# Phosphoesterase domains associated with DNA polymerases of diverse origins

#### L. Aravind<sup>1,2</sup> and Eugene V. Koonin<sup>2,\*</sup>

<sup>1</sup>Department of Biology, Texas A&M University, College Station, TX 77843, USA and <sup>2</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received April 20, 1998; Revised and Accepted June 25, 1998

#### ABSTRACT

Computer analysis of DNA polymerase protein sequences revealed previously unidentified conserved domains that belong to two distinct superfamilies of phosphoesterases. The  $\alpha$  subunits of bacterial DNA polymerase III and two distinct family X DNA polymerases are shown to contain an N-terminal domain that defines a novel enzymatic superfamily, designated PHP, after polymerase and histidinol phosphatase. The predicted catalytic site of the PHP superfamily consists of four motifs containing conserved histidine residues that are likely to be involved in metal-dependent catalysis of phosphoester bond hydrolysis. The PHP domain is highly conserved in all bacterial polymerase III  $\alpha$  subunits, but in proteobacteria and mycoplasmas, the conserved motifs are distorted, suggesting a loss of the enzymatic activity. Another conserved domain, found in the small subunits of archaeal DNA polymerase II and eukaryotic DNA polymerases  $\alpha$  and  $\delta$ , is shown to belong to the superfamily of calcineurin-like phosphoesterases, which unites a variety of phosphatases and nucleases. The conserved motifs required for phosphoesterase activity are intact in the archaeal DNA polymerase subunits, but are disrupted in their eukaryotic orthologs. A hypothesis is proposed that bacterial and archaeal replicative DNA polymerases possess intrinsic phosphatase activity that hydrolyzes the pyrophosphate released during nucleotide polymerization. As proposed previously, pyrophosphate hydrolysis may be necessary to drive the polymerization reaction forward. The phosphoesterase domains with disrupted catalytic motifs may assume an allosteric, regulatory function and/or bind other subunits of DNA polymerase holoenzymes. In these cases, the pyrophosphate may be hydrolyzed by a stand-alone phosphatase, and candidates for such a role were identified among bacterial PHP superfamily members.

#### INTRODUCTION

DNA polymerases are central to genome replication and repair (1-3). In spite of their fundamental role, these enzymes are very

different among different life forms (4). DNA polymerases have been classified on the basis of sequence similarity between their large subunits or domains that catalyze nucleotide polymerization. There are at least four distinct DNA polymerase superfamilies that bear little, if any, relationship to each other, namely: (i) bacterial DNA polymerase I and its eukaryotic orthologs; (ii) B family DNA polymerases including eukaryotic and archaeal replicative and repair polymerases, polymerases of animal viruses and bacteriophages, and Escherichia coli polymerase II; (iii) bacterial replicative DNA polymerase III; and (iv) family X polymerases, which includes terminal nucleotidyltransferases, several polymerases involved in repair, and a variety of other diverse nucleotidyltransferases (1-3 and references therein). Typically, in addition to the subunits or domains directly involved in polymerization, DNA polymerase holoenzymes contain additional subunits (domains) that perform accessory functions in replication or repair (1,2,5).

Here we present the results of detailed computer analysis of the sequences of DNA polymerase subunits. We show that three distinct groups of DNA polymerases, namely the  $\alpha$ -subunit of bacterial DNA polymerase III, family X polymerases and the small subunits of archaeal DNA polymerases and eukaryotic polymerases  $\alpha$ ,  $\delta$  and  $\epsilon$ , contain previously undetected, conserved domains. These domains belong to two distinct, ancient superfamilies of enzymes with phosphatase, and in some cases nuclease, activity. One of these superfamilies, whose only functionally characterized member is yeast histidinol phosphatase, has not been recognized previously. The association of the phosphoesterase domains with DNA polymerases clearly is the result of at least two independent evolutionary events, which suggests an important role for these domains in the polymerase function. We hypothesize that this primary role is the hydrolysis of the released pyrophosphate, which shifts the reaction equilibrium towards nucleotide polymerization.

#### MATERIALS AND METHODS

#### **Databases and sequence analysis**

The non-redundant (NR) protein sequence database at the National Center for Biotechnology Information (NIH, Bethesda) was searched using the gapped BLAST program (6). Database searches were iterated using the PSI-BLAST program, which constructs position-dependent weight matrices on the basis of alignments generated by BLAST and employs them for subsequent

\*To whom correspondence should be addressed. Tel: +1 301 453 5913; Fax: +1 301 480 9241; Email: koonin@ncbi.nlm.nih.gov

search iterations (6). In addition to the automatic generation of profiles 'on the fly', a new option of the PSI-BLAST program was used to generate matrices from alignments constructed independently and seed PSI-BLAST iterations with these (A.Schäffer, L.Aravind and E.V.Koonin, unpublished work). Multiple sequence alignments were constructed using the CLUS-TALW program (7) or the Gibbs-sampling option of the MACAW program (8.9). The program MoST was used for searching the NR database with motifs represented as ungapped alignment blocks (10). Clustering of sequences by sequence similarity was performed using the CLUS program (11). Protein secondary structure was predicted using the PHD program (12), and the PHD-based threading was used to assess possible relationships between aligned protein sequences and known structural folds (13). Globular domains in proteins were predicted using the SEG program with the following parameters: window length 45, trigger complexity 3.4, extension complexity 3.75 (14).

#### **RESULTS AND DISCUSSION**

### PHP domain: a novel family of phosphoesterases associated with polIII $\alpha$ subunit and X-family polymerases

DNA polymerase III  $\alpha$  subunits from bacteria and the DNA polymerases of the X family from Methanobacterium thermoautotrophicum, Aquifex aeolicus and Bacillus subtilis (but not the homologous DNA polymerase from Thermus aquaticus) showed statistically significant sequence similarity to each other when the NR database was searched with any of these sequences as the query using the gapped BLAST program [expectation (e) value:  $\sim 10^{-3} - 10^{-4}$ ]. This was an unexpected finding given that these DNA polymerases belong to two different superfamilies (3). Examination of the distribution of the detected similarity among the predicted globular domains in the two families of polymerases showed that the C-terminal domain of the B.subtilis and the archaeal X family DNA polymerase was related to a domain typically found at the extreme N-terminus of the DNA polymerase III  $\alpha$  subunit. This domain was clearly distinct from the well-defined nucleotidyltransferase domain of the X family (15,16) and the proposed functionally equivalent catalytic region of the polIII  $\alpha$  subunits (17), which do not contain any detectable common motifs. This novel domain shared by the two diverse families of polymerases was subjected to further iterative searches using the PSI-BLAST program, which was run to convergence. These searches resulted in the delineation of a previously uncharacterized superfamily of domains that occur both as stand-alone proteins in a variety of bacterial and archaeal genomes and as fusions with the polymerase domains (Fig. 1). The internal consistency of the derived protein set was assessed by carrying out multiple PSI-BLAST searches with different members of the superfamily and evaluating recovery of the other members at e-value cut-offs below 10-4 within four iterations. This helped to unambiguously define the superfamily as different starting queries retrieved the rest of the superfamily members at a statistically significant level. Additionally, the motif searching program MoST was used to scan the non-redundant database with weighted profiles of the motifs derived from the PolIII  $\alpha$  subunit version of the domain with r values of 0.005 and 0.001. This resulted in consistent recovery of the rest of the superfamily without any additions to the set defined by the multiple PSI-BLAST analysis.

As shown by multiple alignment analysis using a combination of PSI-BLAST derived alignments and Gibbs sampling as

incorporated in the MACAW program, the polymerase and histidinol phosphatase (PHP) domain described here consists of four conserved core regions (motifs I-IV in Fig. 1). The characteristic feature of these motifs is the presence of highly conserved histidines and aspartates, which are likely to participate in catalysis. Motif I has a dyad of histidines which are separated by a single amino acid. A similar arrangement is also observed in two other groups of metal-dependent hydrolases, namely the urease-pyrimidinase superfamily (18) and the calcineurin-like phosphoesterase superfamily (19; see also below), with the conserved histidines apparently coordinating a metal ion such as zinc or nickel. Motifs II-IV contain additional conserved histidines and aspartates (Fig. 1), which may be involved in catalysis by participating in electron transfer and/or through metal coordination. Motif IV is missing in some archaeal and bacterial proteins (Fig. 1), suggesting the possibility that PHP family proteins may differ in the number of coordinated metal ions. The nature and distribution of the conserved residues in the PHP superfamily is analogous to the urease-pyrimidinase superfamily (18). Thus it seems likely that similarly to the enzymes of this superfamily, the activity of the PHP phosphoesterases requires two coordinated metal cations. However, database searches using profiles derived from multiple alignments of each of these superfamilies failed to detect any similarity between them, and furthermore, the distance between motifs I and II in the PHP superfamily is strictly constrained and is much shorter than the distance between the analogous motifs in the urease-pyrimidinase superfamily (18). Thus, in spite of the resemblance between the conserved motifs, which may reflect similar aspects of the catalytic mechanisms, detailed sequence comparisons did not reveal a specific relationship between these two superfamilies of metaldependent hydrolases. Multiple alignment-based database threading did not yield any significant hits for the PHP superfamily, but interestingly, the best hit, albeit with a relatively low score, was urease (2kau), which has a distorted TIM barrel fold and belongs to the urease-pyrimidinase superfamily (18). The relevance of this observation is hard to ascertain. It may indicate that the PHP superfamily has a distinct version of the TIM barrel fold, but a novel fold clearly remains a possibility.

The evidence for the phosphoesterase activity of the PHP domain comes from the yeast histidinol phosphatase encoded by the his-2 gene (20,21). This enzyme is involved in the ninth step of histidine biosynthesis in yeast and cleaves the phosphomonoester bond in histidinol phosphate to release histidinol, which is then reduced to histidine by the his-4c gene product (20). Several mutations resulting in a histidine-requiring phenotype have been identified in the his-2 gene (21), and interestingly, two of them map to the conserved motifs described above. The his2-6 mutation affects motif I and is situated directly N-terminal of the conserved HXH signature, whereas the his2-390 mutation affects the conserved histidine in motif 3 (Fig. 1). The mutation data are compatible with the conclusion that these conserved motifs are required for the enzyme function.

Using the scores produced by the gapped BLAST program as a similarity measure, the PHP superfamily was clustered into five distinct protein families, which may include phosphoesterases with distinct functions (Figs 1 and 2): 1) N-terminal domains of the bacterial DNA polIII  $\alpha$  subunits, which consist of ~210 amino acids; 2) a group of stand-alone proteins from bacteria and archaea, which vary in size from 154 (so far the minimal size of the PHP domain) to 274 residues and show greater similarity to the polIII N-terminal domains than other members of the super-

