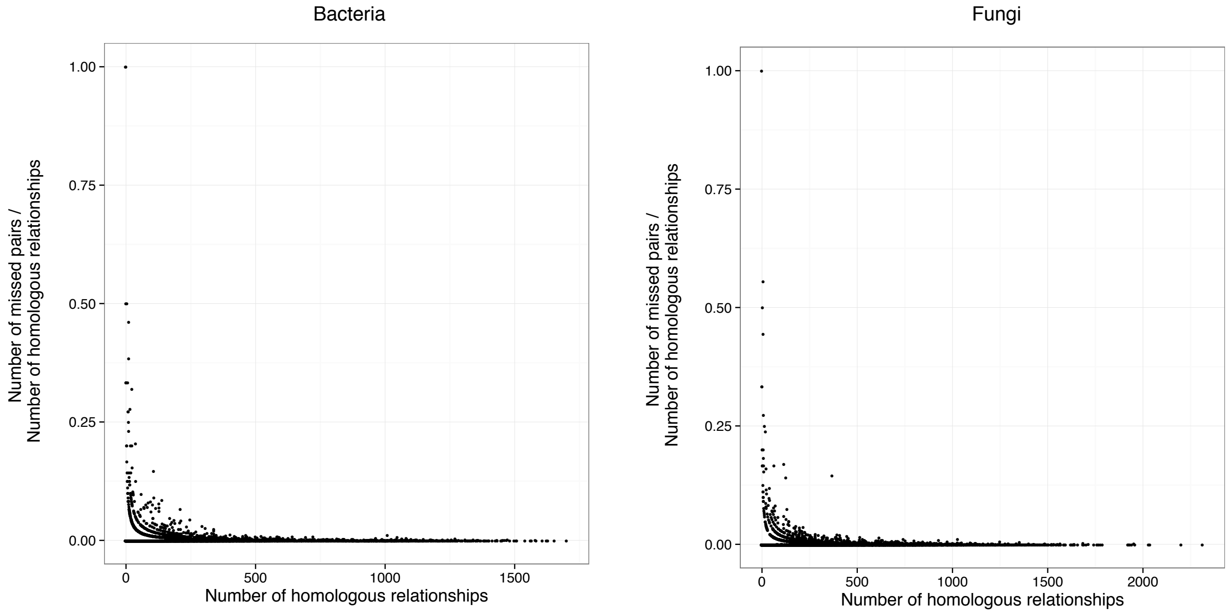
```
PENCW2854_(insight_cluster)   1105 __________________________________________ 1152
PENCH3349_(outside_cluster)   1105 __________________________________________ 1152
PENCW2329_(representative)    1105 ALDEASPEPLPEPMAEVTETQTTKEESNAPGEEPKAEESPAEPEVPMT 1152
SCHPO4904                     1105 A__________________SADRLQKNHP_____VQSSNFNPYTP__ 1152
PENCW2858                     1105 S____________________QQQPHP_____QQQQFMHQQLPTQ 1152
PENCW2873                     1105 _________________TSVDRPTRASPSSAPRSPASLKGTRTPLR 1152

PENCW2854_(insight_cluster)   1153 __________________________________________ 1200
PENCH3349_(outside_cluster)   1153 __________________________________________ 1200
PENCW2329_(representative)    1153 AAERKRAKKAKKKQQKEQEESIGSIAEDTKPVETVPAPDAESAEPSKD 1200
SCHPO4904                     1153 __________________________________________ 1200
PENCW2858                     1153 H_________________________________________ 1200
PENCW2873                     1153 LRSRTRSL___________________________________ 1200

PENCW2854_(insight_cluster)   1201 ____________________________TVSAKP_________ 1248
PENCH3349_(outside_cluster)   1201 ____________________________TVSAKP_________ 1248
PENCW2329_(representative)    1201 IAPAAEIDNVAPADSSAPVSSEPAETEQDVAPTSTTEP____EAAPVT 1248
SCHPO4904                     1201 __________________________APSITVPP____NYIPNT 1248
PENCW2858                     1201 __________________________QPVMQQQPIMRQQQLPTM 1248
PENCW2873                     1201 __________________________SRSRSRSPYRENKTEKRG 1248

PENCW2854_(insight_cluster)   1249 ___________________GASKRTLSPSS______PG______ 1296
PENCH3349_(outside_cluster)   1249 ___________________GASKRTLSPSS______PG______ 1296
