Sequence, structure and functional diversity of PD-(D/E)XK phosphodiesterase superfamily

Supplementary Materials

Kamil Steczkiewicz[1,a], Anna Muszewska[1,a], Lukasz Knizewski[1], Leszek Rychlewski[2], Krzysztof Ginalski[1,*]

[1]Laboratory of Bioinformatics and Systems Biology, CENT, University of Warsaw,
Zwirki i Wigury 93, 02-089 Warsaw, Poland

[2]BioInfoBank Institute, Limanowskiego 24a, 60-744 Poznan, Poland

[a]These authors equally contributed to this work

[*]To whom correspondence should be addressed. Tel: +48 22 5540800; Fax: +48 22 5540801;
Email: kginal@cent.uw.edu.pl

Table of Contents

Materials and Methods

Identification of PD-(D/E)XK families and structures

Known PD-(D/E)XK phosphodiesterases cataloged in SCOP (12) and Pfam (19) databases were used as queries to identify new superfamily members. Detection of new PD-(D/E)XK families (Pfam, COG, KOG) and structures (PDB) was performed with the GRDB system that employs Meta-BASIC (40), a distant homology detection method based on a comparison of meta-profiles. GRDB provides pre-calculated mappings between COG, KOG, Pfam and PDB90 (PDB clustered at 90% sequence identity) databases. The extreme sequence and structure diversity of the PD-(D/E)XK nuclease fold is likely to be reflected by Meta-BASIC scores lower than the confidence threshold of 40 (predictions with scores >40 have <5% chance of being incorrect). To address this issue, we extended our transitive searches by analyzing all Meta-BASIC hits including predictions ranked lower than the first false-positive. All hits were manually inspected for the conservation of secondary structure elements, hydrophobicity patterns and essential active site residues. Eventually, for every hit alternative assignment hypotheses were also considered. Hits that passed manual verification were used for further Meta-BASIC searches until no new families or structures were detected. Non trivial assignments were additionally confirmed with 3D-Jury (41) which uses models provided by a number of fold recognition methods, to select a consensus prediction which is more reliable than those originated from the single modeling server.

Multiple sequence alignments

Structure-based alignment was derived from a manually curated superimposition of all identified PD-(D/E)XK structures. The alignment was built for all of the fold core elements: four β-strands and two α-helices. One should consider that due to extreme diversity of PD-(D/E)XK structures, it was highly non trivial to obtain a consistent structural alignment.

Sequences of proteins belonging to the identified families were collected with PSI-BLAST (42) searches (3 iterations with an E-value threshold of 0.005) against NCBI nr (non-redundant protein sequence) database using Pfam, COG, KOG consensus sequences and PDB structures as queries. Multiple sequence alignments were prepared using PCMA (43). Additionally, PSI-

PRED (164) secondary structure predictions were performed for every aligned sequence to evaluate the secondary structure conservation patterns.

The final alignment for the whole PD-(D/E)XK superfamily was assembled from sequence-to-structure mappings between the considered families and the closest structures. The family-to-structure mappings were built manually using a consensus alignment and 3D assessment approach (44) with respect to Meta-BASIC and 3D-Jury (41) alignments, the predicted secondary structure and conservation of critical active site residues and hydrophobic patterns.

Grouping PD-(D/E)XK realm

The collected PD-(D/E)XK fold proteins were assembled into groups of closely related families and structures. All families and structures within a single group share relatively high sequence similarity detectable with both PSI-BLAST and RPS-BLAST with confident E-value lower than 0.005. Consequently, families within the clusters have very similar PD-(D/E)XK fingerprinting, hydrophobic regions and insertion patterns. Therefore, the sequences belonging to one cluster may have closely related functions.

Structure similarity based searches

A pre-calculated data-set of PDB structures, describing parameterized architecture and topology of the secondary structure elements, (available at http://prodata.swmed.edu/ProSMoS) was analyzed with ProSMoS program (45) in order to identify proteins displaying a PD-(D/E)XK restriction endonuclease-like fold. The conservation of PD-(D/E)XK core elements in the obtained set of structures were judged manually.

Domain architecture analysis

All identified 21 911 sequences were scanned with RPS-BLAST against protein domains stored in the GRDB system (COG, KOG, PfamA) and CDD database (with E-value threshold 0.001), and with HMMscan.pl script (from HMMER3 package) against PfamA version 25 (with E-value threshold 1). Representatives for each domain architecture from each of 121 PD-(D/E)XK clusters were inspected manually. Transmembrane regions were detected with a

TMHMM server (46). Cellular localization for prokaryotic sequences was predicted with PSORTb (47) and for eukaryotic with Cello (48), WoLF PSORT (49) and Multiloc (50).

Taxonomic distribution, genomic context and HGT detection

In order to determine taxonomic distribution for each of the 121 PD-(D/E)XK groups, all sequences were cross-linked with the corresponding taxonomic identifiers (taxonomy database from NCBI). Multiple sequence alignment for each cluster was calculated with MAFFT (standalone, local pair flavor, version 6.861, a maxiterate 500 parameter) (165). The MSA was trimmed according to column evolutionary conservation with trimAl (strict parameter) (166). Maximum likelihood analysis of each cluster was carried out with PhyML with the following settings: the LG model of amino acid substitution, 4 categories in the gamma model with the shape parameter estimated and 5 random starting trees (167). For better performance, instead of bootstrapping replicates a branch support was calculated using an approximate likelihood ratio test (167). Trees were analyzed and visualized using Archaeopteryx package (168). Trees graphics were prepared in iTol (169). Trees of clans were rooted with two sequences belonging to the most similar (in terms of Meta-BASIC scores) yet separate PFAM protein family. Trees were subsequently pruned to branch support values above 0.5. The genomic context was analyzed with The SEED (51), GeContII (52), MicrobesOnline (53) and NCBI genomic resources. For a general sequence comparison a CLANS (54) clustering of all 21 911 sequences was performed, and high resolution figures were drawn with an in-house script based on CLANS scores (run file).

PD-(D/E)XK fold in other unrelated structures

Niv and colleagues suggested that the PD-(D/E)XK structural core might appear in other, unrelated protein structures outlining the so-called “Russian doll” effect (170). Using topology-based searches with ProSMoS (45) we have detected the PD-(D/E)XK nuclease fold in various other structures such as N-acetyltransferase (pdb|1cjw, SCOP classification d.108), lipase (pdb|1ein (171), SCOP classification c.69), carbon monoxide dehydrogenase (pdb|1ffu, SCOP classification d.133), hypothetical protein ybgI (pdb|1nmo (172), SCOP classification c.135), glycerate kinase (pdb|1to6, SCOP classification c.141), ornithine acetyltransferase (pdb|1vz6, SCOP classification d.154) and NEDD8 protease (pdb|2bkq, SCOP classification d.3). All these unrelated proteins contain the substructures retaining the architecture and topology of the PD-(D/E)XK domain, although it lost a characteristic, Y-shaped bend discussed above (Supplementary Figure S2). Noteworthy, neither the whole protein itself nor its PD-(D/E)XK substructure function as a phosphodiesterase.