COMBENES         Implif. 10: hs				MOTIF	I		MOTI	F II			MOTIF	III		MOTIF IV		
DNAROLITI, Ase_293510         21         FUELTON, TOTAL REGISTON, AND SERVICE AND SER	Consensus		.hphH	.Ho.hs.			.hhs	.hh	sho-H			н		shssD.H		
DP3AL_BE_0261         314         KVLLIUGEREQUENY         4 KAZEGAKKORIERAKUTU GAVY         24 GOSNY GOSNY GASRYY MATLAVNS         101         LGEKK KAFYV ARGENTELAREBER           DP3AL_BE_0263427         1         NGCO CITERS NILL         1         LLEKK YAVAREK KOLAY DU KASY SULDA         1         LLEKK YAVAR YAKASY SULDA         1         LLEKK YAVAR YAKASY SULDA         1         LLEKK YAVAR YAKASY SULDA         1         LLEKK YAVAYAKASY SULDA         LLEKK YAVAKASKASY SULDA	DNAPOLIII_Aae_2983510	21	FVHLH	L <mark>H</mark> TO <b>FS</b> I	LLDGA	4	ELVKKAKEY	G <mark>y</mark> ka <mark>v</mark>	GMSDHGNLF	34	KTSEDNITD	YNHHLILIAKD	96	IAKKYDVKLIATQDAHYLNE	EDRY \	
DP3A3_Be_063427         1         MICOL OF DEGRUREL         7         ELEKTAV/REPKOURS OF DELARADOR         11.1         IANOS LOGITAR SALADOR         11.1         11.1	DP3A1_Bs_80261	334	RVELH	L <mark>H</mark> S PMS(	CMDAV	- 4	KLVEQAKKW	GHEAI	ALT DHAVVQ	224	GOSNAYORSF	RPY <mark>H</mark> ATLLAVNS	101	LGEKLNKFVVATONVHYLNE	EDKI	
DRAPOLITI_E_E_2494191         ? FTRE.U PESILDO. 4 ILLARK VERLER KENTTE CIMP 2         OBLESTETKORELCLEARN 101 RACE IT NOT THENDAK           DRAPOLITI_E_E_2688500         21 FTRE.V STRVELLO. 4 DITSK XKCNNSTRINTE CIMP 2         OBLEKTERVERREN 101 LARKEN 195 YEREL VERLEN 11 ALLARKEN 195 YEREL VERLET 195 ALLARKEN 195 YEREL 195 YEREL 195 ALLARKEN 195 YEREL 195	DP3A3_Bs_2634427	1	MIGCH	CHTDRS	VIRLL	7	ELLKTAVKM	e <mark>v</mark> kg <mark>l</mark>	ITOHEVLS	39	RDNYKGGVTP	FPHFLMLAMDP	118	IANGYDLQMIVTTDAHYLRP	EDRA	
DRAPDIII_Esp_1633960         3 FORD USERVICES ALLOS 4         A EXTERATION CONF.         0 6 OTELE KCHERNEQUVILARNE 5         100 IAQUE LITIVATE OSCIETESCHDER           DRAPDIII_C         3 VETRU OSCIENTA         4 DISTAKCHERNEGUVILARNE 5         5 VERLU VETRU OSCIENTA         1           DRAPDIII_C         3 VETRU OSCIENTA         4 KEVARAVENDES COLLARS         DEKKERRAVANDILLICKE         15 TESKEL VETERANDE OSCIENTA         1           DRAPDIIII_C         3 VETRU OSCIENTA         4 KEVARAVENDES COLLARS         DEKKERRAVANDILLICKE         15 TESKEL VETERANDE OSCIENTA         1           DRAPDIIIIC_NU STORELDOS         4 REAVERARLOGICALMICO DENEY         20 DEKKERRAVANDILLICKE         15 TESKEL VETERANDE         16 TESKEL VETERANDE         17 TESKEL VETERANDE         16 TESKEL VETERANDE         17 TESKEL VETERANDE         17 TESKEL VETERANDE         16 TESKEL VETERANDE         17 TESKEL VETERANDE	DNAPOLIII_Hp_2494191	7	FTHLH	LHTEYSI	LLDGA	4	ILAKRVKEL	GMKSV	SVTDHGNMF	27	DNLSSKETK	RFHLCLFAKNQ	105	MSLETGLKIIATNOTHYTMP	NDAK	
DRAPOLITI_Eb.2688500         21         FTHE WERDSELLOC         4         DTESKERKENNELLAKEN         95         YEREE_WELTENK GENVELLOK         1           DRAPOLITI_C         3         VIET (SOUSTLAND         4         KYMARYSYCERLAKING GULF         15         TESKEE UFFENANCE         15 <t< td=""><td>DNAPOLIII_Ssp_1653960</td><td>3</td><td>FVGLH</td><td>IHSDYSI</td><td>LLDGA</td><td>4</td><td>ALIDRAIEL</td><td>GMPAI</td><td>ALTDHGVMY</td><td>26</td><td>GDIEINKRHF</td><td>REFIQVVLAKNN</td><td>100</td><td>IAQELDIKIVATNDSHFISC</td><td>NDVE</td><td></td></t<>	DNAPOLIII_Ssp_1653960	3	FVGLH	IHSDYSI	LLDGA	4	ALIDRAIEL	GMPAI	ALTDHGVMY	26	GDIEINKRHF	REFIQVVLAKNN	100	IAQELDIKIVATNDSHFISC	NDVE	
DNAPOLITI_C         3 WEITE OBSOLUTION         4 KEVAK OVEYSOLPALEAD COLLAR         30 DKKEEKKEVANNULLILICKO         115 TSKELET FSVAR OLTVART SALTUNEN           DNAPOLITIL_D         21 FUNCTIONSNYLLANG         4 KEVAK FKSLOQEALANG ONGP 30 DKSEETGRAFTILVIJVIJANS         95 LARCE IFLAND         50 KARCE IFLAND           DNAPOLITIL_NUL_JOGENG         4 KALVER KSLOQEALANG OLANG OLANG         20 KARCE IFLAND         95 LARCE IFLAND         50 KARCE IFLAND           DP3A.C.11874         6 FUNCTORS OLANG         4 KEVERGUNKASALALGO DUNAF         2 SVPTUBSELEANTLULIANS         91 LSEED IF VIANG ONTTRANCH THRANK           DP3A.K.118794         6 FUNCTORS OF THRANG         4 PUNKAAANGHVARILO FUNCTASAN         2 SVPTUBSELEANTLULIANS         91 LSEED IF VIANG ONTTRANCH THRANG           DP3A.K.118794         7 FUNCTORS OF THRANG         4 PUNKAAANGHVARILO FUNCTASAN         91 LSEED IF VIANG ONTTARS OF THRANG           DP32.MD.143777         4 KEVAGENAANGHVARILO FUNCTOR ONTTA'         2 KVASPLCOBENTELILIANTEN         91 LARCE OLANG ONTO ONTTA'           2105045_Sgp_1001243         5 FUNCTOR DOCHT         2 GUNAF VENTRO ONTTA'         10 CLAGGAVPES SARTANCE         91 LARCE OLANG ONTO ONTA'           2105045_Sgp_1001243         5 FUNCTOR DOCHT         2 GUNAF VENTRO ONTA'         10 LARGE OLANG ONTO ONTA'         10 LARGE OLANG ONTO ONTA'           2105042_Sgp_1001243         5 FUNCTOR DOCHT         2 GUNET VARGE ONTO ONTA'         10 LARGE OLANG	DNAPOLIII_Bb_2688500	21	FIHLH	VHSDYSI	LLDGA	4	DIISKAKKC	NMSHI	ALT DHGNLF	30	LKKQDDLGKN	SYHLILLAKNE	95	YSRELGVPLTAANDSHYVNR	EDAT	1
DNAPOLILI_PP         21         STULTUSENTICLEDA         4 ENVETEXELOGALALINGTODEL'         20         DESENTICOREYULULAREN         96         LACKIGITEUVIDE         Mathematical           DNAPOLILI, MUL_JOGAGE         10         STULTUSANSTELLOA         4 ENVETEXELATION DELVIDE         21         STULTUSANSTELLUSANSTELLOA         4         ENVERDICIAL         70         10         ENVELUSANSTELLOA         4         ENVERDICIAL         70         10         ENVELUSANSTELLOA         4         ENVERDICIAL         70         10         ENVELUSANSTELLOA         4         ENVERSION         4         LAESTONE         70         10         ENVELUSANSTELLOA         4         ENVERSION         21         LAESTONE         70         ENVERSION         70         ENVERSION         70         ENVERSION         20         CONCINTUMENTS         20         FUNCTURE         21         FUNCTURE         20         FUNCTURE         FUNCTURE         FUNCTURE         20         FUNCTURE         20         FUNCTURE         FUNCTURE         FUNCTURE         FUNCTURE         FUNCTURE         FUNCTURE	DNAPOLIII_Ct	3	WIPLH	CHSQYSI	LDAT	4	K <b>FV</b> AKAVEY	QIPAL	ALT DHGNLF	30	DKKKERKSRV	ANHLILLCKDE	115	TSKRLGIPSVATNOIHYLNP	DDWL	
DNAPOLITI2_MUL_2661649 46 YAELMLERAKETLOGA 4 ELVEERAKLOCCALINGTODENY 26 TARTERDPROFILLULARGF 98 LAPREWGINTAGEFADESGR DNAPOLITIX_UL_17064913 10 FUELUMETENSIONA 4 FEASTREENGENDENY 22 SUPERISELEAYTEVLLARGF 97 IGRAITERATEGYNEN DP3A_E_118794 67 VUELUMETENSIONA 4 ELVERAKLOKASLINGTONY 22 SUPERISELEAYTEVLLAKSN 91 IGEETGIFVTANONYTRENAM DP3A_E_118794 77 THERESON L4 FUEXKAALAHANDAUTOTING 22 SUPERISELEAYTEVLLAKSN 95 LAERGUFFVATENYTRESSON DP3A_H_1165392 7 FINERYEDENHIDSI 4 FUEXKAALAHANDAUTOTING 22 SUKSPLGGERYEDITLAKSN 95 LAERGUFFVATENYTRESSON DP3A_M_115777 4 KIDILOUGUTATEDENHIDSI 4 FUEXKAALAHANDAUTOTING 22 SUKSPLGGERYEDITLAKSN 95 LAERGUFFVATENYTRESSON DP3A_M_115777 4 KIDILOUGUTATEDOCH 2 GENERYENGTATION TOWN 27 A KILLGENTRETNINTIK 112 7000LINULAKSN 24YFHHMENE / X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERYENGTATION TOWN 27 A KILLGENTRETNINTIK 112 7000LINULASSON 24YFHMMENE / X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERYENGTATION TOWN 27 A KILLGENTRETNINTIK 112 7000LINULASSON 24YFHMMENE / X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERYENGTATION TOWN 27 A KILLGENTRETNINTIK 112 7000LINULASSON 24YFHMMENE / X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERYENGTATION TOWN 27 A KILLGENTRETNINTIK 112 7000LINULASSON 24YFHMMENE / X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERAL X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERAL X800_H_115777 4 KIDILOUGUTATEDOCH 2 GENERAL X800_H_125151334 98 ROOTUNDANGOR 2 GENERAL X800_H_125151334 98 ROOTUNDANGOR 2 GENERAL X800_KC2465533 7600HWMENDOR 2 BENERL / X800_KC2465533 7600HWMENDOR 2 BENERL / X800_KC2465533 7600HWMENDOR 2 BENERL / X800_KC24645533 7600HWMENDOR 2 BENERL / X800_KC246454410 000KMENDOR 2 BENERL / X800_KC246454533 3 KONTHENESSON 24 KINESSON 24	DNAPOLIII_Tp	21	FVHLH	vhsnysi	LLDGA	- 4	RLVRTAKSL	GQEAL	ALT DHGNMF	30	DRSEHTIGRE	AYYHLIVLAKNE	96	LACREGIPLVVTNDVHYAEQ	EDSV	
DNAPOLITI_Mtu_1706493         10         FWHID_DTEXEMIDGA         4         FWLARDELGAP AVECTO DURY GSGSTENDATEMARANA         97         JGRALDIFFPLATE DEBYTATE DURY FUEXARE           DP3A_E118794         6         FWID_WS SUMMINGL         4         PWXKAAALCHPALATE TWIC         2         NVCCULGBELTPUT/LANN         96         LAERGLEVVATE DURFTERSTOP           DP3A_H_1169392         7         FUND_WS SUMMINGL         4         PWXKAAALCHPALATE TWIC         2         NVCCULGBELTUT/LANN         96         LAERGLEVVATE DWFLINESUMPT           DP3A_H_1169392         7         FUND_WS SUMMINGL         4         PWXKAAALCHPALATE TWYCUNG SUMMID DETNIC         2         NVCCULGBELTUT/LANN         95         LAERGLEVVATE DWFLINESUMPT           DP31_MD1_169392         2         FUND_WS SUMMID DETALDOULS         2         EUMINT WARDOWNELLOOD         10         AKALDERVENT         10         10001 FUND DETALDOULS         2         EUMINT WARDOWNELLOOD         10         AKALDERVENT         10	DNAPOLIII2_Mtu_2661649	46	YAELH	ABSAYSI	FLDGA	4	ELVEEAARL	GLCAL	ALT DH DGLY	26	TARTERPDPE	GPHLLVLARGP	98	LAPRECVGIVATTGAHFADE	SRGR	
DP3A_E_118794         OF UNIT_UNIT_UNIT_UNIT_UNIT_UNIT_UNIT_UNIT_	DNAPOLIII1_Mtu_1706493	10	FVHLH	NHTEYS	MLDGA	4	PMLAEVERL	G <mark>M</mark> PAV	GMT DH GNMF	42	KADDVSGSGS	SYTHLTMMAENA	97	IGRALNIPPLATNDCHYVTF	DAAH	
DP3A_E_118794       6       FUNERWIGHTMENTDE       4       PAXKEAAALGHPALATGPTINLC       22       NVQCULLADBLIGLUTULAANN       96       LAERGELFUVARTOWRIDSTEPP                 DP3A_H_1165392       7       FUNERWIGHTMENTDE       4       KAUNDAWQDQOVANCEGO-FUNE       22       KVXSCLADBYPDLILLANN       95       LAERGELFUVARTOWRINTUMPTOMENTEPPE                 DP31_MDG.1653978       321       EVERTAWERTMENTMENT       4       KAUNDAWQDQOVANCEGO-FUNE       13       IKELIGLSVTWERTWERTMENTMENTMENTMENTMENTMENT       /         VB00_EC.175777       4       EVERTAWERTMENTMENCE       2       ELMHEN MAQONAVALADOUTTA       78       AKALAGERVERSUFARETVE       90       MAKED ALGOUNDAWARE FUNE       /         SLORF1_S1_75777       4       EVERTAWERCEDOUTTA       2       ELMHEN MAQONAVALADOUTTA       78       AKALAGERVERSUFARETVE       90       MAKED ALGOUNDAWARE FUNE       /         SLORF1_S1_S10314       56       FUNERWINDEL       2       GLEGXIHQELOGLATION HAVANALAGENTREPE       /       MARAGUESTAWER       /       MARAGUESTAWER       /        SLORF1_S1_S10314       56       REAGOUNDAWINDUST       REAGOUNDAWERT       /       MARAGUESTAWERT       /       MARAGUESTAWERT       /       /       /       /       /       /       /       /       /	DP3A2_Bs_2293260	3	FVHLQ	VHSGYSI	LLNSA	4	ELVSEADRL	g <mark>y</mark> as <mark>l</mark>	ALT DDHVMY	22	SVFTDDSELE	RAYPLVLLAKSN	91	LSEETGIPVTATCDVHYIRK	EDKA	
DP32_Mp1_169392       7 FITERTEDEPERMEDET       4 PM/KAGANABEVAMEDEPINC       22 KVKSPLCEDSYPELTLAKSNN       95 LAERCELFLAKTOVMPLATEOPEN       1         DP32_Mp1_1673778       321 PVELVERVERMAPDOL       4 EXACFAKERDKYLINATOVAVEDO-PNLF       27 QOCHFELLKERTFINGTIKTIK       1       1       120 DLAERCELFLAKTOVANDE       1         VB00_H1_175777       4 EXACFAKERDKYLINATOVAVEDO-PNLF       81 AKALAGUSTENTATIVUT       100 MARCEPLOGEOXED DEPERCENTEL       2         SLORE_511513194       96 REGOLULADNU BOGSF       2 BAMERTAALDIGERMA DIGTER       4       100 LAEQUATIVENT       31 DLAERCEPLOSCOVED DEPERCENTEL       2         SLORE_511513194       96 REGOLULADNU BOGSF       2 BAMERTAALDIGERMA DIGTER       4 MEDSCULVASINGLARG       97 AVEAACULFSTOTAALOGUADO       31 DLAERCEPLOSCOVENTS       31 DLAERCEPLOSCOVE	DP3A_Ec_118794	6	FVHLR	VIIS D <mark>YS</mark> N	<b>HIDGL</b>	4	PLVKKAAAL	GMPAL	AITDFINLC	22	NVQCDLLGDE	RLT <mark>H</mark> LTVLAANN	96	LAEARGLPVVATNDVRFIDS	SDFD	
DP32_Mpn_2494194       2 FWRETNS YENTLNST       4 KXVDLAVQOQQVANCTATADOLPHIL       13 IKPLIELSUTVENUMELL       ( <t< td=""><td>DP3A_Hi_1169392</td><td>7</td><td>FIHLR</td><td>T<mark>h</mark>t D<mark>f S</mark>N</td><td>AIDSI</td><td>4</td><td>PLVKACAAN</td><td>ENVAN</td><td>GLTDFTNFC</td><td>22</td><td>KVKSPLCGDE</td><td>SYFDLTLLAKNN 1</td><td>95</td><td>LAERCDLPLVATNDVMFLNT</td><td>EDFE</td><td></td></t<>	DP3A_Hi_1169392	7	FIHLR	T <mark>h</mark> t D <mark>f S</mark> N	AIDSI	4	PLVKACAAN	ENVAN	GLTDFTNFC	22	KVKSPLCGDE	SYFDLTLLAKNN 1	95	LAERCDLPLVATNDVMFLNT	EDFE	
DP31_Mpn_1673778       321       FVEX.DEBRYKEMAPDCI 4       EXXD/FXKEMAPDCI 7       ALVIENUEWCKGENCOLUT 7000000000000000000000000000000000000	DP32_Mpn_2494194	2	FUNLH	TNSYYN	FLNST	4	KLVDLAVQD	QQVA <mark>V</mark>	CLTD-PNLF	13	IKPLIGLSVI	VRHYEQNVNLL			I	