PENCW2329_(representative)    1249 E____________DTPLPTDPESIEPTKGIEAPAKAE____PPVKDAAPS 1296
SCHPO4904                     1249 AMMGPSYSSFGDTDPRTYPAGMGPN_____PTAARNGFYPP_______ 1296
PENCW2858                     1249 Q__________________QPSMQQQPPMQQQ____PPMQQ____ 1296
PENCW2873                     1249 Y____________DDLPESYDGGYDRRPRHEPPRRYRSRYDDRRRD____ 1296

PENCW2854_(insight_cluster)   1297 _________________________________________TP 1344
PENCH3349_(outside_cluster)   1297 _________________________________________TP 1344
PENCW2329_(representative)    1297 PEIGATEVAADVPLVSEEAAAPVDVPQDLDQHNDVVKAEEPKSEDLTP 1344
SCHPO4904                     1297 _________________________________________TP 1344
PENCW2858                     1297 _________________________________________QP 1344
PENCW2873                     1297 _________________________________________RP 1344

PENCW2854_(insight_cluster)   1345 K__________________________________________ 1392
PENCH3349_(outside_cluster)   1345 K__________________________________________ 1392
PENCW2329_(representative)    1345 APTQEASADAPETDAPKAQEQSPEPEVLMTAAERKKAKKNKKKQQKQE 1392
SCHPO4904                     1345 AQIHQL_________KAQQQ________________HLQRQSKQ_ 1392
PENCW2858                     1345 LTRQQ_________PPLQQQPP_______________LQPQPEQ_ 1392
PENCW2873                     1345 APNRR____PNSYYDYDREESYGEPLRYTDDYDRHDRHKDKRQRTR_ 1392

PENCW2854_(insight_cluster)   1393 ______________________________PSPKKAARS_____ 1440
PENCH3349_(outside_cluster)   1393 ______________________________PSPKKAARS_____ 1440
PENCW2329_(representative)    1393 AVQLDEQPTREAEPTSAEEKSVEQSGDMLPESEPTPTAETAAGVDDPT 1440
SCHPO4904                     1393 ___________________________MSEPAPINMKSN_____ 1440
PENCW2858                     1393 ______________________________PAPLDFRRGGAPPPN 1440
