Supplementary file

Additional Information about the mcBPPS Sampler. ***The input***. As input the user needs to provide a main alignment, a seed alignment for each subgroup and a hyperpartition table specifying to which of three partitions each subgroup is assigned (see flow chart in Fig. S1 below). Alternatively, in lieu of a hyperpartition, the user can provide a phylogenetic tree corresponding to the seed alignments (in Newick format), which the sampler will then directly convert into the corresponding hyperpartition (see Fig. 2 in the paper).

Main alignment. There is nothing special about the main input alignment *per se*. Hence, it can be built using any available multiple alignment method that will save the output in fasta format—keeping in mind, however, that some methods cannot handle vast numbers of sequences. (The mcBPPS program includes a routine to convert the fasta alignment into the cma format required by the program.) Using a vast number of sequences is important for: (i) increasing the power of the statistical analysis; (ii) analyzing multiple categories of proteins within an entire protein class without the user needing to edit the input alignment (the sampler automatically eliminates unrelated sequences that get included by mistake); and (iii) taking full advantage of the massive amount of sequence data that is now becoming available. Programs capable of aligning vast numbers of sequences include: (i) Sean Eddy's hmmalign program with the A2M output format option; (ii) PSI-BLAST, which was used to create an ~42,000-sequence input alignment for the analysis reported in Supplement 8; and (iii) the MAPGAPS program (1), which can create very accurate alignments of up to a million or more sequences. MAPGAPS requires as input a curated 'template alignment', which are available, for example, from the NCBI CDtree website (http://www.ncbi.nlm.nih.gov/Structure/cdtree/cdtree.shtml). MAPGAPS was used for the analyses reported here.

Seed alignments. The seed alignments can be extracted from the main alignment using a routine that takes as input the seed sequence identifiers for each subgroup. Each seed alignment need include only one sequence but the output is enhanced by including additional sequences. I suggest no more than a dozen. Ideally the selected sequences should be (presumed) orthologs that share high sequence similarity and that are from distinct phyla. This ensures that the conserved patterns that they share are due to strong selective constraints rather than to a recent common origin. The number of seed sequences can differ between contrast alignments.

Note that the mcBPPS sampler does not require that the proteins corresponding to each seed sequence alignment be experimentally characterized in any way. Instead, the primary criterion for selecting the seed sequence(s) is the user's perspective and interests. Using multiple seed sequences produces more informative output alignments inasmuch as the user can directly see that certain residues are conserved across distantly related organisms. This ensures that these residues are subject to strong selective pressure presumably associated with a critical biological function. Using multiple seed sequences also allows the sampler to generate (automatically) a slightly better consensus sequence than would a single sequence. Regardless of whether single or multiple seed sequences are provided for each subgroup, the sampler will focus its search on those divergent patterns and divergent subgroups that are most relevant to the seed sequence(s) and thus presumably of primary interest to the user. By contrast, designing a sampler to identify every statistically significant subgroup and pattern within a major protein class without focusing on specific sequences (besides taking a long time to run) would result in a massive amount of output of unknown relevance to the user. To avoid this, the user needs to obtain some prior information, of course, regarding the phylogenetic (and, ideally, biochemical) relationships between seed proteins. Such information is available in the literature or from various websites (e.g., the NCBI CDtree website) or can be obtained, for instance, by first applying a tree building method or running the scBPPS sampler (2) on the main alignment.

Hyperpartitions and phylogenetic trees.The user determines how the hyperpartition is configured based on the questions he or she seeks to address and based on both the phylogenetic relationships and biochemical properties of the proteins of interest. Fig. 2 in the paper shows how a hyperpartition can be generated from a phylogenetic tree, which is the usual starting point. Table 1 in the paper corresponds (more or less) to the phylogenetic tree for the subgroups associated with P-loop GTPases; (it includes with some minor adjustments to illustrate other aspects of the mcBPPS sampler). However, more flexible, phylogeny-independent hyperpartitions also are allowed in order to characterize subgroups that, based on their overall sequence similarity, are from different branches of the phylogenetic tree, but that, nevertheless, share certain co-conserved residues associated with shared biochemical properties. This is illustrated in Table 2, the interpretation of which is elaborated upon in Supplements 3 and 4. Note that, although the sampler could be modified to create the hyperpartition independent of the user (by treating the number of categories and their partition assignments as additional random variables), this would