YE00_HL_1175777 4 KXDUCQESYNG DEPFCCWIEL 2 ELWHRYAGGWURKUTATODUDTA 78 AKALADEVERTYARYAUVU 90 WAKEFDLQCSYNG DEPFCCWIEL 2 \$10564_Ssp_101243 53 FYNELLERCEDQCF 2 ALWHRYURKUTATODUDTA 71 ALALQGSAVERTYGFARTURU 90 LARCOGLWURAUFARYAGFA 39 LODLYNLFRIGODDUQCEYNG DEPFCCWIEL 2 \$10549_Ssp_101243 53 FYNELLERCEDQCF 2 GLEBORTHOELGCALTONICUA 72 ALHDAGGLWULAUFARYAGFA 39 LODLYNLFRIGODDUQCEYNG DENGFCGWIEL 2 \$107577 4 KXDULLSANGLOGSP 2 EMGRIAALGHEKAALTONICUA 72 ALHDAGGLWULAUFARYAGFA 39 LODLYNLFRIGODSUGAPYNL / YCDL_2_2495593 3 FYDDIALTSANGLOSSP 2 EMGRIAALGHEKAALTONICUA 72 ALHDAGGLWULAUFARYAGFA 69 RAVAAUUFSID DENGFCGWIEL 2 YCCL_2499170 10 EXTINCTING KANN 3 DIXLAOKKGKIKGING CTDG GPM 44 MFDSLDLILIGGNEEVYAFHD 69 AVRADCWAAUUFSID DENGFAFHGEF 1 DNAX_ML_2621626 335 RCDUMMSLFBIDIDS 2 EMERYSUGGYTATOMARYI 37 ALKHEDVUNSVISKLENDSA 71 LARDICVFSIDT DARAGGLDH 3 YCCL_82_177035 335 KCDUMMSLFBIDIDS 2 UMEYSUGGYTATOMARYI 37 ALKHEDVUNSVISKLENDSA 71 LARDICVFSIDT DARAGGLUH 1 yehC_82_177035 335 KCDUMMSLFBIDIDS 2 EMERYSUGGYTATOMARYI 37 ALKHEDVUNSVISKESMOPHE 61 KANUNGCKIAINE DARSKGLENI HIS9_SC_582256 1 MHSHJSBGOYGAHOT 4 SVDOAVANDAWITYDIGEVL 49 VLAEMDIVIASTESSNOPHE 62 KCMREGVYICITYDARSAGLENI HIS9_SC_582256 1 MHSHJSBGOYGAHOT 4 SVDOAVANDAWITYDIGEVL 49 VLAEMDIVIASTESSNOPHE 62 KCMREGVYICITYDARSAGULHM 1 HIS9_SC_3230508 2 FYSSISSOFCHAD 4 UVYOCYNLATIONACUTHALIPPI 74 NNDLLKFCVOSVHINNSIPID 124 VKRKGSRFULDDARVAQWOVC V HIS9_SC_23321 3 KDDIMMSHFBHFGADSES 3 KURISKIKKGESSTETHERD 75 0 DSAFPTEVINFENDELTUN 104 KNRCKANUNGSV 4C_2931Aae_2933501 43 SNAMINGYNYDON V 2000 VXINNHTYDIGEVT 75 0 DSAFPTEVINFERVENNED 104 KNRCKANUNGSV 4CJ23_AAe_2933501 43 SNAMINGYSSISS 2 UVVLVGSRGVUNGTATIONYDEP 55 FOFEVINNTLDIYUNGSFUVEPV 55 FOREWRINGENDED 104 KNRCKANUNGSV 4CSDSJ_AAE_2933500 43 SNAMINGYNYDYDYDYD 20 1 LAREDGVCFLUPPFDFIRK 20 YKKLEWFINNED 104 KNRCKANUNGSV 4CSDSJ_AAE_2933500 13 REDGNESSE 2 UVVLVGSRGVUNGTATODUVEP 55 FOFEVINNTLDIYUNGSFUVEPV 55 LAREFGEVINGNYN 4CSDSJ_AAE_2933500 13 KNDYNFFSHFKKSSSIS 2 UVVLVGSRGVUNGTATODUVEP 55 FOFEVINNTLDIYUNGSFUVEPV 55 LAREFGENCHINGSFUNG 4CSDSJ_AAE	DP31_Mpn_1673778	321	RVELV	F <b>HT</b> KMS/	AFDGI	4	EXAQFAKER	DWKTI	AVT DK DNIH	227	QQCHFELLKF	TFINTGIIYIK	102	TAQRLNKLVAVASDAYFIHP	WENE /	1
YE00_Ec_1175777       4       INDEGETTABDOCLT       2       ALMERYDERKOTATIONETTA       510 (FLACGAVERGERAPELVEC       90       LARCHINASCONDER/DECWIEL       2         S100549_Sep_1001243       51 FNNELINGESCONT       2       GLESCALATONICY       72       ALMERGERALGUERANCAL       39       LEDUNALFRACTON DESIGESINGUL         SLORF1_S1_1518394       96       RODOULUADAGOGGEP       2       BEMARTAALGHEVARTHOLEPRL       49       LEERLDVVVVSUSKLANDAR       69       RAVARACUFSINT       64         YCOX_Ec_2495593       3       PUDUDAGEVARTHAS       3       DVIAONKOKGEKLERATONISPRL       49       MLDRLDVVVVSUSKLANDAR       69       RAVARACUFSINT DARAFOQLEDK       3         YCOX_Ec_2495593       10       SKDDUMUSKINSDARAP       2       BEMARTAALGHEVACATONISPRL       44       MLDRLDVVVSUSKISKLSDAR       11       LARDICVVSLOG DARAFOLDELL       13         YCOX_Ec21703       10       SKDDUMUSKISHADDAR       2       MENETYSKUGVVLICUTOLEQGA       51       VLEQEPTVASKISKLSDAR       61       LANDICVVLICUTADAGARACLENU       14       LANDICVVLICUTADAGARACLENU	YE00_Hi_ 1175777	4	KYDLH(	CHSTASI	DGVLS	2	<b>ELVHRAYAQ</b>	G <mark>V</mark> NVL	ALC DHDTIA	78	AKALADGEVI	RAHYARYLVQI	90	WAKEFDLQGSVGSDFHFPCG	WIEL /	
sl10549_ssp_1001243 53 FYNFILERECGOOCMT 2 GILEONIHOECO ATTOENCYA 72 ALEBAGIVVLABCARVROPA 39 LGDLYBLEFETCOTOSBEESIMYL / SLORFI_SL1518394 98 RGOOLMAANDROGSP 2 EMGENAALGHEVAALGHEVAALGHEVAATTOIGPEN 44 MFDSLDLIIAGENEPVFAPHD 69 AVRAAGVESTUT ABAGGUDA \ YCDX_EC_2495593 3 FVDLMBETVASTHAYS 3 DYAAGKOKGIKLFAITOIGPEN 44 MFDSLDLIIAGENEPVFAPHD 69 AVRAAGVESTUT ABAGGUDA \ YCML_AL_2499170 10 EVENGGIT ISGAHAY 3 EWLEGKKGIKGETGT GPEN 44 ALKELPVINSLEDVCIPSOF 68 LCKKTVVINKOS ANKAAGUSTICKGC I DNAX_MLA_2621626 335 RGDLMMEIFFGIDS 2 QMEYASVLGKGITTOGOPEN 44 ALKELPVINSLDVCIPSOF 68 LCKKTVVINKOS ANKAAGUSTICKGC I DNAX_ALG_2983818 342 KGDLMMEIFFGIDS 2 QMEYASVLGKGITTOTOGNSQSA 51 VLEQEFVVASVDSRFPQDNF 62 KCMEAGVVIGUT DANSAGRALENI DNAX_ALG_2983818 342 KGDLMMEIFFGIDS 2 QMEYASVLGKGITTOTONSQVL 49 VLEDEFVVASVDSRFPQDNF 62 KCMEAGVVIGTADASKGGCCYC NAX_AGG_2983818 342 KGDLMMEIFFGIDS 2 QMEYASVLGKGITTOTONSQVL 49 VLEDEFVVASVDSRFPQDNF 62 KCMEAGVVIGTADASKGGCCYC NAX_AGG_2983818 342 KGDLMMEIFFGIDS 2 WAEVASVLGKGTGTTOTONSQVL 49 VLEDEFVVASVDSRFPQDNF 62 KCMEAGVVIGTADASKGGCCYC H199_SC_585256 1 MHSHSGGCOYCHAHO 4 SVVDOVNLNHTNOTOTOTOTOTOTOTOTOTOTOTOTOTOTOTOTOTOT	YE00_Ec_1175777	4	IYDLH:	SETTASI	DGCLT	2	ALVHRAVEM	RVGTL	AITOHOTTA	81	QRLAQGAV1	RGHFARFLVBC	90	LARQHHLWASQGSDFHQPCF	WIEL	2
SLORF1_S1_1518394       98       RODD1LIADWBOGSP       2       BWGRTAAALGHENAALTGISPEL       49       LERLEVVVVSVBSKLENAAR       69       RAVAACULFSIDTAABAPGLDWQ       \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\	sl10549_Ssp_1001243	53	FYNFH	LHTRCSI	DGQMT	2	GLIEQAIHQ	elqge	AIT DHHCVA	72	ALHDACGLVV	LAHPARYRQPA	39	LGDLYNLFRTCOTDSHGESI	MYRL /	
YCDX_BC_2495593       3       FUDIMETVASTHAYS       3       DYLQACKGKKIFATTOLSPEM       44       MFDSLDLITAGEBEVFAFHD       69       AVRDAGWALGEDSTATMOSPEM       3         YCM1_ALL_21781100       97       RGDLMEENNSUSASA       2       BENVLEACKKIFATTOLSPEM       44       MLDRLDIVVASUSKISHOST       1       LARDICVFSIDTOLARAPGEOPEL       3         YCM_ALL_2499170       10       EVERNORTIASCHAYN       3       ENVLEACKKSIKGICITOLGPEM       44       ALKREDVVIASUBDVCIPSGT       68       LCKKYGVYIANGSDAHACGEDSTAFTMOS         DNAX_AGE_2983818       342       KCDLMENTNSGOVYT       2       EMARCIKKOVYTATTOLSQUA       37       ALKNEDUVIASUBDVCIPSGT       61       MAVDMCCKIAINSDUSSARCINI         JPALA       ALKNEDUVIASUBTVSIDAFS       2       EMARCIKKOVYTATTOLGQUA       51       VLEOPEVVASUSSENDUVGIPI       124       VKHCGSRFVLEDDALSARCINI         JPALA       45       SVDQUVNLNHTYCLOVITICOLSQUA       50       DSAPPEVIASUSSENDUGIPI       124       VKHCGSRFVLEDDALGOVCLAV       HISS         JPALA       45       SVDQUVNLNHTYCLOVITICOLSQUA       40       VUTQEAUCHTONSQUALTINDUFUC       62       TKYNLDYFVCSVHHVNGIPID       124       VKHCGSRFVLEDDALGOVCLAVEGT       144       MANDACCKIAINSDLECONUT       144       MANDACKCKAINSDLECONUT       144       MANDACKCKAINSDLECONUT       144	SLORF1_S1_1518394	98	RGDCHI	LHADWSI	DGGSP	2	EMGRIMAAL	GHEWA	ALT DHSPRL	49	LLERLOVVVV	SVHSKLRMDAR	69	RAVAAGVLFSIDTDAHAPGQ	LDWQ \	
MTCY01A6.12_1781100       97 RCDEIESNNEDGSAP       2 EXAMTALALGAUXCATTORSPEL       44 MLDELDIVVASVDSKLSMDSA       71 LARDICCVPSIDTORALACOLPFL       3         VCR4_ALL_2499170       10 EXETLORIASGANYN 3 EMNEDGAUXCTTORSPEL       44 ALKLEDVISLEDVUTSC       68 LCKKCVVIANGSONPKKADIGCC         DNAX_Mta_2621626       335 RCDLIMISLFEDGIDS       2 UMBEYNSUGREYIATTORARYI       37 ALKNEDUVIASUBEVKORGUTER       61 MAVDMCCKIAINSDARSGALENI         DNAX_Mta_2621626       335 RCDLIMISLFEDGIDS       2 EMPETAYKLOQVIVICOUSSON       51 VLEQFDEVVASVDSRFPDDT       62 KCMERGVVIGUTORDASARCHLM         yshc_Bs_170035       335 KCDLIMISTYBERGEYSANGT       2 EMPETAYKLOQVIVICOUSSON       51 VLEQFDEVVASVDSRFPDDT       62 KCMERGVVIGUTORDASARCHLM         yshc_Bs_12370508       2 FISSESGCPCLIAQ       4 DVJOEATOGCQCGPGEFEREIPER       62 TKYNLDYPCSVGHVNSIPID       104 LMVECGVUCUTORDAGAQQCVC         YHIC_LL_0401577       3 KDVEHSHERMEDSE 3 KHVEATAYGLEZITETEHAPL       61 YGPYLDDSILSVELGDEVINTYDEVVCSV       97 FKRGKFTLTDESHLEPPKOLS         qc_933Aac_2983671       5 LCDFTIFFEREDSE       2 VVDDVVSSKGPDVTATTORVFD       50 SAFPDEVINSVELGDEVINTE       10 EAROKKIPLVFGSDAQAGEVGA         Acf2503L4       3 RAEVEVSSFBORDE       2 VVDLVSSFBORDEVIATTORVFD       50 TSAFPLEVISSFFEDGEN       10 EAROKKIPLVFGSDAQAGEVGA         Ac_2983671       5 LCDFTIFFEREDES       2 UVDLVSSFBORDEVIATTORVFD       50 TSAFPLEVISSFFEDGENLS       2 VVDLVGSFAUGAGEVYL <td>YCDX_Ec_2495593</td> <td>3</td> <td>PVDLH</td> <td>METVAS</td> <td>THAYS</td> <td>3</td> <td>DYIAQAKQK</td> <td>GIKLF</td> <td>ATTOHGPDM</td> <td>44</td> <td>MFDSLDLII</td> <td>GFHEPVFAPHD</td> <td>69</td> <td>AVRDAGGWVALGS DSHTAFT</td> <td>MGEF</td> <td></td>	YCDX_Ec_2495593	3	PVDLH	METVAS	THAYS	3	DYIAQAKQK	GIKLF	ATTOHGPDM	44	MFDSLDLII	GFHEPVFAPHD	69	AVRDAGGWVALGS DSHTAFT	MGEF	
YORA_Ath_2499170       10 EVETBETIASCHAYN       3 EWULSQKKSLKGLCITDIGPEN       44 ALKKLDYVIALEDVCIPSGT       68 LCKKYGVIAKGSDAENKALGKSLKGLCITDIGPEN         DNAX_Mae_2621626       335 RCDLENSILS 2 QMAEYASVUGEVIAIDORSUS       51 VLEQFDFVVASVESRFEQDNT       62 KCMEAGVYIGIVT DAKSARQLENI         DNAX_Aae_2983818       342 KCDLENSTKEDGAFS       2 EMVETOYKLGYQYIVIODISQSA       51 VLEQFDFVVASVESRFEQDNT       62 KCMEAGVYIGIVT DAKSARQLENI         JNAX_Aae_2983818       342 KCDLENSTKEDGAFS       2 EMVETOYKLGYQYIVIODISQSA       51 VLEQFDFVVASVESRFEQDNT       62 KCMEAGVYIGIVT DAKSARQLENI         JNAX_Aae_2983501       315 KCDLENSTKEDGAFS       2 EMVETOYKLGYQYIVIODISQSA       51 VLEQFDFVVASVESRFEQDNT       62 KCMEAGVYIGIVT DAKSARQLENI         JYBC_SS_2370508       2 FISSUESGOFCLHAQ       4 DVIQCAIQOGOGSGFTEHERPRD       62 TKYNLDYFVGSVEHVNSIPID       104 LMVECGAQLTLEDDSHGGPRQVGLN         YHYE_SS_223321       3 KNDHENSTHEGASSE       3 KINTISIIANGUEL       60 SAPPDFVIASVEHVNSIPID       104 LMVECGAQLTLEDDSHGPRQVGLN         aq_1230_Aae_2983501       43 SXARHUHTOFSYNSLG       2 DVVDLYGSHGPDVTAITDHVPD       55 FGFEVINNTDLMEKKYRM       61 LSEGWENNVUGGIVALARDANGAAGVGLA       4         AF1233_Af_2649349       1 MIDUHESNYSDGQGS       2 ELRERAKERGLKATATUONENUT       26 DLIEVODIKVISHFFKKRZ       63 LCRNKLAYSIGSAHASKARAVGLA       4         MJ1295_M_12128MEVYGQA       3 RXELMESSKSERGALENI       2 ELRERAKERGLKATATUONENUT	MTCY01A6.12_1781100	97	RGDLH	UHS NWSI	CGSAP	2	EMMATAAAL	GHQYC	ALT DHS PRL	49	MLDRLDIVV	SVHSKLSMDSA	71	LARDICCVFSIDTDAHAPGQ	LDFL	3
DNAX_Mta_2621626 335 RCDLPMESILFEDGIDE 2 QMEETRSVEGREYIATDGARYI 37 ALKHEDLVIASIEDPCNLTER 61 MAVDMCCKIAINS7ABRGALENI DNAX_Aae_2983818 342 KCDLPMETNKEDGAFS 2 EMPERAVKLGVQYTVIDGESQX 51 VLCQFDFVVASUSRFEDGINF 62 KCMERGVICKIVT 7ANSARQLHLM yshC_Be_1770035 335 KCDLPMESTMEDGAFS 2 EMPERAVKLGVQYTVIDGESQX 49 VLAEMDIVIASIESFMQPEH 67 KANEQGVTLVINT 7ANSARQLHLM His9_SC_585256 1 MHSHESGEGVSAHGT 4 SVVDQVVNLMEHTYGTGEFFEIFERI 74 NNDLKFCVGSVEHVNSIPID 124 VKKKGSRFVLSD 7ANSARQLHLM HIS9_SD_2370508 2 PISSUSSGCGCLHAQ 4 DVICEAIQGEQGSGFGFFEHFERI 74 NNDLKFCVGSVEHVNSIPID 104 LMVECGAQLTED 7ANSARQHQQQVCVC HIC1_L401577 3 KLDYHESIHEADSEB 3 KHVIERIAYGLEELGTERHERFI 74 NNDLKFCVGSVEHVNSIPID 104 LMVECGAQLTED 7ANSARQHQQQVCVC 4q_1230_Aae_2983671 5 LCDFHIETERSDGSLS 2 DVVLYGSHGFDVIAITONVFDP 55 FGFEVINNTLAVHEIGDIEVY 97 FKKLGKFFITTGTSHAKRMLS 4q_931_Aae_29836500 43 SVAFHETCFYDENSDS 4 QVXEAREIGCDDVIVTOREVD 26 DLEVVDIVTUSHEKKNYRM 61 LSEDWENKVCGTDENHKKYYRM AF0505_Af_2650121 3 RAELMESSEDGRGG 2 KLEARAVERKLEVIAITONOTVQ 51 LARELGQCFLAEPFDFIKG 28 YAKKLSKPEIASGANSARAVGLA 4 AF123_Af_2649349 1 MEDLTESNYSDGQGS 2 ELARRAVERKELGVIATTONOTVQ 51 LARELGQCFLAEPFDFIKG 28 YAKKLSKPEIASGANSARAVGLA 4 AF123_Af_2649349 1 MEDLTESNYSDGQGS 2 ELARRAVERKELGVIATTONOTVQ 51 LARELGQCFLAEPFDFIKG 28 YAKKLSKPEIASGANSARAVGLA 4 AF123_Af_2649349 1 MEDLTESNYSDGQGS 2 ELARRAVERKELGVIATTONOTVQ 51 LARELGQCFLAEPFDFIKG 28 YAKKLSKPEIASGANSARAVGLA 4 AF123_Af_2649349 1 MEDLTESNYSDGQGS 2 ELARRAVERKELGVIATTONOTVQ 51 LAREKGREJVINEY 55 IAREFGLKTLINTONSPESDIND MJ1295_Mj_2128772 2 REDFERETVESDGELI 2 ELARRAVERGLKATATIONOTVQ 51 VLVVVSGETVAEPV 55 IAREFGLKTLINTONSPESDIND MJ1295_Mj_2128772 2 REDFERETVESDGELI 2 ELARRAVERGIATIONADFT 42 LAWKKGAEIVVVKGETVAEPV 55 IAREFGLKTLINTONSPESDAN MJ10417_Mj_2495958 2 KVDLHMESIVEKCSLN 2 GLEEKCIKNIVOFSERKKG 77 NVLARGVTLAEFF 55 IAREFGLKTLINTONSPESDIND MJ295_Mj_2495759 4 MVDLHGSTYSGCAGS 7 SILESYRQQVRIIVOTSSERKG 77 NVLARGVTLAEFFENERRER 36 RLVVASOMENUCVAEPPF / VWHEBSISSF55 1 MUDVHSIVEDDGF 7 SILESKYRQQVRIIVOTSSERKG 77 NVLARGVTPLAEFENRER 36 RLVARANFGAEAVVKRNFH 55 CPARAVERAFFYRFFENERFENERFENE	YOR4_Ath_2499170	10	EVETH	OHTIAS	SHAYN	3	EMVLEAQKK	GLKGI	CITCHGPEM	44	ALKRLDFVIA	SLHDVCIPSGT	68	LCKKYGVYIANGS DAHYKAD	IGRC	
DNAX_Ade_2983818       342 KCDLENETWORDSYNT       2 BAVETRYKLGYCYTUCDUSCSA 51 VLECEPTVASVBEREPEDINT       62 KCMEARWYICIVTICTUCTUCTUSSARCLHM         His9_SC_585256       1 MHSHBESGDYSAHGT       2 SVUDONVNINEHTYCLTEHIPRI       74 NNDLLKFCVGSVBHVNSIPID       124 VKKHCGSRFVLSD DARGVAQVGVC         His9_SC_585256       1 MHSHBESGDYSAHGT       4 SVUDONVNINEHTYCLTEHIPRI       74 NNDLKFCVGSVBHVNSIPID       124 VKKHCGSRFVLSD DARGVAQVGVC         HIS9_SD_2370508       2 PISSBESGOFCLHAQ       4 DVICELICQCFCSFBTEHIPRI       62 TKYNLDYPVGSVBHVNSIPID       104 LWPCGAQLTLEDDBHCGHQVGLN         YHTC_LL401577       3 KDDHBESHFBOSBS       8 KHVISTIAVGLEBCTFBERPFF       50 DSAPFDFVILASVBEIGDEVY       97 FKRLGGKFTLGTDBHKRKPLAS         YUTP_BS_223321       3 KRDGHBEFFCHRGSN       4 QXAESALKKGFESITETEHAPLP       61 YGPVLDSILSVBFLKTVSVF       55 TVGLKGCRFVAND OPHIKARDMAR         ac_1230_Aae_2983671       5 LGDFHBETESDSLS       2 DVVDLVGSGGEDVIATDDHVP       55 FGFEVINNTDLMEIVVLDVYE       65 TVGLKGCRFVAND OPHIKKNYIREV         AF1505_Af_2650121       3 RAELPUSSFBORDG       2 KILEAAVEKKLEVIATDDHVQ       51 IARELGGVCFLAMPFDFIKS       28 YAKLEKPELAGDAHIKKPYTREV         AF1505_Af_2649349       1 MTDLHESNSDQQGS       2 BIARRAKREGLKAIATVDRSTE       41 PFDPENIASVEFEVYQQAY       63 LCRDRKIAYSIGGAHSARAVGLA       4         MJ1295_Mj_2128772       2 REDFREGLKSLAIATVDKSKGATTDDAFF       42 LWKKGAEIVVVVGGETVAEVEV       55 IAREFGLKTLINFOTBASEDIAU <td>DNAX_Mta_2621626</td> <td>335</td> <td>RGDLH</td> <td>MESLESI</td> <td>GIDS</td> <td>2</td> <td>QMAEYASVL</td> <td>GREY<mark>I</mark></td> <td>AITDHARYI</td> <td>37</td> <td>ALKNFOLVIA</td> <td>SINDPGNLTER</td> <td>61</td> <td>MAVDMGCKIAINSDAHSRGA</td> <td>LENI</td> <td></td>	DNAX_Mta_2621626	335	RGDLH	MESLESI	GIDS	2	QMAEYASVL	GREY <mark>I</mark>	AITDHARYI	37	ALKNFOLVIA	SINDPGNLTER	61	MAVDMGCKIAINSDAHSRGA	LENI	