Domain architecture

Nucleic acid recognition and processing

Sequences from multiple groups, including among others restriction endonucleases, Holiday junction resolvases and Vsr repair endonucleases co-occur with methylases, helicases, DNA recognition motifs and many other protein domains. Various domain architectures are obvious indicators of adjustment to potentially new functions while others contribute to known functional relationships. For example, the PD-(D/E)XK restriction endonucleases usually belong to Type II which requires a separate endonuclease and methylase providing restriction and protection, respectively. However, in the set of PD-(D/E)XK sequences we observed the occasional presence of methyltransferase both N-terminal and C-terminal to the restriction endonuclease domain. A DNA (cytosine-5-)-methyltransferase from *Sulfurimonas denitrificans* (gi|78778033) possesses the HpaII restriction endonuclease domain flanked with two DNA methylases. *Nitrosococcus oceani* nuclease (gi|77165284) also has a nuclease domain with methyltransferase domains on both sides. Similarly, a cytosine-specific DNA methyltransferase from *Mycoplasma hominis* (gi|269114810) and a Type II restriction enzyme HaeII from *Lyngbya* sp. (gi|119493789) contain the PD-(D/E)XK domain preceded by the methyltransferase domain. A DNA-cytosine methyltransferase from *Cyanothece* sp. (gi|307152624) contains two methylase domains preceding the BsuBI restriction endonuclease domain. Such architectures cannot be analyzed in a wider taxonomic context due to extreme divergence among restriction-modification systems. However, their co-occurrence may be justified by strict co-operation of these enzymes. A very interesting suite of fused domains can be observed within the Vsr/MutH mismatch repair nucleases. The MutHLS repair system recognizes and repairs mismatched bases exploiting their methylation state. A repair endonuclease from extremophilic and radiation-resistant *Truepera radiovictrix* (gi|297624926) includes methyltransferase C-terminal to nuclease, whereas in the *Chlorobium tepidum* protein (gi|21674545) methyltransferase precedes the PD-(D/E)XK domain. Mismatch repair domains can also be assisted with helicase-associated proteins. The *Sideroxydans lithotrophicus* protein (gi|291612916) contains three distinct helicase domains: an ATP-binding region of DEAD helicase (PF04851), a helicase C-terminal domain (PF00271), and an RNA helicase domain (DUF3418, PF11898) along with a Vsr-like domain.

In a hypothetical protein from *Pyrenophora teres* (gi|330944954) the ERCC4 repair endonuclease (PD-(D/E)XK) domain co-occurs with the mitochondrial CBP4 transmembrane protein required for the assembly of the cytochrome bc1 complex. Therefore, the protein may play an essential role in nucleic acid maintenance in the strongly oxidative environment of mitochondria.

Calcium binding

Two PD-(D/E)XK proteins: Rai1 and RAP from *Arabidopsis thaliana* (phosphodiesterase, gi|2245120) and *Aureococcus anophagefferens* (gi|323447941), respectively, are associated with Ca^{2+} channel formation. The *A. thaliana* protein contains the Ca^{2+} binding domain (EF hand, PF00036) followed by multiple pentatricopeptide repeats (PPR, PF01535), two transmembrane elements and the Rai1 phosphodiesterase domain. The PPR repeats are found to take part in RNA editing (173), while the EF helix-loop-helix structural domain is involved in Ca^{2+} binding (174). Altogether, these domains may provide a mechanism for customizing the ion channel function as described by Tan and colleagues (175). On the other hand, an RAP domain from *A. anophagefferens* (gi|323447941) accompanied by an inositol 1,4,5-trisphosphate/ryanodine receptor domain (Ins145_P3_rec, PF08709), an RIH domain (RYDR_ITPR, PF01365), an EF-hand domain (PF00036), an epidermal growth factor domain (EGF-like, PF00008) and a glycoside hydrolase family 5 domain (Cellulase, PF00150), which together form a 5000 amino acid long protein. These domains are separated by transmembrane, low complexity and repeated regions. Four of these domains (Ins145_P3_rec, RIH, EGF, EF-hand) are involved in calcium channel formation and signaling, which additionally explains the presence of multiple transmembrane regions.

Regulatory proteins

Interestingly, we observed that the RAP (PD-(D/E)XK) domain ordinarily co-occurs with two kinase domains (FAST_1, PF06743 and FAST_2, PF08368), which is common within Metazoa e.g. in the *Trichoplax adhaerens* hypothetical protein (gi|196000781). The FAST kinases are essential for cellular respiration, they are also involved in stress sensing and play a role in splicing regulation (151). Noteworthy, all these proteins harbor the RAP domain. The human genome, however, also encodes FAST kinases which however contain only a single FAST domain (e.g. gi|119628500). The discussed architecture may be decorated with additional repeat motifs as in *Taeniopygia guttata* protein (gi|224050319).

The proteins containing the PD-(D/E)XK domains may regulate proper protein folding. For instance, the ERCC4 domain can be encountered with the Mannosyl_trans3 domain (PF11051). This domain catalyzes O-mannosylation reaction, where mannose is transferred from mannose-p-dolichol to a serine/threonine residue in secretory pathway proteins (176). In yeast O-mannosylation of aberrant proteins reduces the load for ER chaperones (177). Thus, the potential role of this protein is to induce a degradation signal for misfolded proteins. The RPB5_N domain is present in *Trichinella spiralis* retinol dehydrogenase (gi|316979012), where it is fused to a short chain dehydrogenase domain (PF00106). This architecture is limited to Trichinella, proteins sequences from the second nematode *Caenorhabditis elegans* lack the aforementioned architecture. Retinoic acid plays an essential role in development and is involved in all stages of embryogenesis and morphogenesis (178). Moreover, retinoic acid regulates proteins that maintain a specific open chromatin conformation (179). Additionally, a hypothetical protein from *Vitis vinifera* (gi|147821195) combines two domains: ERCC4 and cyclophilin (proline isomerase, PF00160). This is an interesting combination of DNA structure specific repair endonuclease with proline isomerase. The latter domain may function as a chaperone (180) and is reported to prevent the aggregation of folded histone proteins with DNA during the assembly of nucleosomes (181). This enzyme may represent a versatile repairing entity that addresses both DNA and protein issues.