PENCW2873                     1393 ___________________________SRSRSPYRESR______ 1440
```

Supplementary figure 5: **Case study #2: Representative extract of the multiple sequence alignment of the first cluster to which PENCH3349 should be included to recover the missing pair PENCW2854-PENCH3349.** Alignments drawn with JalView (Waterhouse et al. 2009 DOI:10.1093/bioinformatics/btp033)

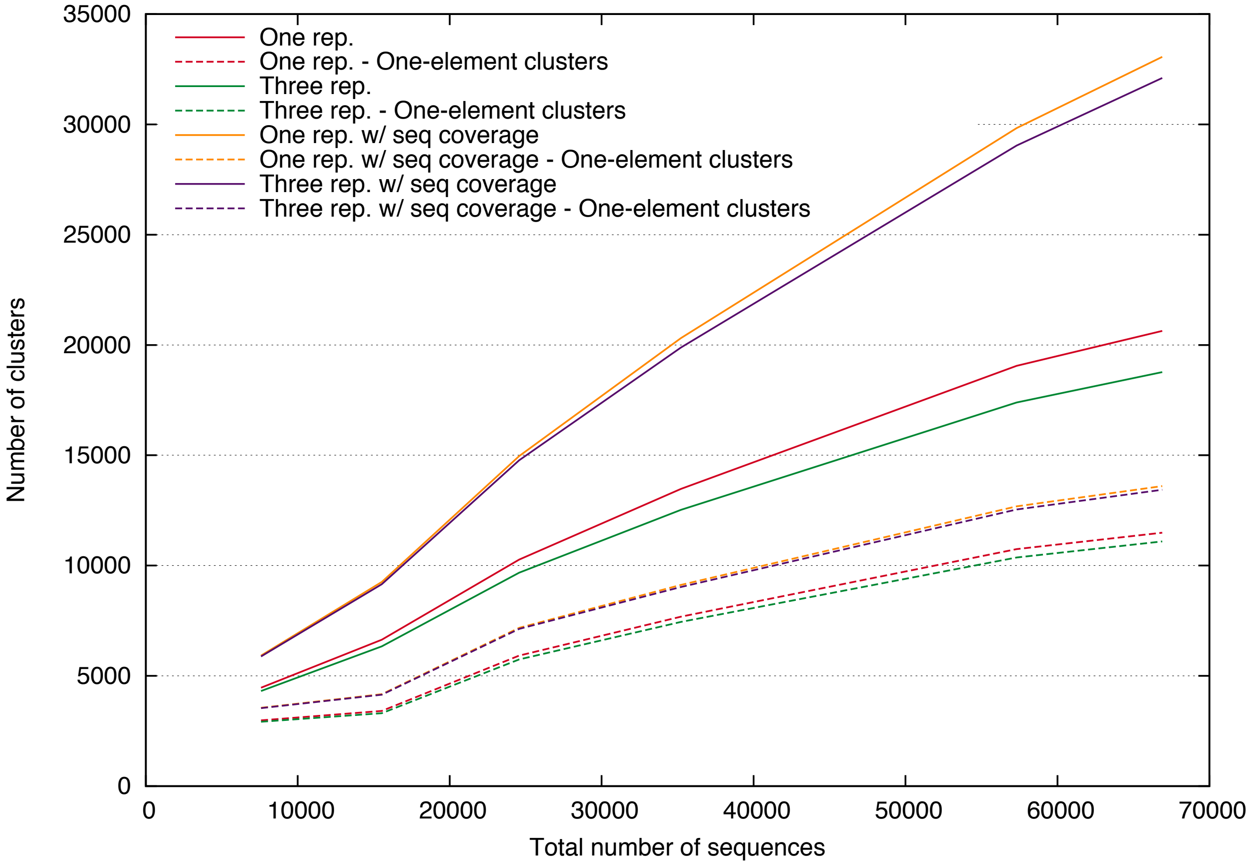


```
PENCW2858                       289 NFAKDDDSSSDIDEADGYSVAQFQRTMNPAFRTSSP_____QPSTFDS 336
PENCW2854_(inside_cluster)      289 ASTPPPPSPSTL_______________R__VP____RAP____RHGAKYDD 336
PENCH3349_(outside_cluster)     289 ASTPPPPSPSTL_______________R__VP____RAP____RHGAKYDD 336
PENCW2605_(representative)      289 YIVKDPSAPPAL_______________QYAPP____QRPMVHYGHPNNFQQ 336
PENCW2911                       289 GVTPD_STTAAL_______________AANVP____KEP_________NGDL 336
SCHPO3779                       289 AATNAPQQHPQL_______________QRMMPILSSNQPIQQLPLPNQASP 336

PENCW2858                       337 HHD_PHSDLAAQMGHPSPPVVNNAPPPQQQPLPSQQQPHPQQQQFMHQ 384
PENCW2854_(inside_cluster)      337 YEPYPTRYSARLAGQRGSRVAQTTPPPRHATVSAK___PGASKRTLSP 384
PENCH3349_(outside_cluster)     337 YEPYPTRYSARLAGQRGSRVAQTTPPPRHATVSAK___PGASKRTLSP 384
PENCW2605_(representative)      337 YRPYPTP______SNQQRPAQYSPAP_PGRPGYTGS___HPSPQQARGP 384
PENCW2911                       337 PGSFPET______PGQESEQTFSVAPIPASGGY____________________ 384
SCHPO3779                       337 YIPVPLQ______QQQQSQPQ__________________Q___QPQQQQHQQP 384

PENCW2858                       385 QLPTQHQPVMQQQPIMRQQQLPTMQQPSM____________________QQQ 432
PENCW2854_(inside_cluster)      385 SSPGTPKPSPKKAARSHRNAASRISPFDS________________________ 432