yshC_BS_1770035       335       KCDLHENSTYNDKAFS       2 BENERALIKKGXYMALTAIDISON1       49       VLAENDIVIASHESSMOPEN       67       KANEQUVILINE DAUVILSID ANGVAQVGVC         HIS9_SC_S52256       1 MHSHBSGCYSANGT       4 SVUDOVNINNEHTYGTJERFTPINS       71       NADLIKFCVSSVBHVNSIPID       104       LMVKKICSSRPTUED ANGVAQVGVC         HIS9_SC_S52256       1 MHSHBSGCYSANGT       4 SVUDOVNINNEHTYGTJERFTPINS       62       TKYNLDYFVGSVBHVNSIPID       104       LMVKICSSRPTUED ANGVAQVGVC         YHYC_BS_223321       3 KIDYBHBSGCYSANGT       4 SVUDOVNINNEHTYGTJERFTPINS       60       DSAPPDFVIASVBEIGDIEVY       97       FKRLGGRFTTDCRSANGAQVGVC         aq_1230_Aae_2983501       5 LCDFHBTETEMBASSE       2 DVVDLVSGGLGCDYVINDEVT       50       DSAPPDFVIASVBEIGDIEVY       97       FKRLGGRFTTDCRSANGAQVGVC         Ag_1230_Aae_2983501       43       SXAPHVHTOFSYNSLG       3 UVVLAVGHVEND       55       FGFEVTNINTDLMEVTVLDVYE       65       TVGLKGCRFTANG ONHILKHEFAM         Ag_1230_Aae_2983501       43       SXAPHVHTOFSYNSLG       3 UVVLAVGHVEND       51       LGPHILTEREDCHANAGAGUAVA       4         AF1233_Af_2649349       1       MTDHHBSNSDQCGS       2 ELARRAYRAGNKGAITOHADFY       42       LAWKEGRETVAEVES       54       LAENYGPELVIND ONAGAGUAVA       4         MJ1295_M_12128772       2 REDFHENTSYSDQCGS       2 ELARR	DNAX_Aae_2983818	342	KGDLH	MHTNWSI	DGVNT	2	EMVETAYKL	GYQYI	TGDHSQSA	51	VLEQFOFVV/	SVHSRFEQDNT	62	KCMEAGVYIGIVT DAHSARO	CHLM	
H199_SC_S45256       1 MHSHBBSGOYSAHGT 4 SWDLWINNEHTWICTGHTPHI 74 NNDLLKPCVGSWBHWNSTPID 124 VKRKGSKFLBDQAUGVGVC (         H199_SC_2370508       2 PIS9BSGOFCHLAQ 4 DUTORIDCOFCSFTETEITED 2 TKURLDYVGSWBHWNSTPID 124 VKRKGSKFLBDQEVGQUN         YHUC_L1_401577       3 KLDYHPESHFGADSEB 3 KHVIEALAYGLESIGTEBRAPP 6 2 TKURLDYVEGSWBHWNSTPID 104 LWVPCAGLTEBDSHQVGVGVA         aq_1230_Aae_2983671       5 LOFFIETEFRENDSS 4 QXAESGLEKGEESITETEBRAPP 61 YGYTLDDSLSVBELGDIEVY 97 FKRLGEKFILEDOPHIKKYLEFFAM         aq_93_Aae_2983500       43 SXAENUHTORSYDSLG 3 DVKKAELGEDVYTVTDEVDT 26 DLIEVODIKVISBHFKKYRW       61 LSEGWERNVLGIDEBHKKYTREV         AF2639_Aae_2983671       5 LOFFIETEFRENDGSLS 2 DVVDLYGSHGEDVTATDBVPD 55 FGFEVNINTDLWIVUDVYE 65 TVGLKGCRFVABSOHNKVTREV         AF2649349       1 MIDLHESNYSDGQGS 2 KIERAYKERGLKAIATVOHSIEL 43 PDFDFDFILASVBEFVYQQAY 63 LCRDKLAYSTGSDAHSLSGYGEV         AF1550_Af_2649349       1 MIDLHESNYSDGQGS 2 ELARRWXERGLKAIATVOHSIEL 43 PDFDFDFILASVBEFVYQQAY 63 LCRDKLAYSTGSDAHSLSGYGEV         MJ2295_Mj_2128772       2 REDEFRETVS SDGELI 2 ELARRWZAGNKERATORADFT 42 LAWKEGAEIVVVEGETVAEPV 54 LAEXYGEPLINTD 0HSPSDIIND         MJ0417_Mj_2495968       2 KVDLHVESIVEKCSLN 2 GLDEKFCIKNIVATISTSRKK 77 NVLRACVTPIVHEFFUNER 38 YAEXYDPAMAFGBANHILGPRPFF (         YWRE_BS_199755       1 MIDLHOSTVESCEL 2 ELARRWZAGNKERATORADFT 42 LAWKEGAEIVVVEGETVAEPV 55 IAREFGLKTLINTDFBAPEDLIDD         MJ0417_Mj_2495968       2 KVDLHVEDVDGP 7 DLIGEFYAQOVRITYSTSKK 77 NVLRACHTVFWEGTVAEPV 55 IAREFGLKTLINTDFBAPEDVENA /             YWRE_BS_199743       1 MIDLHOSTVESCELN 2 GLEKFFQLKKILVINTSTSRKKG 77 N	yshC_Bs_1770035	335	KGDLH	MESTWEI	GAFS	2	EMAEACIKK	G <b>Y</b> QYM	TTDRSQYL	49	VLAENDIVIA	SINSENQPER	67	KANECOVILVINICANNIEM	LDDM /	
H159_5D_2370508       2 PHSDB0500FCLHAQ       4 DICCEDUDGECSENTIAITENTPHONED       2 TKYNLDYFVCSWHAWNSEPID       104 LAWDCLAQDITLENDSHAKRENLS         YHTC_L1_401577       3 KDCHHESTERDSEB       3 KLDYHESHHESDESE3       4 QYAEESALKKGFESITTEHAPLP       61 YGFYLDSILSVEFLETDSSY       109 EAKOKKIPLVTGSDAHQAGDVGYA         ac_1230_Aae_2983671       5 LCDFHEBTERDSDSLS       2 DVDLYGSGEFUTTEHAPLP       61 YGFYLDSILSVEFLETDSY       109 EAKOKKIPLVTGSDAHQAGDVGYA         ac_q_931_Aae_2983671       5 LCDFHEBTERDSDSLS       2 DVDLYGSGEFUTTEHAPLP       61 YGFYLDSILSVEFLETDSY       65 TVGLKGCRFYANS DHEHLKREFAW         ac_q_931_Aae_2983670       43 SXAFHUTCYSUSLG       2 DVVLNGSGEFUTTEHAPLP       61 YGFYLDSILSVEFLETDSF       55 FGFEVTNNTDLMEVVLDVYE       65 TVGLKGCRFYANS DHEHLKREFAW         AF10505_Af_2650121       3 RAELEVESSFSDGRDG       2 KILEAAVEKKLEVIAITDHDTVQ       51 IARELGVCFLANFPOPTIASVEFEVYQQAY       63 LCRDRKIAYSICGBAHLSAVGEA       4         AF1550_Af_264910       1 MTDHHESVNSDQGS       2 EXRRKKREGIKAIAITDHAPF       4 KAKDLGAEIVVVEGETVAEPV       55 IAREFGLKTLINTDHESPEDIIND       MJ0417_Mj_2495968       2 KNDLHWESVFBCGLL       2 ELEVRRAVLKHRAIAITDHAPS       4 LAEXYGFELVINTD VESPSDIIND       MJ0417_Mj_2495968       2 KNDLHWESVFBCGAGIKXIVJSSERAKG       7 NVLARAVTPHINTEN ENTPEDIARRER       38 YAEKYDFAMAFGSDAHLSAPKGEV       ////////////////////////////////////	H189_SC_585256	1	MHSH	SISCOYS	SAHGT	4	SVVDQVVNL	NFHTY	LTHUPRI	74	NNDILKFCVG	SVEHVNGIPID	124	VKKHCGSRFVLSDOALGVAQ	VGVC \	1
YH1C_LL_0117/       3 KEDYHENSIHBADSEE       3 KEDYHENSIHADSEE       3 KEDYHENSIHADSEE       3 KEDYHENSIHADSEE       10 EAKOKKIPLUT       65 TVGLMCCRFVANSOHHENSIHADSEE       42 DYFLDDSLINHELKINHELE       61 LSEGWENNVLGELDENVENTED         ac_1230_Aae_2983500       43 SXARHUNTOPSYDSLG       3 DVKKAREIGGDYYTODEVD7       26 DLIEVODIKVISHFKKKYRM       61 LSEGWENNVLGELDENVKVIREV         AF1050_Af_2650121       3 RAKLEYSTEDGELS       2 DVKKLARESCANATATOHENUT       26 DLIEVODIKVISHFKKKYRM       61 LSEGWENNVLGELDENVKUGELDENVKUGEL       4         AF1233_Af_2649349       1 MIDUHUSINSDOGGS       2 ELARRAYAGNKGATOHADF       42 LAWKEGAEIVVUKGETVEPV       54 LAEKYGPELVINDOHSPEDIEND       4         MJ1295_M_2128772       2 REDURINTY EDGELI       2 ELARRAYAGNKGATOHADF       44 KANDLAREIVVUKGETVEPV       54 LAEKYGPELVINDOHSPEDIEND       4         MJ1295_M_212868       2 KVDIHUSSITVEDGELI       2 ELARRAYAGNKGATOHADF       44 KANDLAREIVVUKGETVEPV       54 LAEKYGPELVINDOHASPEDLIDD       44 KANDLAREIVVUKGETVEPV       54 LAEKYGPELVINDOHASPEDLIDD       44 KANDLAREIVVUKGETVEPV       54 LAEKYGPELVINDOHASPEDLIDD       44 KANDLAR	HIS9_Sp_2370508	2	PISSE	SISCOFO	CHAQ.	4	DVIQEAIQQ	GFQSF	FTEHTPRD	62	TRYNLDYFVC	SVEHVNSIPID	104	LMVECGAQLITLEDSHGPHQ	VGLN	
ICVP_DS_229321       3 KROWHEETFECHNSK       QARESDLYARDEDUTATDEVPD       5 ICOFHLETEENDAGSDCIA         aq_123_Aae_2983671       5 LCOFHLETEENDSLS       2 UVDLYGSHGEDUTATDEVPD       55 FOREVINITULINEIVVLDEVVE       61 LSEGWENKVLGGLDEHVKVYIREV         aq_993_Aae_2983671       3 SXAPHVETCPSYDSLG       3 DVKKAREICGLDYVITUTEVEVPT       26 DLIEVCDIKVISHFKKKYRW       61 LSEGWENKVLGGLDEHVKVYIREV         AF0505_Af_2650121       3 RADUHUSSVEDGELI       2 KIEPAVEKLEVIATDDITVQ       51 IARELGQCFLAHFFDFIRG       28 YAKKLEVETAGGAAHSARAVGLA       4         AF1233_Af_2649349       1 MIDLHUSSVEDGELI       2 EIARRAYAGSKGRATDIADFT       42 LAWKEGAEIVVVEGETVAEPV       54 LAEKYGPELVINTOHSSPSDIND         MJ2455_Af_2649349       1 MIDLHUSSVEDGELI       2 EIARRAYAGSKGRATDIADFT       42 LAWKEGAEIVVVEGETVAEPV       54 LAEKYGPELVINTOHSSPSDIND         MJ2417_Mj_2495968       2 KVDLHVESIVENCELN       2 GLEKFCIKKNIVFAICDHNKI       38 RVEQGALIYLPEPPLARR       38 YAEKYDFAMEGGAAHFIWEVGGAA       ////////////////////////////////////	YHIC_L1_401577	3	KLDYH	PESHES	ADSEB	3	KHVIBAIAY	GLEEI	FTERRDFY	50	DSAPPDFVIA	SWIEIGDIEVY	97	FKRLGGKFITLGTOSHIAKE	DWLS	
ac_1230_Aac_2935671       5 ECEMENDICERSUSSES 2 DWDLIGSECULATIONAPDF 35 FOREWARDULARIUMVEDVIE       65 TVGLAGE_KYARDUPHENTARIUMTORSUSCIES         ac_933_Aac_2935500       43 SXAPHUNTORSUSCIG 3 DVKKAREIGGLDVVITUDEVUT 26 DLIEVODKVVLDVIE       61 ECEMENTARIEGADSANAGLA       4         AF1233_AC_2649349       1 MTDLHESNISDGGS 2 ETARRAKERGIKALATIDURDTVQ 51 IARELGGVCFLANPFDFIRKG 28 YAKALEKPETAGSANISARAVGLA       4         AF1233_AC_2649349       1 MTDLHESNISDGGS 2 ETARRAKERGIKALATIDURDTVQ 51 IARELGGVCFLANPFDFIRKG 26 YAKALEKPETAGSANISARAVGLA       4         AF1233_AC_2649349       1 MTDLHESNISDGGS 2 ETARRAKERGIKALATIDURDTVQ 51 IARELGGVCFLANPFDFIRKG 26 YAKALEKPETAGSANISARAVGLA       4         AF1250_AC_2649010       1 MTDLHESVISDGGEL1 2 ELARRAKERGIKALATATUDHENFT 22 LAWKERGAETVVLGENTPHENFT 25 LAKKERGAENVINTO MERGATUPAAGNERGATUDAADFS 44 KAKDLGAEIVVVLGETVAEPV 55 IAREFGLKTLINTO MERSPEDIND       MJ0417_M_2495968       2 KVDLHMESTVERCELL 2 ELARRAVIKHRATATUDHADFS 44 KAKDLGAEIVVVLGETVEPV 55 IAREFGLKTLINTO MERSPEDIND       MJ0417_M_2495968       2 KVDLHMESTVERCELL 2 ELARRAVIKHRATATUDHADFS 44 KAKDLGAEIVVVLGETVEPV 55 IAREFGLKTLINTO MERSPEDIND       MJ0417_M_2495968       2 KVDLHMESTVERCELL 2 ELARRAVIKATIVSTEGRAKG 77 NVLARAVTPHINTENTER       38 YAEKYDPAMAFGBOAMHUWECKNA /         MJ0417_M_2495968       2 KVDLHMESTVERCELL 2 ELARRAVIKOGIRTITATENHING 77 DLQLKGYIPVLAHERTVAH E KVDRAVIKASOMHUWECKNA /       5       FOREFORMER 36 RUKKAKENA /       5       FOREFORMER 36 RUKKAKENA /       6       FUEAGUATINTO MERSPEGASANA /       5       FOREFORMER 37       6       FUEAGUATINTO MERSPEGAANANAKEN	YtvP_Bs_2293321	3	KRDGH.	INTPECI	PHGSN	4	QYAEBALKK	GFESI	TFTDHAPLP	61	YGPYLDDSII	SVHFLRTDSSY	109	EARORRIPLVFUERADQAGD	VGYA	
ad_99_Add_954       astranumicus inste       backetsus       backetsus       backetsus       astranumicus inste       backetsus         AF0505_Adf_2650121       3 RARLESSUS       astranumicus inste       3 Backetsus       backetsus       astranumicus inste       astranumicus inste </td <td>ag_1230_Aae_2983671</td> <td>43</td> <td>LCDPH.</td> <td>THEFEMOL</td> <td>JUSLS (</td> <td>4</td> <td>DVVDLIGSH</td> <td>GEDVI</td> <td>TUTDAVEDP</td> <td>22</td> <td>PGPEVENNTL</td> <td>LIFIVVLDVIE</td> <td>65</td> <td>TVGLKOCKFVANS FHHIKH</td> <td>TDEN</td> <td></td>	ag_1230_Aae_2983671	43	LCDPH.	THEFEMOL	JUSLS (	4	DVVDLIGSH	GEDVI	TUTDAVEDP	22	PGPEVENNTL	LIFIVVLDVIE	65	TVGLKOCKFVANS FHHIKH	TDEN	
AF0503_AF_250161       3 RABLAGESSEDERGS 2 RELEASEDERGE 1 A PRODUCTION 100 51 FARLESCENAMEPEDERGE 20 FARLESCENAMEPEDERGE 100 FOR 100 F	ag_993_Aae_2983500	43	STAPH	VET CES	TUSLG	3	DVKKAREIC	GLDYV.	TTDAEVDT	20	DELEVODIKY	ISPHEKKKINW	20	LSEGWENKVLGGLDHHVKVI	IREV L	
AF1253_AL_200310       1 MEDINESSIGEQUES       2 BARKMARSDERGENERGENERGENERGENERGENERGENERGENERGE	AFUDUD_AE_2600121	3	MTDT	THENDER	ARIA	2	PERPER	OT VAT	TUDECTPL	13	DEPENDENT	CUMPPUPIKKG	63	LODDRYTAYCTORDAUCI CO	VGLA	G.
AF135_AL_208010       1 HEDRINGS VEDGEL 2       2 BARKHINGSKOLDDEL 7       2 BAR	AF1233_AE_2049349	1	MIDLA	THE NEED	AGG S	2	ETARNAR SN	CNIVOR	TOPHADET	40	PDPDPDPTTP	SVALEVIGUAI	6.0	LCRDRATAI SICALAHSESG	TIMD	
MJ2417_M2495968       2 KULENETITESOBEL 2 ELLENETIKANUPATODALIS 4 KARADANUVENDALIS 3 INTERESOLUTION SI INTERESOLUT	MT1205 M4 2128222	2	PEDE		VIELT	2	ETARCHIAN ETARDDEDUT	CHDAT	TOPRADES	44	KAKDLGAETU	UUCETUVEDU	54	LARPEAL KTL TNTDTHA DED	LIND	
Discreting	MT0/17 M4 2/05060	2	VUDT	UPC THOS	AGED I	2	CLIEVECTY	INTUD	TODENT	30	DUDEOGALTY	T.D. DEDI NDDD	38	VARYYORAMAROCOMPTWE	ucua /	,
MIDD/GLID/MUDGA       1       MIDD/GLID/MUDGA       1       EMARAAVRGGRTITATPHINNG       77       DLOLGYITYHAHERNINEIR       36       RLUEANLIHVAJGARNVKTINFH       5         CPSA_Sag_S85000       1       MIDD/GLID/MUDGA       7       EMARAAVRGGRTITATPHINNG       77       DLOLGYITYHAHERNINEIR       36       RLUEANLIHVAJGARNVKTINFH       5         CPSA_Sag_S85000       1       MIDD/BEIVENVDDGP       7       SLEESYRQOVRIUSTSPRAKG       77       NULMIGTFVVAHERNNEIR       36       RLUEANLIHVAJGARNVKTINFH       5         MJ0043_Mj_2495759       4       NVDLHJESRIGGRSK       5       NILKYGKLKGLNI GTGDC CHPD       68       IVRD/VGLIGPAHCVPDTLL       283       IPELRD/PFS/SDAMSYHPHRLG       5         MTH1911_Mta_26223045       15       LVD/HUTAPDVKERI       4       ELANALDEGMERAVVIKS/TEEPT       55       GDLDAVLSAVAKHEMVLJIGH       6         MTH1941_Mta_2622593       3       RIDPHHESVISCDARG       3       ELERASVUCDAVAVADNITMIK       49       EHDOGVALIPHEPRVKRQG       4         M1587       1       ALUEANLANDERNIK       1       RUEANLANDERNIK       49       EHDOGVALIPHEPRVKRQG       6         MTH191       4       2       SALUEANLANDERNIK       49       EHDOGVALIPHEPRVKRQG       6         MTH191       2       XLUEA	EDGB St 1276875	1	MTDUE	CHILD I VIEN	TDDCP	7	DLTGEGVAO	URKT	CONSTRACT	77	NULPAGUTPI	VALTERVDALE	41	FELEXMLAHAVASIMUMLOP	REFE 1	
Julies	1007E Be 1894743	1	MTDT	TLD	(DDGA	-	EMARAGENTRO	2TDTT	TATENNA	22	DLOLXOVIDI	TANDERNDETR	36	PLUEAMLTHEVA SDAHNNKT	DALER	5
MJ0043_MJ_2495759 4 NVDLTHESRFSGSTSK 5 NITKYEKLIGLNIIGTGDCTHPD 68 IVRDVGGLIGFAECVPPDTLL 283 IPELRDLPFLSHSDAHSYHPHRLG MTH1911_Mta_2623045 15 LVDTHVHTAPDVKERI 4 ELAHAALDEGHEAVVIKSKTEPT 55 GDLDAVLSAVAENEMVLGTGH 6 MTH1478_Mta_2622593 3 RIDPHHESVEGCARG 3 EIIRRHSSNVLDAVAVADNTMK 49 EIHDDGVAIIPHPFVRVRQG MI587_Mt_2496191 2 KNJUTERVEGCARG 3 EIIRRHSSNVLDAVAVADNTMK 49 EIHDDGVAIIPHPFVRVRQG	CPSA Sag 585000	1	MIDI	SHIVEN	TDDGP	7	SLIERSYRO	VRIT	STSTRENG	77	NVLMLGITPU	VALLERYNALE	41	YFLEENLVHFVASDMINLDU	RPPF /	, °
MTH1911_Mta_2622045 15 LVDTRUETAFDVKERI 4 ELAHAALDBCHEAVVIKSTEPT 55 GDLDAVLSAVABHENVLSTGH 6 MTH1478_Mta_2622593 3 RIDPHHESVYSGDARG 3 ELIRASAVCLDAVAVADBNTMK 49 EIHDDIGVALIPHPFVRVRQG 4 MI587_Mt_2406101 2 KDUHTERVYGCDARG 3 ELIRASAVCLDAVAVADBNTMK 49 EIHDDIGVALIPHPFVRVRQG	MT0043 MT 2495759	Â	NVDLH	THERE	CALSK	5	NILKYGELK	GLNIT	TGDCTHPD	68	IVEDVOGLTO	PARCUPPOTLL	283	IPELEDLPFLSNSDAUSYHP	HRLG \	
MTH1478_Mta_2622593 3 RIDPHTESVYSGDARG 3 EILRRASAVGLDAVAVADLWTMK 49 EIHDOLGVAIIPHFFVRYRQG MT1587 Mi 2406191 2 KDDILTRYKGGCKE 13 MILKUNKKKGTEUNATTAUWTME 49 KIKEGGCLADADDYSDICKA	MTH1911 Mta 2623045	15	LADT	UNTAPAT	KERT	4	FLAHAALDR	GMEAN	TKSHTEPT	55	GDLDAVLSAU	AREENVLOTOR		a contract of the property of the	1	6
MILET ALL AND	MTH1478 Mta 2622593	3	RTDPH	THEVYS	CDARG	3	ETTREASAU	SLDAV	NUNDANTME	49	ETHDODGVAL	TRUPEVEVEN				9
CALL ON COLUMN THE OWNER AND ADDRESS OF A DEPENDENCE OF A DEPE	MT1587 Mi 2496191	2	KADL	THURKYS	TGKE	13	NTLEVAKER	GIEVY	ATTDHNTTR	49	KIKEOGGLAT	APPYSPICKA				
HP1573 Hp 2314758 2 FIDTHCHLDHKDYEND 2 EVEKELEKGYTOCYTPGADMKD 8 EKFEGYPFAIGABPYDVESFD	HP1573 Hp 2314758	2	FIDTH	CHLDHKI	YEND	2	EVLKESLEK	GVTOC	PGADMKD	8	EKFEGVFFAL	GARPYDVESFD				