Another, ERCC4-containing and probably regulatory protein from a sea anemone *Nematostella vectensis* (gi|156390228) exhibits a unique domain arrangement comprising the UIM (PF02809), EGF epidermal growth factor (PF00008), DEK_C (PF08766), and SWIB (PF02201). The SWIB/MDM2 domain binds to the transactivation domain and downregulates the ability of p53 to activate transcription (182). DEK is a chromatin associated protein that is linked with cancers and autoimmune disease. This domain is found at the C terminal of DEK and is of clinical importance since it can reverse the characteristic abnormal DNA-mutagen sensitivity in fibroblasts from ataxia-telangiectasia (A-T) patients (183).

The presented domain architectures (Supplementary Figure S3) demonstrate that the PD-(D/E)XK proteins, in addition to nucleic acid metabolism, may also be involved in stress sensing, cell cycle regulation and many other important biological processes.

Taxonomic distribution and HGTs

HGT between animal related bacteria

Transfers involving human or animal pathogens were mostly found in bacteria inhabiting or infecting the respiratory tract and oral cavity. A prominent example of such an HGT event can be observed for FokI restriction endonuclease from *Haemophilus influenzae* 22.4-21 (gi|145641810). Both the tree topology and the genomic context of this protein family support the HGT hypothesis (see Supplementary Figure S4). *H. influenzae* sequence is present in a well-supported clade (aLTR score: 0.99) along with *Streptococcus equi* subsp. *zooepidemicus* (gi|225868181), *S. sanguinis* VMC66 (gi|323350332) and *S. equi* subsp. *equi* 4047 (gi|225870881) sequences. The genomic organization in the proximity of the gene encoding FokI in *H. influenzae* includes mobile elements (ME). ME are known to mediate lateral transfer and are one of the major factors facilitating HGT (184). *S. sanguinis* is a human opportunistic pathogen, *S. equi* subsp. *equi* is a specialized horse pathogen and *S. equi* subsp. *zooepidemicus* is an opportunistic pathogen of domestic animals, also dangerous for humans. *H. influenzae* is a naturally competent nasolaryngeal pathogen (185). All of these share similar ecological niches, which additionally favors the HGT possibility. High sequence similarities between the analyzed sequences as well as very limited distribution of FokI enzymes make HGT even more probable. The other restrictase transferred between human pathogens is HinP1I (see Supplementary Figure S4). The observed taxonomic distribution and phylogenetic tree topology justifies us to hypothesize about multiple transfers from

Proteobacteria to *Leptotrichia goodfellowii* (Fusobacteriales) (gi|262038310), and between *Moraxella catarrhalis* (Pseudomonadaceae) (gi|326560964) and some Haemophilus species (Pasteruellaceae). *H. somnous* is a bovine pathogen, *L. goodfellowii* is found in dental plaque, and *M. catarrhalis* was recently described as a respiratory pathogen (186). *Campylobacter upsaliensis* and *Helicobacter pylori* are human pathogens causing gastrointestinal infections. Noteworthy, all bacteria encoding the HinP1Iare related to animal hosts. The genomic context of the PD-(D/E)XK proteins between all these bacteria is not conserved and MTase is a single co-occurring gene.

HGT from Prokaryota to Eukaryota

Transfers from Prokaryota to Eukaryota have also been spotted for Oomycota, insects and nematodes (187-189). Within the PD-(D/E)XK superfamily we observed possible transfers to a chytrid, plants, pelagophyte algae and a nematode. A chytrid, *Batrachochytrium dendrobatidis* JAM81 protein (gi|328765959) has no eukaryotic homologs, but locates with *Desulfotomaculum nigrificans* DSM 574 protein (gi|323703925) on the phylogenetic tree of LlaJI, McrBC restriction endonucleases. *B. dendrobatidis* is the causal agent of chytridiomycosis responsible for the depletion of amphibian global population. Recently, Sun and colleagues reported a study concerning CRN and serine peptidase acquisition by *B. dendrobatidis* (190). Our results demonstrate that not only typical virulence effectors, but also endonucleases can be transferred into chytrids. However, the role of restriction endonuclease close homologs in Fungi needs further studies.

The castor oil plant, *Ricinus communis* provides an outstanding example of HGT, because its genome harbors more than one potentially transferred gene encoding a PD-(D/E)XK protein domain. It probably acquired a TnsA transposase (gi|255595804) and proteins of unknown function: DUF1016 (gi|255589015) and PDDEXK_4 (gi|255614689) from an unknown Proteobacterial source. Other plants also possess genes with homology to bacterial PD-(D/E)XK sequences. A *Populus balsamifera* subsp. *trichocarpa* protein (gi|222874199) shares 100% sequence identity with the *Ralstonia metallidurans* protein (gi|291434870) suggesting a very recent horizontal transfer. Balsam poplar is a dynamically expanding tree from North America and *R. metallidurans* is a metal resistant bacterium found in contaminated waste and soil. To elucidate the significance of this transfer it is indispensable to assess the function of this gene encoding a protein of unknown function (DUF4263). Another detected transfer in plants concerns EXOV exonuclease from rice, which forms a consistent clade (aLTR score 1) with Enterobacteriaceae, including an aphid endosymbiont *Serratia symbiotica* str. Tucson (gi|320539715), *Serratia odorifera* 4Rx13 (gi|270264844), *Serratia odorifera* DSM 4582 (gi|293394677) and *Serratia proteamaculans* 568 (gi|157372049). *S. odorifera* is a rhizobacterium; *S. proteamaculans* 568 was isolated as a root endophyte from poplar *Populus trichocarpa*; and *S. odorifera* were isolated as a part of Human Microbiome Project (HMP) from human samples. This is a very intriguing example of the HGT, which involves interaction between rice, soil bacteria and animals. We also observed a possible transfer that might have occurred in sea waters – the HGT prone environment (191). A Cas-like protein (gi|323451003) from 'brown tide' blooms causing algae *Aureococcus anophagefferens* shares homology with a protein from *Planctomyces maris* DSM 8797 (gi|149173160). Both organisms inhabit the sea, which makes possible direct interaction between *A. anophagefferens* and some Planctomyces.

A *Caenorhabditis remanei* protein (gi|308446462) groups with proteins from evolutionary distant species, including *Acinetobacter lwoffii* SH145 (gi|262377357), *A. johnsonii* SH046 (gi|262370624) and many similar *Acinetobacter baumannii* sequences. The transferred protein, a 'Secreted endonuclease distantly related to Holliday junction resolvase' has no function determined yet. *A. baumannii* is a drug resistant pathogen currently considered responsible for major health problems. *Caenorhabditis elegans* is used as a model organism in the studies concerning infection caused by *A. baumannii* (192).