prevent the user from focusing on those properties of the sequence data of primary interest and from examining the data from multiple perspectives (as is illustrated by the clamp loader analysis in Supplement 3). Note that if a user asks a meaningless question (via the hyperpartition) such as: "show me the residues co-conserved in subgroups A and B but absent from subgroups C and D?", when the answer is none, then the sampler will return nothing as expected—so no questions are off limits (as long as the hyperpartition restrictions described in Methods are satisfied). Finally, subgroups that contain relatively few sequences or sequences that are nearly identical are essentially pre-defined by the seed sequences. Hence, a Gibbs sampling strategy is inappropriate for hyperpartitions with such a high degree of phylogenetic detail.

Biological relevance. When the seed sequences for each subgroup are obtained from distinct phyla, this ensures that consensus residues have been conserved for 1-2 billion years. In the light of spontaneous mutation rates and the fact that structure is often conserved in the absence of significant pair-wise sequence similarity, these residues are subject to strong selective pressure to maintain associated (and typically unknown) biological functions. Thus the issue is not whether these conserved residues play functionally important roles, but rather whether the sampler can provide useful clues regarding these roles. To accomplish this, the sampler uses an indirect statistical approach similar to that used by classical geneticists prior to the emergence of more direct, biochemically-based approaches. Just as classical genetics seeks to infer, from patterns of inherited traits, the existence and various properties of previously unknown biological components (genes, chromosomes, the relative locations both of genes on chromosomes and of gene products within implicit biological pathways, etc.), the mc-BPPS sampler's aim is to infer, from patterns of co-conserved residues, the existence and various properties of currently unknown molecular components of proteins. This is illustrated in Supplements 2, 3 and 4, which also illustrate another application of the sampler: automated (and statistically-based) PROSITE-like classification of protein subgroups and annotation of the divergent patterns most distinctive of each subgroup.

Properties of the protein class that are determined by the sampler. Based on the input provided by the user, the sampler optimally assigns the sequences within the main input alignment to the various subgroups and determines the patterns that most distinguish each subgroup from the other subgroups. To do this, the sampler first generates both a consensus sequence for each subgroup (as a guide to pattern selection) and an aberrant sequence subgroup consisting of random sequences (implemented as a high number of random prior pseudocounts for that subgroup; this number is specified within the hyperpartition).

The mcBPPS sampler requires no prior patterns. Instead, because the seed sequences are (by definition) *bona fide* members of their corresponding subgroups, the sampler requires that patterns characteristic of the subgroup be conserved within the consensus collectively defined by the seed sequence(s). This constraint is not as stringent as it may first appear: For example, if the consensus at position *j* is an asparagine residue ('N'), then the ten following pattern sets are allowed at that position: 'N', 'SN', 'TN', 'ND', 'NQ', 'NH', 'STN', 'NDE', 'NEQ', 'NQH', 'NDEQ'. Without such a requirement, users would neither be able to examine the protein class from a relevant perspective nor to ask specific questions regarding those proteins of primary interest to them. Note, however, that the size and composition of each subgroup (i.e., whether it corresponds to a subfamily, family, superfamily, subclass, *et cetera*) and which sequences belong to each class is determined by the sampler albeit from the perspective defined by the user-supplied hyperpartition and seed alignments.