**Figure 1.** The four conserved motifs in the PHP domain. The alignment was constructed using the MACAW program and adjusted on the basis of the PSI-BLAST search results. The numbers indicate the distances from the protein N-termini to the first aligned block and the distances between the blocks. The sequences are grouped by similarity-based clusters identified using the CLUS program: 1) N-terminal domains of the bacterial DNA polIII α subunits; 2) group of stand-alone proteins from bacteria and archaea that show greater similarity to the polIII N-terminal domains than other members of the superfamily; 3) C-terminal domains of family X DNA polymerases from *B.subtilis* and *M.thermoautotrophicum* and highly similar stand-alone proteins from bacteria and archaea; 4) histidinol phosphatases from two yeast species and their apparent orthologs from *Llactis* and *B.subtilis*; 5) a group of proteins encoded in operons involved in capsular biosynthesis in *Streptococcus*, *Staphylococcus* and *B.subtilis*. Group 6 includes all sequences lacking a detectable motif IV. The shading is based on a 85% consensus. The consensus is shown above the alignment; h indicates hydrophobic residues (A,C,F,I,L,M,V,W,Y; yellow background), s indicates small residues (A,C,S,T,D,N,V,G,P; blue background), p indicates polar residues (D,E,H,K,N,Q,R,S,T; brown coloring), o indicates hydroxy residues (S,T; green coloring), and '–' indicates negatively charged residues (D,E; purple coloring). The putative active site residues or those involved in metal chelation as described in the text are shown by inverse red shading. The yeast histidinol phosphatase encoded by the his-2 gene is designated HIS9 after the SWISS-PROT database since this is the enzyme catalyzing the ninth step of histidine biosynthesis. The species abbreviations are: Aae, *A.aeolicus*; Af, *Archaeoglobus flugidus*; Ath, *Anaerocellum thermophilum*; Bb, *Borrelia burgdorferi*; Bs, *B.subtilis*; Ct, *Chlamydia trachomatis*; Ec, *E.coli*; Hi, *H.influenzae*; Hp, *Helicobacter py* 