Supplementary Tables, Figures and Datasets

LINK TO SUPPLEMENTARY TABLE S1

Supplementary Table S1 Meta-BASIC pair-wise score matrix for all Pfam, COG, KOG families and PDB90 structures belonging to PD-(D/E)XK superfamily.

Supplementary Table S2 Human genes encoding proteins belonging to the PD-(D/E)XK superfamily.

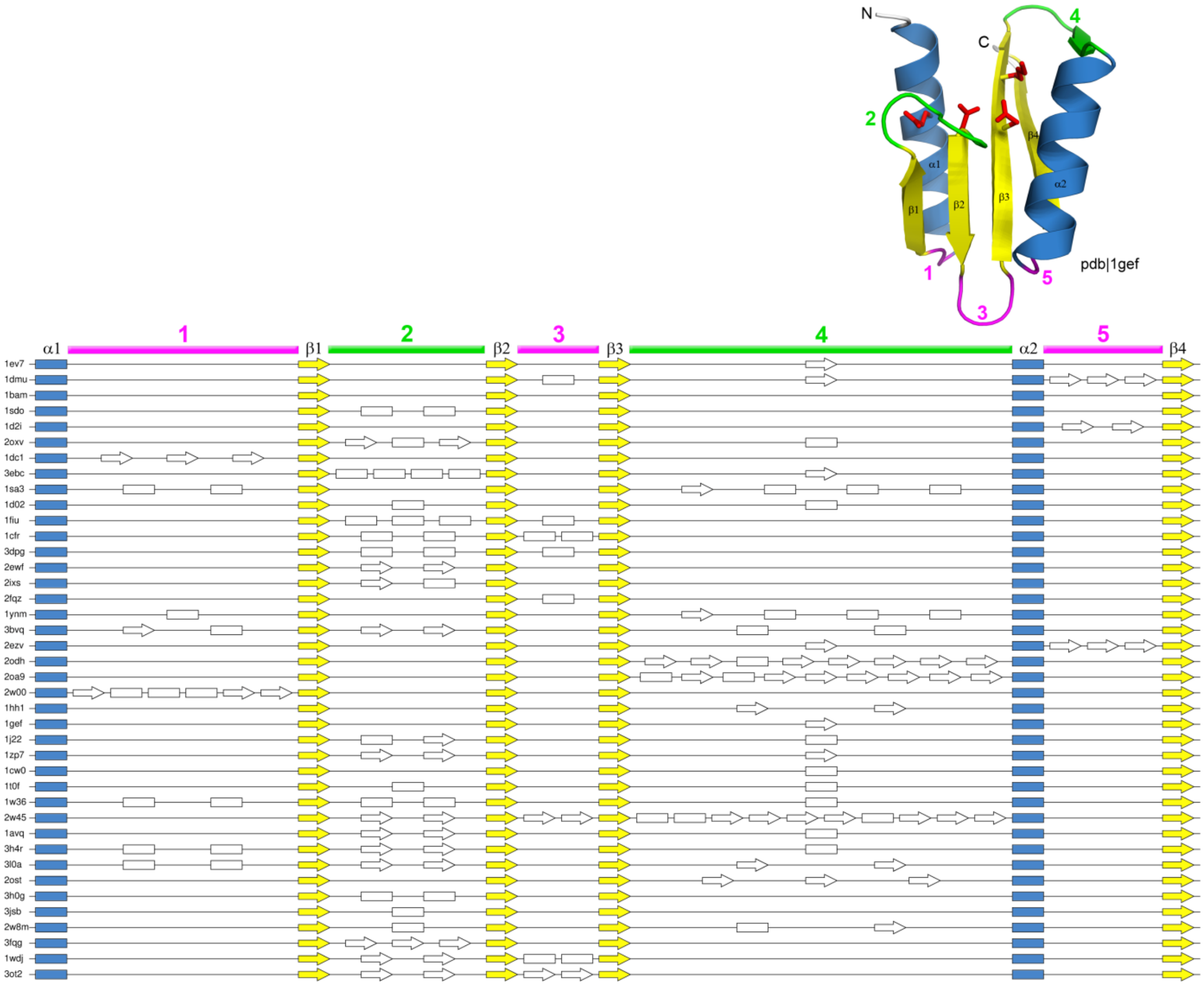
Location	Symbol	Name	Gene features	Gene malfunction effects (OMIM)
1p34.2	DEM1, C1orf176, Exo V	defects in morphology 1 homolog (*S. cerevisiae*)	Defects in morphology protein 1 homolog; probable exonuclease V	-
5p15.31	FASTKD3	FAST kinase domains 3	FAST kinase domain-containing protein 3	-
19q13.32	ERCC1, COFS4, RAD10, UV20	excision repair cross-complementing rodent repair deficiency, complementation group 1	DNA excision repair protein ERCC-1	Cerebrooculofacioskeletal syndrome 4; very severe clinical manifestations, including pre- and postnatal developmental failure and death in early infancy; moderate hypersensitivity to ultraviolet rays and mitomycin C (193)
16p13.12	ERCC4, ERCC11, RAD1, XPF	excision repair cross-complementing rodent repair deficiency, complementation group 4	DNA excision repair protein ERCC-4; DNA repair endonuclease XPF; DNA repair protein complementing XP-F cells; excision-repair, complementing defective, in Chinese hamster; xeroderma pigmentosum group F-complementing protein; xeroderma pigmentosum, complementation group F	Xeroderma pigmentosum group F (194); XFE progeroid syndrome1 (195); cell death and senescence in response to DNA damage
19p13.3	POLR2E, RPABC1, RPB5, XAP4, hRPB25, hsRPB5	polymerase (RNA) II (DNA directed) polypeptide E, 25kDa	DNA directed RNA polymerase II 23 kDa polypeptide; DNA-directed RNA polymerase II 23 kDa polypeptide; DNA-directed RNA polymerase II subunit E; DNA-directed RNA polymerases I, II, and III subunit RPABC1; RNA polymerases I, II, and III subunit ABC1; RPB5 homolog	-
10q21.3-q22.1	DNA2, RP11-9E13.1, DNA2L	DNA replication helicase 2 homolog (*S. cerevisiae*)	DNA replication ATP-dependent helicase-like homolog; DNA2-like helicase	Reduced capacity of mitochondrial DNA to replicate and repair H_2O_2-induced oxidative DNA lesions; reduced cell growth, accumulation of cells in the G2/M phase of the cell cycle, and the appearance of aneuploid cells and internuclear chromatin bridges in human osteosarcoma cells (196)
22q13.1	EIF3D, RP5-1119A7.12-003, EIF3S7, eIF3-p66, eIF3-zeta	eukaryotic translation initiation factor 3, subunit D	eIF-3-zeta; eIF3 p66; eukaryotic translation initiation factor 3 subunit 7; eukaryotic translation initiation factor 3 subunit D; eukaryotic translation initiation factor 3, subunit 7 zeta, 66/67kDa; translation initiation factor eIF3 p66 subunit	-
6p21.3	DOM3Z, DAMC-178C20.2, DOM3L, NG6, RAI1	dom-3 homolog Z (*C. elegans*)	Protein Dom3Z	Rai1-deleted yeast cells with wild type capping enzymes showed significant accumulation of mRNAs with 5' end capping defects under nutritional stress conditions of glucose starvation or amino acid starvation (197)