PENCH3349_(outside_cluster)     385 SSPGTPKPSPKKAARSHRNAASRISPFDS________________________ 432
PENCW2605_(representative)      385 NQGNAPSPQGGQKPPQGQKGQPAKPSPDPVIQMLATRAAADPELKALM 432
PENCW2911                       385 __GNPVSLKPGEEVPHHETLH_______NNTVDSNATLDKESYEKGQTL 432
SCHPO3779                       385 QQPQPPQ_QPLQQQQQQRQLHSGIQQPVSTIVS___________QNGTYY 432

PENCW2858                       433 PPMQQQPPMQQQPLTRQQPPLQQQPPLQPQPEQPAPLDFRRGGAPPPN 480
PENCW2854_(inside_cluster)      433 _____________________________ELSLPRTSAHHIPQT____________ 480
PENCH3349_(outside_cluster)     433 _____________________________ELSLPRTSAHHIPQT____________ 480
PENCW2605_(representative)      433 RVVASSQASQEQLRAF________QAHIDEL__NAIIKS_______R___ 480
PENCW2911                       433 PLGAGTQNTDTNPGAF_________AFAIPPVTNNMIPES_______SLPI 480
SCHPO3779                       433 SIPAVNHPMAGQPIAIAPVPAPNQAALPPIPPQALPANGTPNTLASPV 480

PENCW2858                       481 YDPEQHGEIGAVPHNAYPT__________DGMTMFCRAGPPSERSSATSA 528
PENCW2854_(inside_cluster)      481 _________SA_____EQALPT__________PAKTPSKKKIINSDSSTSRSL 528
PENCH3349_(outside_cluster)     481 _________SA_____EQALPT__________PAKTPSKKKIINSDSSTSRSL 528
PENCW2605_(representative)      481 EQQEQRQQS_____SAGQPL______AQQPTPTSQTQTQKQQDSGTKSVS 528
PENCW2911                       481 GGPAQRAAT_____SATEPTYTGGPAQQAAMTDPGYHIQSAGPNSTTAA 528
SCHPO3779                       481 TLPAANSAV_____QNAQPV__________PMTSSPAMAVVPQNKTAATST 528

PENCW2858                       529 YRPSSRDSQSEVSNPTSISSVE_______________________________ET 576
PENCW2854_(inside_cluster)      529 F________________PST_____________________________________ST 576
PENCH3349_(outside_cluster)     529 F________________PST_____________________________________ST 576