family; 3) C-terminal domains of family X DNA polymerases from *B.subtilis*, *A.aeolicus* and *M.thermoautotrophicum* and highly similar stand-alone proteins from bacteria and archaea; 4) histidinol phosphatases from two yeast species and their apparent orthologs from *Lactococcus lactis* and *B.subtilis* (for the *L.lactis* protein, this functional prediction is compatible with its location in the vicinity of and probably within the same operon with other histidine biosynthesis genes) (22); 5) a group of proteins encoded in operons involved in capsular biosynthesis in *Streptococcus, Staphylococcus* and *B.subtilis* (23).

The presence of the PHP domains in two distinct polymerase families, especially its ubiquity in the  $\alpha$ -subunits of PolIII, is of particular interest and may have important functional implications. The polymerase reaction involves the transfer of a nucleotide monophosphate from a deoxynucleoside triphosphate to a primer,

accompanied by the release of inorganic pyrophosphate. The equilibrium of this reaction is shifted in the direction of polymerization if the pyrophosphate is removed as it is produced by the polymerase reaction. Pyrophosphate hydrolysis coupled with polymerization is thought to be catalyzed by a phosphatase that has not yet been definitively identified in any system; inorganic pyrophosphatase, which is one of the most abundant phosphatases in all cells, has been tentatively implicated (1). We hypothesize that the PHP domain fused with the domain responsible for nucleotide polymerization may carry out this phosphatase function in bacterial polIII. PolIII  $\alpha$ -subunits from a wide phylogenetic spectrum of bacteria, including *Aquifex, Deinococcus, Chlamydia, Spirochaetaceae, Helicobacter, Synechocystis*, and at least one copy in the Gram-positive bacteria, contain all the conserved residues of the PHP domain that are predicted to participate in the phosphatase



Figure 2. Distinct domain architectures of proteins containing the PHP domain. The figure is roughly to scale, except for the large DNA polymerase III α subunits. PHP N-terminal and PHP C-terminal refers to the N-terminal portion of the PHP domain (motifs I–III) and the C-terminal part (motif IV), respectively. HhH indicates a helix–hairpin–helix DNA-binding domain (35).

reaction (Fig. 1). However, in the polymerase from Proteobacteria, such as Rickettsia, Escherichia coli and Haemophilus influenzae, some of these residues are replaced, and in the case of Mycoplasma, some of the conserved motifs are additionally disrupted by deletions (Fig. 1). In Gram-positive bacteria, there are multiple copies of the polIII  $\alpha$ -subunit. In *Mycobacteria*, both copies have intact proposed catalytic sites, but in one of the three Bacillus polymerases, they appear inactivated, and in the Mycoplasma, both copies are predicted to be inactive. Even though the PHP domain in *E. coli* polIII is predicted not to possess phosphatase activity, deletion mutagenesis experiments have shown that the N-terminal 60 amino acids are strictly required for the polymerase activity (17). It seems likely that the apparently inactive copies of the PHP domain perform an allosteric regulatory function, perhaps including pyrophosphate binding. Notably, Proteobacteria contain stand-alone versions of the PHP domain (e.g. E.coli YE00) that are predicted to possess the phosphatase activity and might supply this function to polIII in trans.

The PHP domain in the X-family polymerases may have the same function as proposed for the PolIII  $\alpha$  subunit. However, this may not be the only explanation for their presence. Given the possibility that these DNA polymerases could be involved in repair, it cannot be ruled that in this case, the phosphoesterase domains function as nucleases. In the case of the PolIII  $\alpha$  subunits, a functional nuclease is unlikely given the lack of evidence for any nuclease activity except for that of the well

characterized 3'-5' exonuclease domain (Fig. 2), in spite of the detailed experimental studies on this protein (1,16).

## The small subunits of the archaeal DNA polymerases and eukaryotic DNA polymerases $\alpha$ and $\delta$ are members of the calcineurin-like phosphoesterase family

The small subunits of the eukaryotic DNA polymerases  $\alpha$ ,  $\delta$  and  $\epsilon$ showed moderate but statistically significant sequence similarity to their apparent orthologs from the four completely sequenced archaeal genomes (in a typical search, an e-value below  $10^{-3}$  was observed in the second or the third iteration, but some eukaryotic polymerase subunits showed a greater similarity to the archaeal homologs; for example, the human polymerase  $\delta$  subunit produced an e-value below  $10^{-4}$  in the first pass of the search). Inspection of the alignment of these archaeal proteins revealed four conserved motifs, which define the vast superfamily of calcineurin-like phosphoesterases (19,24,25; Fig. 3). These motifs contain conserved histidine and aspartate residues involved in metal coordination and catalysis (24). These catalytic motifs, however, are disrupted in the eukaryotic polymerase subunits, in spite of the co-linear alignment throughout the entire eukaryotic and archaeal sequences (Fig. 3). These disruptions notwithstanding, when iterative PSI-BLAST searches were performed with the eukaryotic polymerase subunit sequences as queries, members of the phosphoesterase superfamily were retrieved from the NR database

	MOTIT I MOTIT II
DNApolIIs1_Pf_2967436 DNApolIIs1_Ph_3130366 WTH1405_Mta_2622517 AF1790_Af_2648756 MJ0702_Mj_2128452	303 KVLPDAVVAFKOVYSKRG-ILVANKFYLPDVPLY-RRQXPPLEKVYALIJSDIHVGSKE-FCENAFIKFLEWLNGNVETKEEEEIVSRVKYDIIAGUVUD 312 KVLFDAVVAFKOVYSKRG-IFFANRFYLPDVFLY-RKQXPFLEKVYALISDIHVGSKE-FCEKAFIKFLEWLNGVVSKEEEEIVSRIKYLIAGUVUD 183 KIVFDAVVAFKOTYSKGR-VVASEIFHFGVPRUCKENDFSVAFISDVHISQV-FLEDAFMKVKVKUNGDGSEQORSLAADVXLVVAGUIVU 186 ELLGDEVIGVTCLLKGSS-LYANRIVFPDVPIN-GNGEKKRDFYIVFLSDTHFGSKE-FLEKEWEMFVRWLKGEVGGKKSQNLAEKVXIVIAGUIVU 280 DILLDEVIGAIGTVSKSGSSIVVDEIIRPALPPKEPKRIDEBIYMAFLSDIHVGSKE-FLKKEFKFIRFLNGDVDNELEEKVVSRLXYICIAGULVD
DFD2_At_2673906 DFD2_Hs_1352307 DFD2_Mm_2398725 HYS2_Sc_1170421 Cdc1_Sp_2052244 DFB2_Hs_2637123 DFB2_Hs_2632262 DFB2_Sc_2506366 DFD2_Mm_461957 DFD2_Sc_585063 DFD2_Sp_3169066 consensus/90%	148 YUTCYVVALHEKETNAG-EFFVEDULEAGLEPQ-TERFIDLOEDKYVYLISGLCIESKS-ANPLOPCLLVDHITGHLGDEEDQGLAADIV:VVIAGNSFE 152 LVTCTVLAVFSVRDG-KFLVEDYCFADLAPQKPAPPLDTDRFVLSGLGLGGGGGESLLGTQLLVDVTGQLGDGGQCSAAHVSRVIAGNLLE 154 LVTCT-VLAVLGSAKDDG-KFQVEDHCFADLAPQKPAPPLDTDRFVLSGLGLGGGGGESLLGTQLLVDVTGQLGDGGQCSAAHVSRVIAGNLLE 155 LVTCT-VLAVLGSAKDDG-RFQVEDHCFADLAPQKPAPPLDTDRFVLSGLGLGGGGGESLLGTQLLVDVTGQLGDGGQCSAAHVSRVIAGNLLE 166 FITGV-VUGILGMERAEG-TFQVLDICYFTPLPQ-NPFPAPIATCFTRGKIAVSGLNLKNTSPDRLLRLEILKEFLKRFLMGRINNKIDDISLIGRLICGNSVD 138 VVTCV-VLAVLGHEDQG-RFVVDVCPGIFSHSIPMTDESQ-AQPEYIAAVSGLSLSNDG-IEGIQVHQLVDFLRGTLPH-VSSFSPSSIRKLIILGNCLA 147 TEFSSTTRAYY GNINFFG-GFSNASVKTYKKLKQLEEENKDAMFVFLSDVNLQVE-VLEKLHIMFAGYSPAPPTCFLLCGNFSS 244 TEFSSTTRAYY GNINFFG-GFSNASVKTYKKLRQLEDENKDAMFVFLSDVNLQVQ-VLEKHNYPGSYSPAPPTCFLLCGNFSS 245 ESAEVTRKELSSNENNFGG-GDKIAFRCSDRLRSALAKQEDTSLVFLSDVNLDQVG-VLEKVFLLQGYKDQPPVATVFCGNFCS 256 ESICGNLDLGI-HGS-NNNFIARLDKDLKIRLHLLEKEL-TDHKFVILSDVLDVLRVKLLQKLNDQPPTLING GFTS 257 IFGNLDLGI-HGS-NNNFIARLDKDLKIRLHLLEKEL-TDHKFVILSDVFLDDLK-IMTALSKLQKLNDQPPTLING GFTS 258 LFFCQ-VVIMEGFNTTGRRLTATFKLYESVELFYQPTEEBGAGEGYXVVVRAF GYTTSDS-ITTDPLLDLIAIINRQCPDVCIFG GPELD 259 NANGO-YFTWSILFLEYFN-NSPVSTSGLQEFVANTNNQPISTYXAGFWSLRDG-LSFSPLKSMISYNKNPVDLVICGFFLD 254 NPSCN-MFIAKEILPIPPL-PFPSSSQEHATFVANTNNQPISTYXAGFWSLRDD-LSFSPLKSMISYNKNPVDLVICGFFLD 254 NPSCN-MFIAKEILPIPPL-PFSSSQEHATFVANTNNQPISTYXAGFWSLRDD-LSFSPLKSMISYNKNPVDLVICGFFLD 256 N
DNApolIIS1_Pf_2967436 DNApolIIS1_Ph_3130366 MTH1405_Mta_2622517 AF1790_Af_2648756 MJ0702_Mj_2128452	Motif III GUGVYPGQYADLTIP
DPD2_At_2673906 DPD2_He_1352307 DPD2_Mu_2398725 HYS2_Sc_1170421 Cdc1_Sp_2052244 DPE2_Hs_2697123 DPE2_Hs_2697123 DPE2_Hs_2697123 DPD2_Mu_461957 DPD2_Mu_461957 DPD2_Sc_2506366 DPO2_Sc_385063 DPO2_Sp_3169066 consensus/90%	PPRKLINGQVLITFNLASKDQSTLYEP-IKELDIMLSQIAAGVSVDTYPTND-PANPALPQQPLNRCLPPGSSPYNTFRSCTNPHSPDVDNI HSTQSRDSINKAKY-LTXKTQASVEAVKMLDEILLQLSASVFVDVPGEFD-PINYTLQQPLHPCMPPLATAYSTLQLVTNPYQATLGV HNTQSRDSINKAKY-LTXKTQASVEAVKMLDEILQLSASVFVDVPGEFD-PINYTLQQPLHPCMPPLATAYSTLQLVTNPYQATLGV FDIKSVNKDSVEAVKMLDEILQLSASVFVDVPGEFD-PINYTLQQPLHPCMPPLATAYSTLQLVTNPYCATLGV PSIELADSASASVTGKKKVKNGVDTSAVNNNPTFQLDNFLDQVCSSILVTXPAQFVG-PSD-PGFGSILFQPLATAYSTLQTVTNPTFLSCAH APYGKNQVQALKDSL
DNApolIIs1_Pf_2967436 DNApolIIs1_Ph_3130366 MTH1405_Mta_2622517 AF1790_Af_2648756 MJ0702_Mj_2128452	Motif V         Motif V           DFLIMIGRGIEDVVGSVPGLTH         6         MVELLKMERVAPMFGGKVPIAPDPEDL-LVIEVPDUMGGVUVVDAVVRGVQLVNSATWQAQTEFQKMVNIVPTPAK         21           DFLIMIGRGIEDVVGSVPGLTH         6         MVELLKMERVAPMFGGKVPIAPDPEDL-LVIEVPDUVGGGVUVVDAVVRGVQLVNSATWQAQTEFQKMVNIVPTPAL         21           RTLIVHGRSFDDMAMSVNGLSH         6         MVELLKNERVAPTGGKVPIASEEDH-LVIEVPDUVGGGVUVVDAVVRGVQLVNSATWQAQTEFQKMVNIVPTPAL         21           KVLIVHGRSFDDIJSKIPRLSY         6         MEELLKRERVAPTGGKVPIASEEDH-LVIEUPPULHCGUILWGGTGYRGVPMVNSSTWQAQTEFQKKVNLNPMFGN         14           KVLIVHGRSFDDLVGQIRAASY         6         MEELLKRERVAPTGGRCPLAPEEDV-LVIEUPPULHCGUILWGTGYRGVPMVNSSTWQAQTEFQKKVNLNPMFGN         14           DTLIVHGRSFDDLVGQIRAASY         6         MKELIKRRULCPTYGGRCPLAPEEMV-LVIEUPPULHCGUILWGTGYRGVPMVNSSTWQAQTEFQKKVNLNPMFGN         14
DPD2_At_2673906 DPD2_Hs_1352307 DPD2_Mm_2398725 HYS2_Sc_1170421 Cdc1_Sp_2052244 DPE2_Hs_2697123 DPE2_Hs_2697123 DPE2_Ce_1255340 DPD2_Ks_2506366 DPO2_Mm_461957 DPO2_Sc_585063 DPO2_Sc_3169066 consensus/90%	RFLGTSGONIDDLGKYSEAKSK       3       VERTLRWRHLAPTAPNTLGCYPTDRDPFLIETCPHVYFVGNQDKYDNRLIKGSBGQLVRLICI-PKFCET-GIAVAVNL 17         RFLGTSGONYSDIFRYSSMEDH       3       LEWTLRVRHSPTAPDTLGCYPFYKTDPFIPECPHVYFCGNTPSFGSKIIRGPBQUVLLVTV-PDFSAT-GTACLVNL 27         RFLGTSGONYSDIFRYSSMEDH       3       LEWTLRVRHSPTAPDTLGCYPFYKTDPFIPECPHVYFCGNTPSFGSKIIRGPBQUVLLVTV-PDFSAT-GTACLVNL 27         RFLGTSGONYSDIFRYSSMEDH       3       LEWTLRVRHSPTAPDTLGCYPFYKTDPFIPECPHVYFCGNTPSFGSKIIRGPBQUVLLVTV-PDFSAT-GTACLVNL 27         DVLAVSGNINDICKYVIPSND       26       MECTMKWCNIAPTAPDTLGCYPFYKTDPFIPECPHVYFCGNTPSFGSKIIKGPBQUVLLVAV-PDFSST-GTACLVNL 27         FVLATSGONINDLKYYPFXS       MENTLUMNETPTSPDTLWCYYPTDKDTFYWHEEMPLJLCGNQFKFGCKTUINE-GNRIGLVSV-PEFRKT-GVLVLINK 27         BITVFREDLWNNKCRNCVRFPS       9       VKTILSQGHLPLPLYVCFYMWAY-DVALRVYPVPDLLVIADKYDPFYTNTECLCINFGSFPR-SGFSFKVFY-PSNKT 52         BIVFFRDLINNKCRNCVRFPS       9       VKTILSQGHLPLPLVCFYMWAY-DVTLKVYPVPDLLVIADKYDPFYTNTECLCINFGSFPR-SGFSFKVFY-PSNKT 52         BIVFFRDLSGRFKRRLEFFF       9       VKTILSQGHLPLPLVCFYMWAY-DVTLKVYPVPDLLVIADKYDPFYTNTECLCINFGSFPR-SGFSFKVFY-PSNKT 52         BIVFFRDLSGRFKRRLEFFF       3       NEFSTLSWDL-DHTLFLPPTSSWDL-DHTLFLCPIPSTWLCCTTAREFTFEKVGADTUSNPGSFSRNTFHVYYPCONR 52         BIVFFRDLSGRFKRRLEFFF       3       DRFSRVLKRLLQRRY-YPFYFPGSERF-SGIRY-FKYVPYPDLFUNDFVSQLFSAQHDLYNNCGCVNNPGSFIRNRYPHYPYPCORR 52         SUFFSNILLTDUCHTSS       3       DRFSRVLKRLLQRRY-YPFYFPGSERF-SGIRY-FKYVPYPDUPIVFVSQADTUSNPGSFSNSNTFHVYYPYPCOR 52