11q13	MUS81, SLX3	MUS81 endonuclease homolog (*S. cerevisiae*)	SLX3 structure-specific endonuclease subunit homolog; crossover junction endonuclease MUS81	Severe chromosome abnormalities, such that sister chromatids remain interlinked in a side-by-side arrangement and the chromosomes are elongated and segmented (198)
3p25.2	TSEN2, PCH2B, SEN2, SEN2L	tRNA splicing endonuclease 2 homolog (*S. cerevisiae*)	hsSen2; tRNA-intron endonuclease Sen2; tRNA-splicing endonuclease subunit Sen2	Pontocerebellar hypoplasia type 2B (33)
19q13.4	TSEN34, XXbac-BCX105G6.5, LENG5, PCH2C, SEN34, SEN34L	tRNA splicing endonuclease 34 homolog (*S. cerevisiae*)	hsSen34; leukocyte receptor cluster (LRC) member 5; leukocyte receptor cluster member 5; tRNA-intron endonuclease Sen34; tRNA-splicing endonuclease subunit Sen34	Pontocerebellar hypoplasia type 2C (33)
1q25	TSEN15, C1orf19, sen15	tRNA splicing endonuclease 15 homolog (*S. cerevisiae*)	SEN15 homolog; hsSen15; tRNA-intron endonuclease Sen15; tRNA-splicing endonuclease subunit Sen15	-
14q21.2	FANCM, FAAP250, KIAA1596	Fanconi anemia, complementation group M	ATP-dependent RNA helicase FANCM; Fanconi anemia group M protein; fanconi anemia-associated polypeptide of 250 kDa; protein Hef ortholog	Fanconi anemia, complementation group M (35); aberrant DNA repair and cancer predisposition
15q22.2	NARG2, UNQ3101/PRO10100, BRCC1	NMDA receptor regulated 2	NMDA receptor-regulated gene 2; NMDA receptor-regulated protein 2; breast cancer cell 1	-
17q21.33	EME1, MMS4L, SLX2A	essential meiotic endonuclease 1 homolog 1 (*S. pombe*)	MMS4 homolog; SLX2 structure-specific endonuclease subunit homolog A; crossover junction endonuclease EME1; essential meiotic endonuclease 1 homolog 2; hMMS4; homolog of yeast EME1 endonuclease	Increase in chromosomal abnormalities following treatment with the DNA-crosslinking agent mitomycin C (199)
16p13.3	EME2, SLX2B, gs125	essential meiotic endonuclease 1 homolog 2 (*S. pombe*)	SLX2 structure-specific endonuclease subunit homolog B; homolog of yeast EME1 endonuclease 2; probable crossover junction endonuclease EME2	-
15q13.2-q13.3	FAN1, KIAA1018, MTMR15	FANCD2/FANCI-associated nuclease 1	Coiled-coil domain-containing protein MTMR15; fanconi anemia associated nuclease 1; fanconi-associated nuclease 1; myotubularin related protein 15; myotubularin-related protein 15	-
17q25.1	TSEN54, PCH2A, PCH4, SEN54L, sen54	tRNA splicing endonuclease 54 homolog (*S. cerevisiae*)	SEN54 homolog; hsSen54; tRNA-intron endonuclease Sen54; tRNA-splicing endonuclease subunit Sen54	Pontocerebellar hypoplasia type 2A, Pontocerebellar hypoplasia type 4
20p11.23	C20orf72	chromosome 20 open reading frame 72	Gene of unknown function	-
15q14	C15orf41	chromosome 15 open reading frame 41	Gene of unknown function	-