PENCW2605_(representative)      529 LEK___________SPPTS_____________________________________QT 576
PENCW2911                       529 LAA___________GVPLE_____________________________________SK 576
SCHPO3779                       529 LAAQQGANVLPPNAPESVRHLISLNEETWIQIGRLAELFDDQDKALSA 576

PENCW2858                       577 PTKKSVVKPASSAPVPSS___________________________________GDD 624
PENCW2854_(inside_cluster)      577 SKRRKMDNPKYSAETPAI______________________________________ 624
PENCH3349_(outside_cluster)     577 SKRRKMDNPKYSAETPAI______________________________________ 624
PENCW2605_(representative)      577 PSKGSKQSPVKTEAQPQTP_____________________________________ 624
PENCW2911                       577 EKQTNGDKPV__EEVPQR______________________________________ 624
SCHPO3779                       577 YESALRQNPYSIPAMLQIATILRNREQFPLAIEYYQTILDCDPKQGEI 624
```

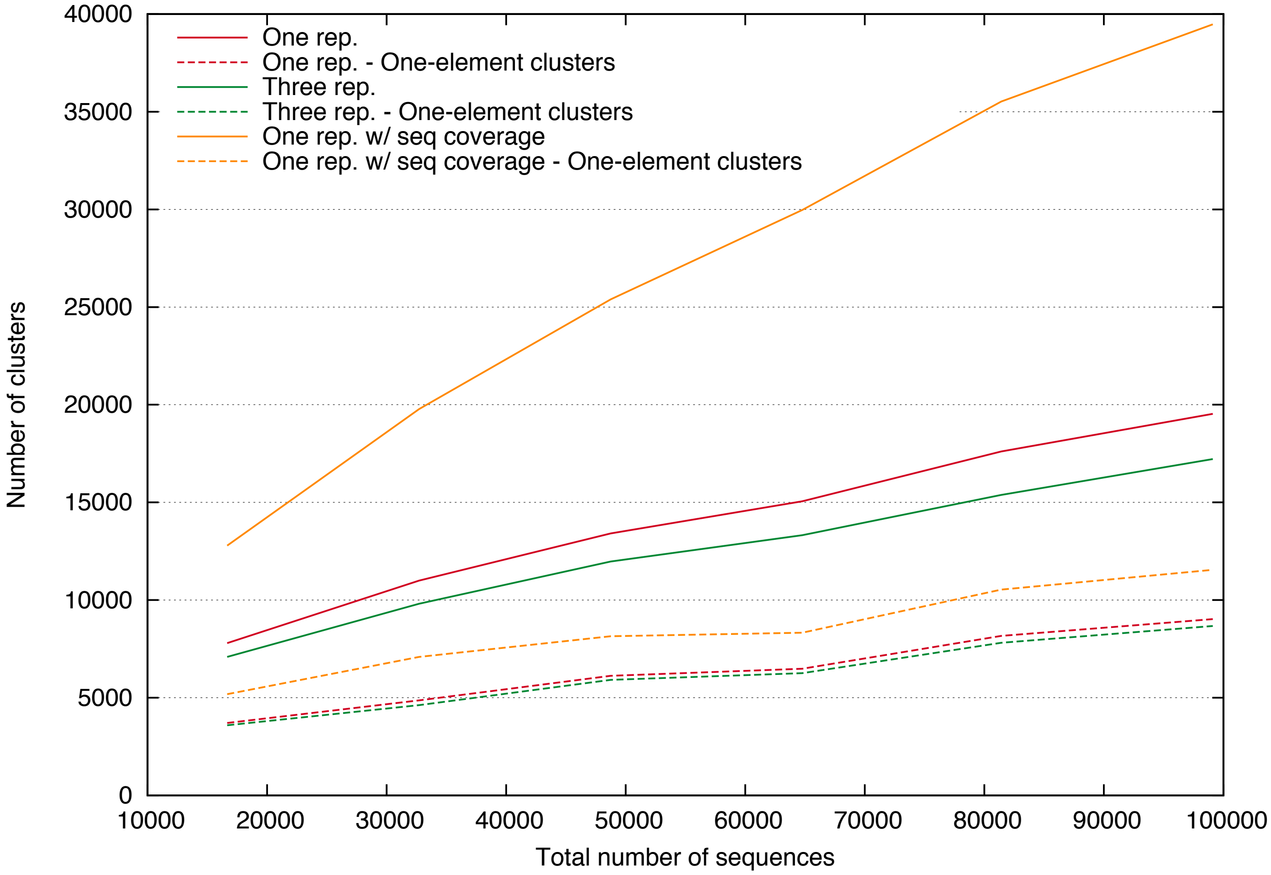
Supplementary figure 6: **Case study #2: Representative extract of the multiple sequence alignment of the second cluster to which PENCH3349 should be included to recover the missing pair PENCW2854-PENCH3349.** Alignments drawn with JalView (Waterhouse et al. 2009 DOI:10.1093/bioinformatics/btp033)



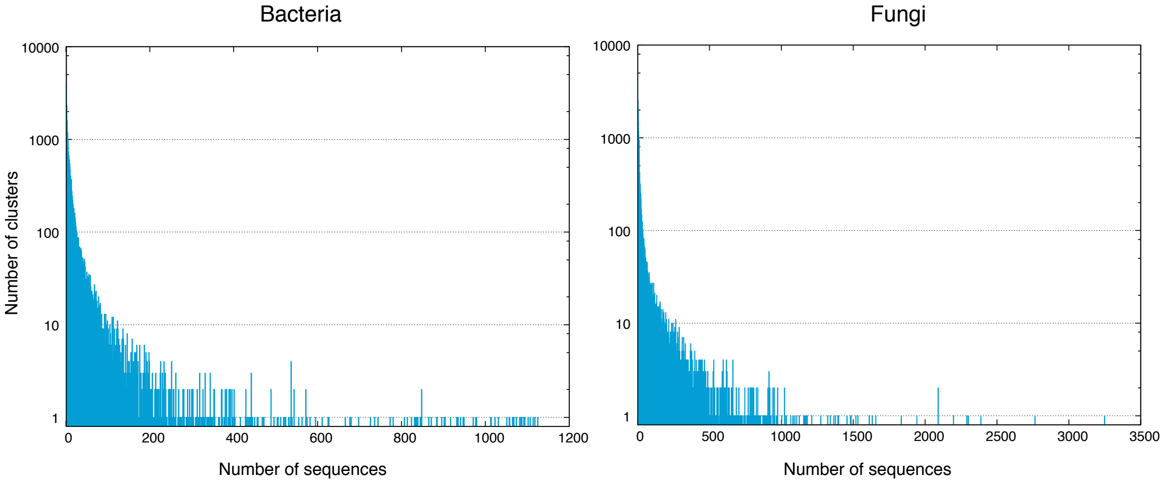
Supplementary figure 7: **Fraction of missing homologous relationships for each gene, as a function of the number of homologous relationships, for the full bacteria (14 proteomes) and fungi (12 proteomes) datasets with one representative and subsequence homology.** In large gene families (>100 members), virtually all homologous relationships are recovered compared to full all-against-all. This is also largely true for small families. The plot suggests that errors are distributed quite evenly across all types of genes.