**Figure 3.** A complete alignment of the archaeal and eukaryotic DNA polymerase small subunits. The consensus was derived under the 90% rule. The designations and the convention used in shading are as in Figure 1; additionally, c indicates charged residues (D,E,K,R,H; magenta coloring) and u indicates 'tiny' residues (G,A,S; green background). The conserved motifs typical of the calcineurin-like phosphatase superfamily are indicated, and the five predicted catalytic and/or metal-chelating motifs in the archaeal polymerases (top five lines) and the residues mutated in the two Hys2 alleles are shown by inverse type. The protein name abbreviations: DPD2, DNA polymerase  $\delta$  small subunit; DPO2, DNA polymerase  $\alpha$  small subunit. The archaeal species abbreviations are as in Figure 1. Eukaryotic species abbreviations: At, *Arabidopsis thaliana*; Ce, *Caenorhabiditis elegans*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Sc, *S.cerevisiae*; Sp, *S.pombe*.

at statistically significant levels [for example, in a search initiated with the sequence of mouse polymerase  $\delta$  small subunit (26), an Icc family member from *M.autotrophicum*, MTH1722 (Fig. 4), was retrieved with an e-value below  $10^{-3}$  in the fourth iteration].

To evaluate the phosphoesterase connection statistically and to determine the details of the phylogenetic distribution, we constructed a profile based on the multiple alignment of the known members of the calcineurin-like superfamily and carried out a systematic search of the NR database and available complete genomes. This procedure resulted in the retrieval of the entire superfamily, including the small subunits of the eukaryotic and archaeal polymerases, at a statistically significant level (at least  $e = 10^{-4}$  at the third iteration). In order to assess the robustness of the observed relationships, multiple searches were performed as described above for the PHP superfamily. The three-dimensional structure of calcineurin and the purple acid phosphatase have been determined (24,25), and the observed pattern of residue

conservation in the polymerase subunits as well as the results of alignment-based secondary structure prediction are fully compatible with the presence of a calcineurin-like phosphoesterase core domain (Fig. 4).

The calcineurin-like superfamily consists of enzymes with diverse functions, including protein phosphoserine phosphatases, nucleotidases, sphingomyelin phosphodiesterases and 2'-3' cAMP phosphodiesterases, as well as nucleases such as bacterial SbcD or yeast MRE11 (19,27). The superfamily is defined by five conserved blocks, which center around the metal-chelating residues; the four most conserved blocks are shown in Figure 4. We detected all the members of this superfamily from completely sequenced genomes and classified them into groups on the basis of sequence similarity; the polymerase subunits belong to a family typified by the bacterial Icc-like proteins whose functions are not known (Fig. 4). Given the absence of evidence for nuclease activity associated with the accessory subunits of archaeal or

		Motif I		Motif II		Motif II	II		Motif V	
Consensus 90%		hsD.H		hhGDh		h	3N		hhh.uH.H	
AF1790_Af_2648755 MTH1405 Mta 2622517	227	FYIVFLEDTHFGSK	24	NLAEKVKYIVIACDIVDGIG	28	PKEIKIIVSP	HDAVROAEP	88	EDVPDILHCCHIETYGTG	Archaean
MJ0702_Mj_2128452	323	IYMAFLS INVGSX	24	KVVSRLKYICIACDLVDGVG	28	PEHISIIISP	AVRPAEP	89	DRDIDILHTONINGYG	DNA polymerase
PolIIs1_Ph_3130366	355	VYAVLTSDIRVGSX	24	EIVSRIRYLIIACOVUDGIG	28	PKHITIFIGP	NHDAARPAIP	90	REVPOLYCMCHVMVYDTA	subunits
Mp11 Mg 1300175	12	VIALLINI INVOSA	29	STARAKINI MATANA AND AND AND AND AND AND AND AND AND	28	PARTIMPLAP	DDDDDDDDDDD	90	BEVPDVVHRMMVPVIDAV	<i>'</i>
SbcD_Ec_140195	1	MRILHTEDWHLOON	20	AQTHQVDAIIVAD VFDTGS	17	QTOCHLVVLA	SVATLNB	119	FPFADYIAL BURRAQII	
CDC1_Sc_550426	47	ISI <mark>V</mark> WILWLGLISY	70	QYYLDPDSNFFLODLFDGGR	21	PLRRTVMSLP	IN HE IGFGETV	111	KIQPE <mark>ILF</mark> SCDDMDHCQI	
ICC_EC_547704 F18F2.1 Ce 1418505	15	VRIDQITETELFAQ	22	CLACOFDETWHICTAYDLA	12	SFRAPCVWLP	HDFQPAMYS	112	PPHVRYLLCCHIHQELDL OCSUDECENERSYERE	
F21A3.2_Ce_2276162	84	FIYAVYCELGVENG	9	AQKGQLDMVLHVCDFAYNMD	17	AGYIPYMATV	IN HEYYNNFTH	102	EYGVDVELWAHEHSYERL	
T23G7.2_Ce_1132531	224	LKLAVISCLMAGAS	12	VLENPVDAILIVODMVDAPV	13	PARAPTY	AND YYYGDVQ	79	LSEID <mark>IVL</mark> SCHTHAGOFY	
YLR361c Sc 1077358	248	FRIVOLADLHLGVG	27	LDIEKPOLVVFTSPOINGDR	17	ARKIPWAMVW	HDDEGSLTR	141	RLSVDVVSCGHDHCNDYC	
MTH178_Mta_2621221	2	VLIAHISDLHVGAP	13	INKLEPDATVIT DUTDNGY	12	DLKGPHIFVP	MHDARHVGNE	89	SKGADINI SGEREVPPRL	
MTH1722_Mta_2622852	3	KKIAQISOVHFGEK	13	LENENPOLIIVSCOLTTEGY	14	RTITSTYTIP	IN HDARNVGLI	88	EYGVDFVLNGERHVPNVW	
MTH1002_Hta_2023020 MTH541_Mta_2621616	171	YRFAHLSCHLGAO	19	ALOKDVDFMIIACLFHSNI	17	EAGVPIYVNY	1911 YSPSSTS	94	PROFEYYAGEIVIERKGCY	
MTH1179_Mta_2622286	130	LRF <mark>VQLSDIHLGTV</mark>	12	VRRVEAAAVLITEDLLDGSR	10	ETGAPIF	NHDTY SGDFR	64	ERGVELQLSCHTRGGQFY	
MTH1774_Mta_2622904	1	MLIGVISDTHIPDR	. 9	DAFRDVELILHAS DLTSPDI	4	ETLAPVECVQ	MORHYGIET	34	ELGADVLISCHTHQPFIR CORVERNMENTER FR	
MTH1803_Mta_2622935	20	EDSMVIADLALGYE	25	RDVSGASGIIINGLKHEFG	18	ENPRDIT	MEDALIPYMS	26	DLDAETIICHE PCVGL	
MTH1492_Mta_2622608	1	MNDTHGYIK	29	IRKERPALLDCS DTINGTY	13	LNEMGFRAMT	AHNEFAYGPER	104	VEGICVLLSANTWNRLER	
AF0876_Af_2649728 AF1790 Af 2648756	24	ITILHTNDEHSALI	18	EKDKONVILVSASDYIGGTP	16	QKVGYDVVTI PKEIKIIVSP	HEFDYTSKA	101	VPBIDVIIACHDHRVLEE EDVEDINGELIHTVGTG	
AF2433_Af_2650659	1	MRIVHISDLHFGEE	12	INEIEPDLLVITSCLSCWGI	12	DLKPDYIAVP	MHDARNIGYE	47	REDRENTLASHELIAPIP	
AF1031_Af_2649563	1	MKFAHIADVELGYE	21	AVESNADP <mark>VVIAGPL</mark> FHRSL	17	KENI PVF <mark>A</mark> VE	NHOKT SRDIS	97	LPEAS <mark>YYA</mark> F <mark>GH</mark> I <mark>H</mark> LPKIY	
AF1821_Af_2648726 AF2181_Af_2648343	17	RETLAVARTELEEV	11	IEREIEGLAIFLEFYADRGD ELMEGADFVVHAGFFRSYKV	12	FSDYELYAVA	NINSTGVYPH NSDDDXIKEE	106	ELGVINIIRAHEPOKVLK ELGVINIIRAHEPOKVLK	
AF1154_Af_2649433	15	GNTAVIADLELGFE	25	VERYGIAKIVVACELKHEFS	15	SVNVELEVVR	MEDNYLAAIL	23	DCREEKVIMOREPAVKV	
AF0799_Af_2549807	4	MRILIFEDTHIPER	.9	DYLVDFDMVVITEDLTSERV	.5	RVAESVIAVR	MDDLPLPHS	32	EMDVDVLISOFTHLPDVY	
AF0919 AE 2649682	1	MRFAVVSTNDNLS	8	LSKEKLDFVVHADVIAPFT	5	REGVELYTAF	NNDGERRILT	33	KKLETINVFOITHEVLKE	
AF0533_Af_2650087	1	MRVLLLEDIKSSTQ	5	LKAAEYDA <b>VF</b> IAC <mark>CL</mark> TQFRP	11	EQTDSCLAVE	INCRHEAILGE	84	MIEFDVILCONINESHGV	
MJ0912_Nj_2128575	4	SMIAVISTIESNLE	8	IKNRGIKKIFCLEDIVGYGA	8	IRDLNCLSVV	YGVLGKE	68	FNYGDLIFVERSTIPFVN	
MJ1323_Nj_2128792	1	MMFVHIALNELGYR	21	ILEIKPDVVLHSEDLFNDLR	17	ENNIKVYIVA	MEMPRRLGE	82	LPKF SYYALCHIEKRI LE	
MJ0037_Mj_2495754	13	KDY <mark>AIIADTHIGFD</mark>	25	IDKYKINNLIINE IKENFK	16	REVINVILIE	IN HOTFISSAG	26	LLKERTWILGHERPSIKL	
MJ1445_NJ_2128862 MJ0936 NH 2501611	4	REITLINGTOPHA	25 8	NVVRDKDIVYPLADDILSKN ENDENVETVIHCEDEVSLEV	10	LLNGEIVFIR	AND GEBCKLK	20	SQLVENUT VALUE SVER	
MJ1162_Mj_2129069	ĩ	MKIVGITELGKLP	4	EFKDFADVLVVC	11	SDYMEVICVP	NCOTKEVIDE	92	NENIR <b>FCA</b> CONTRESECT	
MTCY07H7A.4c2117277	14	YVLLHISTHLIGG	22	QSGLRPDAIVFT DLADKGE	16	QLGAELVWVM	MHD DRAELRK	90	GTDVRAILACHLHYSTNA	
HP1044_Hp_2314191 YARI Ec 1552742	49	PKILFLAULWSRF	12	GIRCKPDLTLLGG TVLEDM	14	AECAPTEACE	NERPVGTEK	69	DRPNDIMLOGHTHGGOLR	
TPase_Ec_1786234	1	MATYLICOWGCYD	9	EFTPGKDTLMLTCOLVARGP	10	SLGDSVRLVL	NHOLHLLAVF	35	IDEEKKLVMANAGITPQW	
FRPA_Ec_2499737	18	RHIWLSCHINGCLE	9	RFDFWRDLLISV: DVIDRGP	8	LEQHWVCAVR	NHEQUANDAL	88	ITGADHEWF HTPLRHRV	
PRPB_EC_2499738	15	REINVVGLIEGEYO	- 21	SFFFKIDLLISVENIDRGP	8	LNOPWFTSVK	NHEAMALEAF	43	TNDNIKYAIAHADYPGSE	
ushA_Ec_1786687	34	ITVLHTNDHIGHFW	23	AABGGSVLLLSGCDINTGVP	14	NLVGYDAMAI	IN HEFDNPLTV	116	AGSLAMINGERSODFVCH	
CpbD_Ec_1790658	24	LRIMETTOLISNMM	45	LADYMSAKGLKACDIHPVYK	10	NTLDYTVGTL	IN HEFNYGLDY	121	IPGVNAIMFCHAHAVFPG	
Y207_Mpn_2146381	2	TKVLVLSTTGYND	5	MKLHNPDVVIHAGHL/TTKK	0	MDQNATFWVA	NHEVVGEEIQ	34	SYLCOVLIVENSHIEHYE	
YhaO_Bs_2226129	4	DLTPHAADLHLDSP	29	AVREHVDPILLAC DLFDEAN	17	ECGISVYVIF	IN HIDHLGGEWT	84	KSGMD <mark>YWALCHIH</mark> KRQVL	
YHCR_BS_1724012	390	LRILSMNFLEGKID	30	RAEKKNSLIVHACTMIGGSS RRESDEDELLUTGTAKKGD	15	EDIGPDVGTV	SNHEFDEGTDE	124	DSEIDVIFAARNHOVVNG DVRIDIVIAARNHOVVNG	
YjbP_Bs_2633517	1	MAYDIISTIGCYD	17	FVHEEGRVLVFAC DLTDRGP	12	YERGAVRYVP	INHCNELYRYL	97	YNGKAWWYYNTEWTPVKEPR	
YokB_Bs_2529446	1	MKIDYVSDLHINHW	21	ISNGNGEV <mark>LV</mark> IAGPTEWNQ	20	RQYERVYPTY	IN ROLYLLSKS	126	FINARHWYCCHDHLQAEP	
YSNB_BS_2501608 VkuE Bs_2632226	3	MNVLIIS SIGLEE	12	RHEAEVOLMINCS SELETR	15	PALEPYAVVK NAPECEL CTV	NOFAGDFRD	33	ELGADVICFCHSBIAGSE EXPANIOLSER/THCCOTO	
YkoQ_Bs_2632055	46	LNILHLEDLHLENI	10	TKDQFVDIIALTOPLDRKR	15	KPAYGMYAVF	NHEYVLXEED	71	CVHYDYLLS PF0GGQIH	
YPBG_Bs_1730889	36	LTIFFICTURELI	7	ARSHAPHLVIIG DLAEGGV	13	HFGVPIVEVW	NN YEVROHK	70	DDGIDVILSCHTRGGQIR	
YTRA_88_2618853 YfkN Es 2626826	920	LTVMHTNTTAHLD	16	RSETNHNILLDA DVF SQDL	16	NMMGYDAMPP	SNHEFDRGPTU	135	VEGIDLIIGEPTETLVDK	
SbcD_Bs_548899	1	MRILHTADWHLGKT	20	VKDEQIDAIVMACDAFDTVN	18	RGKRPIVVIA	IN HIDNFORLSA	116	PADAAYVALCHLERPOTI	
YunD_Bs_2635734	5	LRLYHTNELHSHFE	16	QSDGEETLVFDIG HLDRFQ	15	NRLHIDGAAI	SNN GITLPHE	101	VPEIDVILENTTHLLED	
laui_Hs_306477	83	APVTVCCTTIGOPP	8	GGSPANTRYLFLEDYVDRGY	14	LYPKTLFLLR	CRHL/TRY	96	HNNLLSILRAFEAODAGY	
Pag Strang		PPPP UNU		PPPPP		PPEPP	UUUUUUUU		UUU PPPP PP	