Supplementary Table S3 Results of distant homology detection (Meta-BASIC) for newly identified PD-(D/E)XK families.

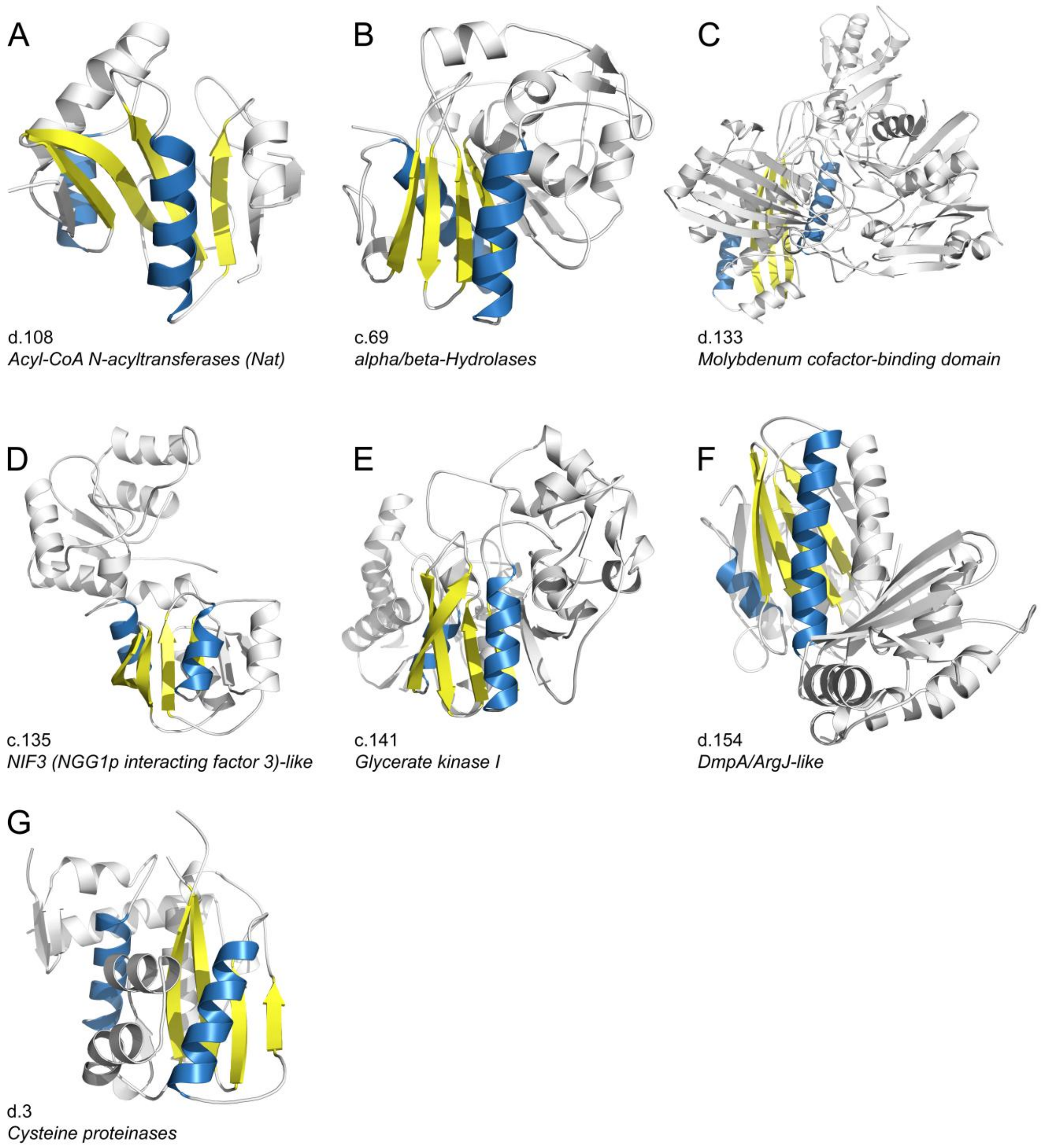
New PD-(D/E)XK family		**First PD-(D/E)XK hit**		
ID	***Description***	***ID***	***Description***	***Meta-BASIC score***
PF09569	ScaI restriction endonuclease	PF04373	DUF511	25.6
PF09553	Eco47II restriction endonuclease	pdb|1wtd	Restriction endonuclease EcoO109I	36.0
PF09554	HaeII restriction endonuclease	PF06317	Arenavirus RNA polymerase	34.3
PF06300	Tsp45I type II restriction enzyme	PF04556	DpnII restriction endonuclease	38.0
PF09561	HpaII restriction endonuclease	PF09208	Restriction endonuclease MspI	34.1
COG5482	Uncharacterized conserved protein	PF04373	DUF511	46.5
COG1395	Predicted transcriptional regulator	PF01870	Archaeal holliday junction resolvase (hjc)	33.1
PF14082	DUF4263	PF01939	DUF91	60.3
PF13020	DUF3883	pdb|1hh1	Archaeal HJC resolvase	37.1
PF14390	DUF4420	PF13635	DUF4143	23.6
PF13814	Replication-relaxation	COG4636	Uncharacterized protein conserved in cyanobacteria	38.1

LINK TO SUPPLEMENTARY TABLE S4

Supplementary Table S4 Brief summary of additional domains co-occurring with PD-(D/E)XK domain.



Supplementary Figure S1 Insertions to the commonly conserved core of the PD-(D/E)XK fold. Secondary structure elements forming the core are colored blue (α-helices) and yellow (β-strands). Locations of insertions placed in the proximity of the active site and on the opposite side of the protein are denoted in green and magenta, respectively.



d.108
Acyl-CoA N-acyltransferases (Nat)