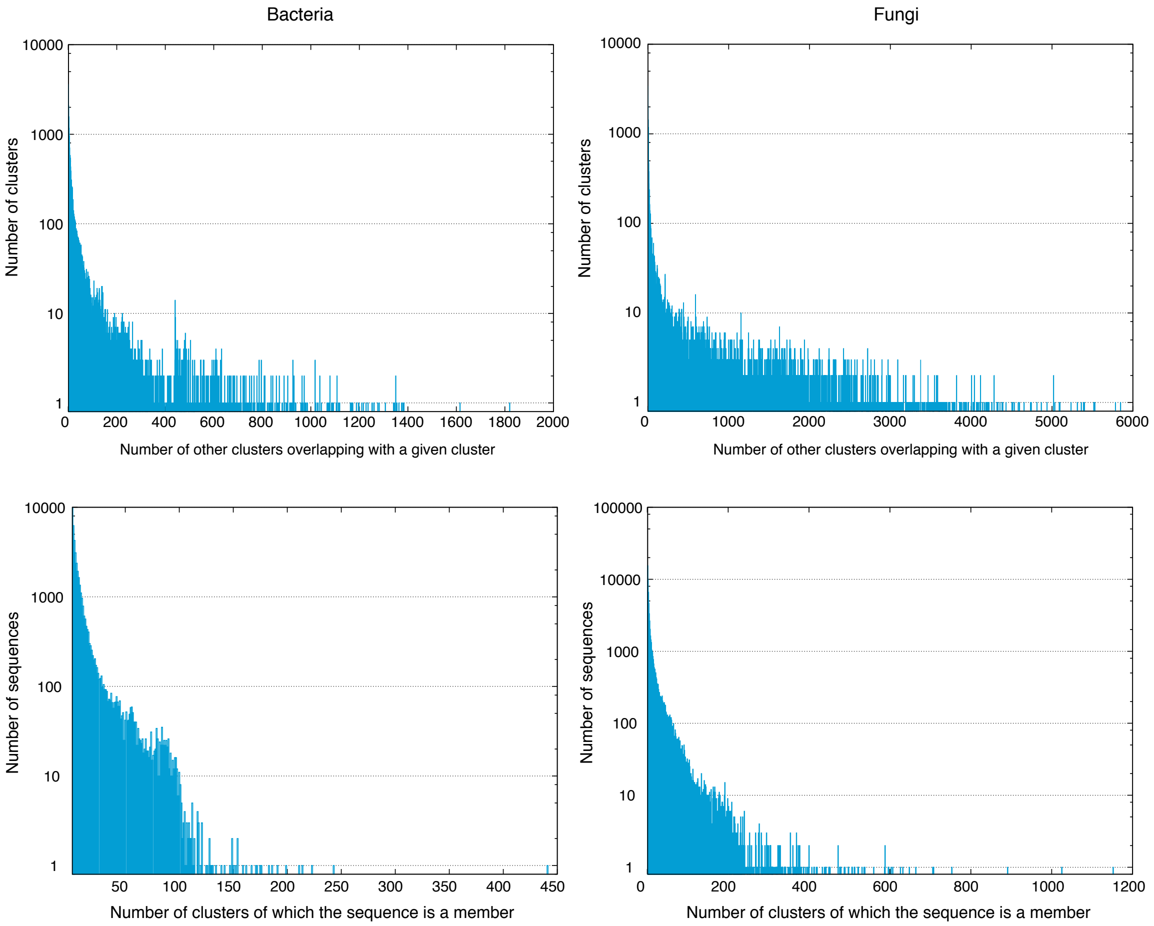
Supplementary figure 8: **Growth of number of clusters on bacteria dataset.** No tapering is observed in the growth in the number of clusters generated by the new method



Supplementary figure 9: **Growth of number of clusters on fungi dataset.** No tapering is observed in the growth in the number of clusters generated by the new method.



Supplementary figure 10: **Distribution of cluster size for the full bacteria (14 proteomes) and fungi (12 proteomes) datasets with one representative and subsequence homology.** The distribution is heavily skewed toward small clusters.



Supplementary figure 11: **Histogram of the number of clusters overlapping with each cluster (top row) and of the number of clusters in which each sequence is involved (bottom row) for the full bacteria (14 proteomes) and fungi (12 proteomes) datasets with one representative and subsequence homology.** Some clusters overlap with thousands of other clusters, which suggest potential to merge some of them (see *Discussion*).