Figure 4. The four most conserved motifs in the calcineurin-like phosphoesterase superfamily. The rules for consensus derivation and shading are the same as in Figures 1 and 3. A fifth, short conserved motif was omitted for compactness. All the representatives of this superfamily from the three completely sequenced archaeal genomes as well as *E.coli* and *B.subtilis* are shown in addition to the archaeal DNA polymerase subunits (the top three lines), in order to indicate the diversity and the comparable representation of this superfamily in archaea and bacteria. The bottom two lines are the sequences of purple acid phosphatase from *P.vulgaris* (Pv) and human calcineurin, for which X-ray structures have been determined. The consensus secondary structure derived by comparison of these two structures is shown underneath the alignment; E indicates extended conformation ( $\beta$ -strand) and H indicates  $\alpha$ -helix.

eukaryotic DNA polymerases (1,5), it seems likely that similarly to the PHP domain in bacterial polIII, the archaeal polymerase subunits possess phosphatase activity and may be required for pyrophosphate hydrolysis as discussed above. Recent studies on archaeal DNA polymerases have shown that the orthologs of the eukaryotic small subunits are in fact the small subunits of a novel DNA polymerase (designated DNA polymerase II) whose large, catalytic subunit is unrelated to any other known polymerases (28,29). Notably, it has been demonstrated that the archaeal DNA polymerase II possesses a 3'-5' exonuclease activity that requires a complex of the two subunits. These striking new findings call for caution in the interpretation of the exact nature of the predicted phosphoesterase activity, and make experimental analysis of these proteins particularly interesting. The eukaryotic DNA polymerase subunits with disrupted catalytic motifs have clearly lost the enzymatic activity. In mammalian DNA polymerase  $\delta$ , the small subunit is required for the interaction with the clamp protein PCNA (30). The high conservation of PCNA in archaea (31) suggests that this interaction is an ancestral function in archaea and eukaryotes, and accordingly, the archaeal DNA polymerase small subunits may possess dual function, namely pyrophosphate hydrolysis, resulting in increased polymerization rate and PCNA binding. In eukaryotes, the phosphatase function may have been lost due to displacement by alternative phosphatases, but selection for the interaction with the clamp subunit could have resulted in the conservation of a structurally intact protein without enzymatic activity. Recently, the yeast polymerase  $\delta$  small subunit, which is encoded by the

Hys2 gene, has been characterized (32). The mutation in the Hys2-2 allele affects the conserved aspartate in motif II of the phosphatase superfamily (Fig. 3). This suggests that despite the disruption of most of the predicted metal-chelating sites in the eukaryotic polymerase subunits, coordination of a divalent cation by this aspartate together with the conserved histidine in motif V could be important to maintain the structure of these proteins. Similar functions may be proposed for the less thoroughly characterized small subunits of DNA polymerases  $\alpha$  and  $\varepsilon$  that are paralogs of the polymerase  $\delta$  small subunit (33 and the present work) and like the latter, appear to be an inactivated phosphatases (Fig. 3).

## Functional and evolutionary implications of the association of phosphoesterase domains with DNA polymerases

The analysis described here resulted in the identification of two unrelated phosphoesterase domains in four distinct types of DNA-dependent DNA polymerases, namely: (i) the N-terminal PHP domain in bacterial polIII  $\alpha$  subunits; (ii) the C-terminal PHP domain in bacterial and archaeal X family DNA polymerases; (iii) an apparently active calcineurin-like phosphatase superfamily domain in the small subunit of archaeal DNA polymerase II; and (iv) an inactivated calcineurin-like phosphatase superfamily domain in the small subunits of eukaryotic DNA polymerases  $\alpha$ ,  $\delta$ and  $\epsilon$ . A striking parallel between the PHP domain and the calcineurin-like phosphatase domain is the existence, in addition to the forms containing all predicted catalytic residues, of apparently inactivated versions. The homologous relationship between the domains predicted to possess phosphatase activity and the inactivated ones is strongly supported statistically. Given the multiple motifs that are required for catalytic activity of phosphoesterases but are disrupted in the apparently inactive forms (e.g. Fig. 3), it is clear that in evolutionary terms, the active enzyme is the ancestral state. It appears most likely that in both instances, the polymerase-phosphatase association has evolved through the recruitment of an active phosphatase. For the PHP domain, this scenario is further supported by the presence of a domain predicted to possess phosphoesterase activity in the polIII  $\alpha$  subunits from all major bacterial lineages, including the ancient Aquifex branching, whereas the apparently inactive forms are found in only two, namely Proteobacteria and Gram-positive bacteria (Fig. 2). In the case of the eukaryotic polymerase subunits, it appears that there had been an ancestral inactivation of the calcineurin-like phosphoesterase domain, followed by a series of duplications resulting in the extant subunits of the  $\alpha$ ,  $\delta$  and  $\varepsilon$  polymerases.

Fusion of phosphoesterase domains with polymerases is not limited to the cases discussed here. We had previously described a fusion of the polyA polymerase domain (polymerase X superfamily) to a newly defined phosphoesterase domain called the DHH domain (34). In addition, we also noticed the fusion of another novel phosphoesterase domain to the tRNA-CCA adding enzyme, also of the polymerase X family (L.A. and E.V.K., unpublished observations). These multiple and obviously independent fusions suggest some selective advantage(s) conferred by the phosphoesterasepolymerase association. Pyrophosphate hydrolysis resulting in a shift of the reaction equilibrium towards nucleotide polymerization is a plausible unifying explanation that deserves experimental testing, though alternative interpretations also should be considered. Specifically, the possibility exists that some of the phosphoesterase domains actually possess nuclease activity. Furthermore, the independent disruption of the motifs implicated in catalysis in two

classes of polymerase-associated phosphoesterase domains, whose function, however, remains essential for the polymerase activity, may even suggest that their primary role may be non-enzymatic. Such roles may include regulation of polymerase activity, perhaps through pyrophosphate binding and/or interaction with other subunit of the holoenzyme, like the clamp subunit in archaeal and eukaryotic polymerases. Experimental elucidation of the roles of active and inactivated phosphoesterase domains in polymerases will hopefully reveal fundamental but hitherto unnoticed aspects of DNA replication in all known systems, and clarify the relationship between the predicted enzymatic and regulatory functions of these domains.

#### REFERENCES

- 1 Kornberg, A. and Baker, T. (1992) DNA replication. 2nd edition, WH Freeman, San Francisco.
- 2 Friedberg,E.C., Walker,G.C. and Siede,W. (1995) DNA Repair and Mutagenesis. ASM Press, Washington, DC.
- 3 Braithwaite, D.K. and Ito, J. (1993) Nucleic Acids Res., 21, 787-802.
- 4 Mushegian, A.R. and Koonin, E.V. (1996) Proc. Natl Acad. Sci. USA, 93, 10268–10273.
- 5 Hindges, R. and Hubscher, U. (1997) *Biol. Chem.*, **378**, 345–362.
- 6 Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W.
- and Lipman, D.J. (1997) *Nucleic Acids Res.*, 25, 3389–3402.
  7 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) *Nucleic Acids Res.*,
- **22**, 4673–4680. 8 Schuler G D. Altschul S F. and Lipman D I. (1991) *Proteins* **9**, 180–190
- Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) *Proteins*, 9, 180–190.
   Neuwald,A.F., Liu,J.S. and Lawrence,C.E. (1995) *Protein Sci.*, 4, 1618–1632.
- Neuwau, A.F., Eud. S. and Lawrence, C.E. (1995) Protein Sci., 4, 1016–1032
   Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1994) Proc. Natl Acad. Sci.
- USA, 91, 12091–12095.
  Koonin,E.V., Tatusov,R.L. and Rudd,K.E. (1996) *Methods Enzymol.*, 266, 295–322.
- 12 Rost, B. and Sander, C. (1994) *Proteins*, **19**, 55–72.
- Rost,B., Schneider,R. and Sander,C. (1997) J. Mol. Biol., 270, 471–480.
- 14 Wootton, J.C. and Federhen, S. (1996) *Methods Enzymol.*, 266, 554–573.
- 15 Holm,L. and Sander,C. (1995) Trends Biochem. Sci., 20, 345–347.
- 16 Yue, D., Maizels, N. and Weiner, A.M (1996) *RNA*, 2, 895–908.
- 17 Kim, D.R., Pritchard, A.E. and McHenry, C.S. (1997) J. Bacteriol., 179, 6721–6728.
- 18 Holm,L. and Sander,C. (1997) Proteins, 28, 72-82.
- 19 Koonin, E.V. (1994) Protein Sci., 3, 356–358.
- 20 Jones, E.W. and Fink, G.R. (1982) In Strathern, J.N., Jones, E.W. and Broach, J.R. (eds), *The Molecular Biology of the Yeast Saccharomyces*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. Vol. 2, pp. 264–299.
- 21 Malone, R.E., Kim, S., Bullard, S.A., Lundquist, S., Hutchings-Crow, L., Cramton, S., Lutfiyya, L. and Lee, J. (1994) *Genetics*, **137**, 5–18.
- 22 Delorme, C., Ehrlich, S.D. and Renault, P. (1992) J. Bacteriol., 174, 6571-6579.
- 23 Roberts, I.S. (1996) Annu. Rev. Microbiol., 50, 285-315.
- 24 Goldberg, J., Huang, H.B., Kwon, Y.G., Greengard, P., Nairn, A.C. and Kuriyan, J. (1995) *Nature*, 376, 745–753.
- 25 Strater, N., Klabunde, T., Tucker, P., Witzel, H. and Krebs, B. (1995) Science, 268, 1489–1492.
- 26 Zuo, S., Gibbs, E., Kelman, Z., Wang, T.S., O'Donnell, M., MacNeill, S.A. and Hurwitz, J. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 11244–11249.
- 27 Sharples, G.J. and Leach, D.R. (1995) Mol. Microbiol., 17, 1215–1217.
- 28 Uemori, T., Sato, Y., Kato, I., Doi, H. and Ishino, Y. (1997) *Genes Cells*, **2**, 499–512.
- 29 Ishino, Y., Komori, K., Cann, I.K. and Koga, Y. (1998) J. Bacteriol., 180, 2232–2236.
- 30 Zhou,J.Q., He,H., Tan,C.K., Downey,K.M. and So,A.G. (1997) Nucleic Acids Res., 25, 1094–1099.
- 31 Brown,J.R. and Doolittle,W.F. (1997) Microbiol. Mol. Biol. Rev., 61, 456–502.
- 32 Hashimoto, K., Nakashima, N., Ohara, T., Maki, S. and Sugino, A. (1998) Nucleic Acids Res., 26, 477–485.
- 33 Li,Y., Asahara,H., Patel,V.S., Zhou,S. and Linn,S. (1997) J. Biol. Chem., 272, 32337–32344.
- 34 Aravind, L. and Koonin, E.V. (1998) Trends Biochem. Sci., 23, 17–19.
- 35 Doherty, A.J., Serpell, L.C. and Ponting, C.P. (1996) *Nucleic Acids Res.*, 24, 2488–2497.