c.69
alpha/beta-Hydrolases

d.133
Molybdenum cofactor-binding domain

c.135
NIF3 (NGG1p interacting factor 3)-like

c.141
Glycerate kinase I

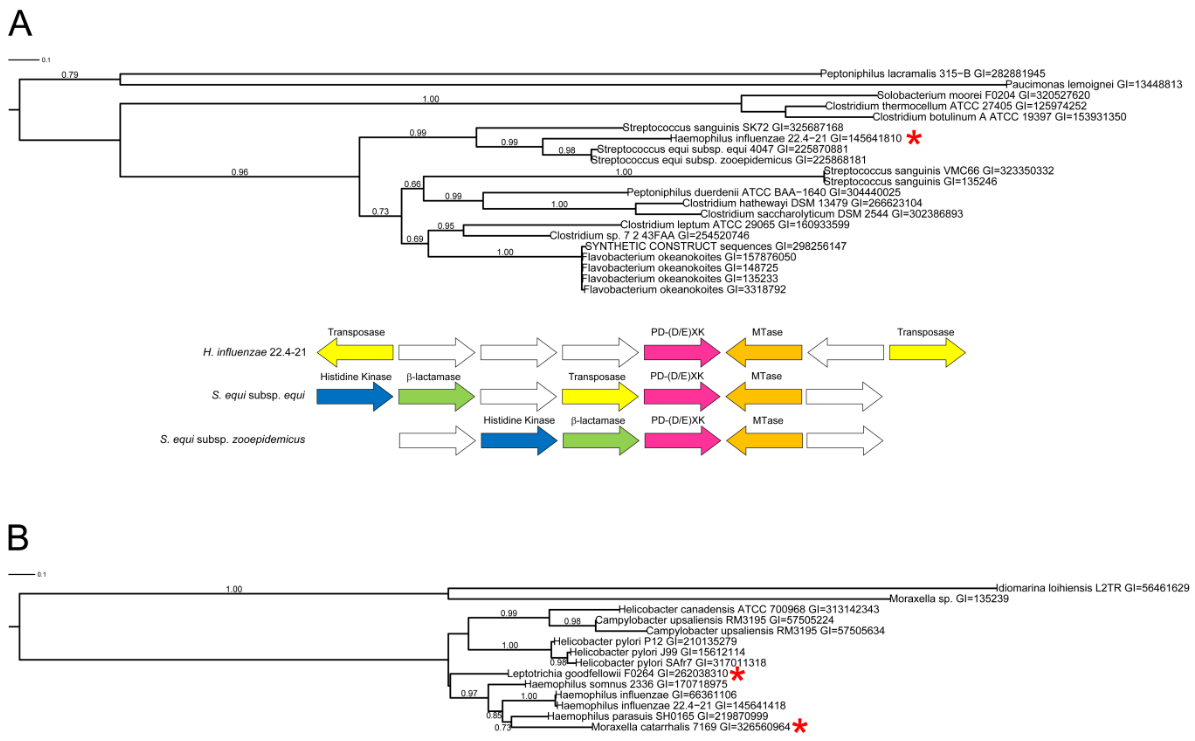
d.154
DmpA/ArgJ-like

d.3
Cysteine proteinases

Supplementary Figure S2 Unrelated structures retaining the PD-(D/E)XK nuclease fold as a substructure. Secondary structure elements corresponding to PD-(D/E)XK nuclease fold core are colored yellow (β-strands) and blue (α-helices). (A) *Ovis aries* N-acetyltransferase (pdb|1cjw); (B) *Thermomyces lanuginosus* lipase (pdb|1ein); (C) *Hydrogenophaga pseudoflava* carbon monoxide dehydrogenase (pdb|1ffu); (D) *Escherichia coli* hypothetical protein ybgI (pdb|1nmo); (E) *Neisseria meningitides* glycerate kinase (pdb|1to6); (F) *Streptomyces clavuligerus* ornithine acetyltransferase (pdb|1vz6); (G) *Homo sapiens* NEDD8 protease (pdb|2bkq).

LINK TO SUPPLEMENTARY FIGURE S3

Supplementary Figure S3 Domain architectures identified within proteins having PD-(D/E)XK domain. Presented architectures are unique within the group. Groups and sequences lacking domains other than PD-(D/E)XK are omitted. Protein domains are colored as follows: PD-(D/E)XK domain, in red; nucleic acid interacting domain, in blue; methylase, in green; internal repeat, in yellow and square-shaped; transferase, in light pink; kinase, in dark green; helicase, in yellow; ATPase, in cyan; coiled-coils, small rectangles in pink; transmembrane elements, in dark pink and narrow rectangle-shaped.


A
0.1
0.79
Peptoniphilus lacramalis 315−B GI=282881945
Paucimonas lemoignei GI=13448813
1.00
Solobacterium moorei F0204 GI=320527620
Clostridium thermocellum ATCC 27405 GI=125974252
Clostridium botulinum A ATCC 19397 GI=153931350
0.99
Streptococcus sanguinis SK72 GI=325687168
0.99
Haemophilus influenzae 22.4−21 GI=145641810 *
0.98
Streptococcus equi subsp. equi 4047 GI=225870881
Streptococcus equi subsp. zooepidemicus GI=225868181
0.96
1.00
Streptococcus sanguinis VMC66 GI=323350332
Streptococcus sanguinis GI=135246
0.66
0.99
Peptoniphilus duerdenii ATCC BAA−1640 GI=304440025
1.00
Clostridium hathewayi DSM 13479 GI=266623104
Clostridium saccharolyticum DSM 2544 GI=302386893
0.73
0.95
Clostridium leptum ATCC 29065 GI=160933599
0.69
Clostridium sp. 7 2 43FAA GI=254520746
1.00
SYNTHETIC CONSTRUCT sequences GI=298256147
Flavobacterium okeanokoites GI=157876050
Flavobacterium okeanokoites GI=148725
Flavobacterium okeanokoites GI=135233
Flavobacterium okeanokoites GI=3318792
Transposase
PD-(D/E)XK
MTase
Transposase
H. influenzae 22.4-21
Histidine Kinase
β-lactamase
Transposase
PD-(D/E)XK
MTase
S. equi subsp. equi
Histidine Kinase
β-lactamase
PD-(D/E)XK
MTase
S. equi subsp. zooepidemicus
B
0.1
1.00
Idiomarina loihiensis L2TR GI=56461629
Moraxella sp. GI=135239
0.99
Helicobacter canadensis ATCC 700968 GI=313142343
0.98
Campylobacter upsaliensis RM3195 GI=57505224
Campylobacter upsaliensis RM3195 GI=57505634
1.00
Helicobacter pylori P12 GI=210135279
Helicobacter pylori J99 GI=15612114
0.98
Helicobacter pylori SAfr7 GI=317011318
Leptotrichia goodfellowii F0264 GI=262038310 *
Haemophilus somnus 2336 GI=170718975
0.97
1.00
Haemophilus influenzae GI=66361106
Haemophilus influenzae 22.4−21 GI=145641418
0.85
Haemophilus parasuis SH0165 GI=219870999
0.73
Moraxella catarrhalis 7169 GI=326560964 *

Supplementary Figure S4 HGT in FokI (A) and HinP1I (B) restriction endonucleases. The phylogenetic trees were calculated in PhyML with the following settings: LG model, gamma estimated, 5 random starting trees, aLTR branch support (en estimate for topology likelihood, given above the corresponding branches). Tree images were prepared using in iTol. The asterisks point at probable horizontally transferred sequences. Schematic representation of the genomic context (analyzed using MicrobesOnline and The SEED) for FokI nucleases is shown with colored arrows.

LINK TO SUPPLEMENTARY DATASET S1

Supplementary Dataset S1 A list of all 21 911 identified PD-(D/E)XK superfamily sequences (3 iterations of PSI-BLAST for each family; NCBI gene identifiers).

LINK TO SUPPLEMENTARY DATASET S2

Supplementary Dataset S2 Phylogenetic trees of 118 protein clans belonging to the PD-(D/E)XK superfamily. Protein alignments were generated with MAFFT (linsi flavor, 500 iterations) and trimmed with trimAl. The trees were calculated in PhyML with the following settings: LG model, gamma estimated, 5 random starting trees, aLTR branch support. The zipped archive contains 118 trees in XML format, rooted, pruned to 0.5 branch support with annotated taxonomy.

Supplementary References

164. Jones, D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol*, **292**, 195-202.

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

166. Capella-Gutierrez, S., Silla-Martinez, J.M. and Gabaldon, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.

167. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*, **59**, 307-321.

168. Han, M.V. and Zmasek, C.M. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.

169. Letunic, I. and Bork, P. (2011) Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res*, **39**, W475-478.

170. Niv, M.Y., Ripoll, D.R., Vila, J.A., Liwo, A., Vanamee, E.S., Aggarwal, A.K., Weinstein, H. and Scheraga, H.A. (2007) Topology of Type II REases revisited; structural classes and the common conserved core. *Nucleic Acids Res*, **35**, 2227-2237.

171. Brzozowski, A.M., Savage, H., Verma, C.S., Turkenburg, J.P., Lawson, D.M., Svendsen, A. and Patkar, S. (2000) Structural origins of the interfacial activation in Thermomyces (Humicola) lanuginosa lipase. *Biochemistry*, **39**, 15071-15082.

172. Ladner, J.E., Obmolova, G., Teplyakov, A., Howard, A.J., Khil, P.P., Camerini-Otero, R.D. and Gilliland, G.L. (2003) Crystal structure of Escherichia coli protein ybgI, a toroidal structure with a dinuclear metal site. *BMC Struct Biol*, **3**, 7.

173. Kotera, E., Tasaka, M. and Shikanai, T. (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature*, **433**, 326-330.

174. Nakayama, S., Moncrief, N.D. and Kretsinger, R.H. (1992) Evolution of EF-hand calcium-modulated proteins. II. Domains of several subfamilies have diverse evolutionary histories. *J Mol Evol*, **34**, 416-448.

175. Tan, B.Z., Huang, H., Lam, R. and Soong, T.W. (2009) Dynamic regulation of RNA editing of ion channels and receptors in the mammalian nervous system. *Mol Brain*, **2**, 13.

176. Lommel, M. and Strahl, S. (2009) Protein O-mannosylation: conserved from bacteria to humans. *Glycobiology*, **19**, 816-828.

177. Nakatsukasa, K., Okada, S., Umebayashi, K., Fukuda, R., Nishikawa, S. and Endo, T. (2004) Roles of O-mannosylation of aberrant proteins in reduction of the load for endoplasmic reticulum chaperones in yeast. *J Biol Chem*, **279**, 49762-49772.

178. Coste, K. and Labbe, A. (2011) [A study of the metabolic pathways of vitamin A in the fetal human lung]. *Rev Mal Respir*, **28**, 283-289.

179. Chuang, Y.S., Huang, W.H., Park, S.W., Persaud, S.D., Hung, C.H., Ho, P.C. and Wei, L.N. (2011) Promyelocytic leukemia protein in retinoic acid-induced chromatin remodeling of Oct4 gene promoter. *Stem Cells*, **29**, 660-669.

180. Wang, P. and Heitman, J. (2005) The cyclophilins. *Genome Biol*, **6**, 226.

181. Laskey, R.A., Honda, B.M., Mills, A.D. and Finch, J.T. (1978) Nucleosomes are assembled by an acidic protein which binds histones and transfers them to DNA. *Nature*, **275**, 416-420.

182. Kussie, P.H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A.J. and Pavletich, N.P. (1996) Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science*, **274**, 948-953.

183. Meyn, M.S., Lu-Kuo, J.M. and Herzing, L.B. (1993) Expression cloning of multiple human cDNAs that complement the phenotypic defects of ataxia-telangiectasia group D fibroblasts. *Am J Hum Genet*, **53**, 1206-1216.

184. Aminov, R.I. (2011) Horizontal gene exchange in environmental microbiota. *Front Microbiol*, **2**, 158.

185. Hogg, J.S., Hu, F.Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J.C. and Ehrlich, G.D. (2007) Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol*, **8**, R103.

186. Gupta, N., Arora, S. and Kundra, S. (2011) Moraxella catarrhalis as a respiratory pathogen. *Indian J Pathol Microbiol*, **54**, 769-771.

187. Richards, T.A., Soanes, D.M., Jones, M.D., Vasieva, O., Leonard, G., Paszkiewicz, K., Foster, P.G., Hall, N. and Talbot, N.J. (2011) Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc Natl Acad Sci U S A*, **108**, 15258-15263.

188. Zhu, B., Lou, M.M., Xie, G.L., Zhang, G.Q., Zhou, X.P., Li, B. and Jin, G.L. (2011) Horizontal gene transfer in silkworm, Bombyx mori. *BMC Genomics*, **12**, 248.

189. Haegeman, A., Jones, J.T. and Danchin, E.G. (2011) Horizontal gene transfer in nematodes: a catalyst for plant parasitism? *Mol Plant Microbe Interact*, **24**, 879-887.

190. Sun, G., Yang, Z., Kosch, T., Summers, K. and Huang, J. (2011) Evidence for acquisition of virulence effectors in pathogenic chytrids. *BMC Evol Biol*, **11**, 195.

191. McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B. and Paul, J.H. (2010) High frequency of horizontal gene transfer in the oceans. *Science*, **330**, 50.

192. Peleg, A.Y., Tampakakis, E., Fuchs, B.B., Eliopoulos, G.M., Moellering, R.C., Jr. and Mylonakis, E. (2008) Prokaryote-eukaryote interactions identified by using Caenorhabditis elegans. *Proc Natl Acad Sci U S A*, **105**, 14585-14590.

193. Jaspers, N.G., Raams, A., Silengo, M.C., Wijgers, N., Niedernhofer, L.J., Robinson, A.R., Giglia-Mari, G., Hoogstraten, D., Kleijer, W.J., Hoeijmakers, J.H. *et al.* (2007) First reported patient with human ERCC1 deficiency has cerebro-oculo-facio-skeletal syndrome with a mild defect in nucleotide excision repair and severe developmental failure. *Am J Hum Genet*, **80**, 457-466.

194. Sijbers, A.M., van Voorst Vader, P.C., Snoek, J.W., Raams, A., Jaspers, N.G. and Kleijer, W.J. (1998) Homozygous R788W point mutation in the XPF gene of a patient with xeroderma pigmentosum and late-onset neurologic disease. *J Invest Dermatol*, **110**, 832-836.

195. Niedernhofer, L.J., Garinis, G.A., Raams, A., Lalai, A.S., Robinson, A.R., Appeldoorn, E., Odijk, H., Oostendorp, R., Ahmad, A., van Leeuwen, W. *et al.* (2006) A new progeroid syndrome reveals that genotoxic stress suppresses the somatotroph axis. *Nature*, **444**, 1038-1043.

196. Duxin, J.P., Dao, B., Martinsson, P., Rajala, N., Guittat, L., Campbell, J.L., Spelbrink, J.N. and Stewart, S.A. (2009) Human Dna2 is a nuclear and mitochondrial DNA maintenance protein. *Mol Cell Biol*, **29**, 4274-4282.

197. Jiao, X., Chen, H., Chen, J., Herrup, K., Firestein, B.L. and Kiledjian, M. (2009) Modulation of neuritogenesis by a protein implicated in X-linked mental retardation. *J Neurosci*, **29**, 12419-12427.

198. Wechsler, T., Newman, S. and West, S.C. (2011) Aberrant chromosome morphology in human cells defective for Holliday junction resolution. *Nature*, **471**, 642-646.

199. Abraham, J., Lemmers, B., Hande, M.P., Moynahan, M.E., Chahwan, C., Ciccia, A., Essers, J., Hanada, K., Chahwan, R., Khaw, A.K. *et al.* (2003) Eme1 is involved in DNA damage processing and maintenance of genomic stability in mammalian cells. *EMBO J*, **22**, 6137-6147.