Analysis of Ras-like GTPases. The Ras-like GTPases—which include Ras, Rab, Rho/Rac, Ran, Arf, and Arf-like (Arl) GTPases and α subunits of heterotrimeric G proteins—function as on-off switches within eukaryotic signaling pathways regulating diverse cellular processes, including vesicle transport, embryonic development, the sensation of vision, odor, taste and pain, microtubule assembly and cell division. These GTPases are associated both with guanine nucleotide exchange factors (GEFs), which turn them on by mediating the exchange of GTP for GDP, and with GTPase activating proteins (GAPs), which turn them off by stimulating hydrolysis of GTP to GDP. The on- or off-state of Ras-like GTPases is communicated through conformational changes within their switch I and II regions, which detect the presence or absence of the γ phosphate of bound guanine nucleotide. Fig. S2 contains four contrast alignments that were generated by the mcBPPS sampler for Rab GTPases (the subgroup specified in row 4 of the hyperpartition in Table 1of the paper). Here I consider the biological significance of the residues identified in these alignments. See Table S1 for the functional/structural relevance of pattern residues. To obtain the actual output for this and the other examples, readers of this article can obtain the mcBPPS program, the input data, and a script for running the program on this data at the mcBPPS website (http://www.chain.umaryland.edu/mcbpps/).

P-loop GTPases vs other proteins. The contrast alignment in Fig. S2A identifies residues that are characteristic of all P-loop GTPases. These residues bind to GDP or GTP and show up at the sequence level as several highly conserved motifs (3), including: the Walker A (G-K-[ST]) motif, which corresponds to the P-loop; the Walker B (D-x-x-G) motif, the conserved aspartate (D) of which interacts (indirectly through a water molecule) with the Mg++ ion that coordinates with nucleotide phosphate groups; and an [NT]K.D motif, the residues of which bind to guanine.

TRAFAC subclass of P-loop GTPases. The contrast alignment in Fig. S2B highlights the same motifs generally conserved within all GTPases. However, unlike the other major subclass (SIMIBI GTPases) the TRAFAC (transcription factor-related) GTPases also conserve a threonine within the switch I region and a serine near the C-terminal end of the GTPase domain. The threonine coordinates with a Mg++ ion that coordinates with bound guanine nucleotide whereas the serine forms a hydrogen bond to the aspartate residues of the NK.D (guanine binding loop) motif.

Ras-like GTPases. The contrast alignment in Fig. S2C identifies co-conserved residues characteristic of Ras-like GTPases. In addition to residues known to perform critical functions (see below), these include three additional patterns: (i) the pattern [RK]-x-[ILV] preceding the P-loop, (ii) the pattern [WF] directly preceding the Walker B aspartate, and (iii) the pattern [YF]-[YF] at the C-terminal end of the switch II region. Even though these patterns occur within three distinct regions in the sequences, within available crystal structures the corresponding residues cluster near the switch II region's C-terminal (swII-CT) end and exhibit diverse structural configurations, which appear to be facilitated by alternative interactions involving the aromatic residues of the [WF] and [YF]-[YF] patterns. (Such aromatic interactions play a central role in molecular recognition and self-assembly (4, 5) and can accommodate diverse conformational forms by interacting in four distinct geometric orientations (6).) One configuration can explain why the swII-CT residues are conserved inasmuch as it is characterized by atomic interactions (Fig. S2D) that form a pocket around the negative-dipole moment of the switch II α helix with the positively-charged residue of the [RK]-x-[ILV] pattern inserted into this pocket. A connection between this **charge-dipole pocket** and switch II restructuring is suggested by comparisons between typical monomeric forms of Rab family GTPases and *an unusual, homodimeric form*, in which the region connecting switch I to switch II is exchanged between subunits so that, for each subunit, this region is donated by the other subunit (Fig. S3A,B). As a result, the switch II region forms a long α helix that is directed away from the structural core of one subunit and toward the structural core of the other subunit. For each subunit the switch II helix presumably lacks the conformational strain typically imposed on monomeric forms (for which the switch II region needs to bend around and reconnect to the structural core) and the swII-CT residues form the charge-dipole pocket configuration with ideal geometry. The charge-dipole pocket thus appears structurally compatible with formation of this unusual, outward-directed switch II helix—a structural theme that is observed repeatedly in other Ras-like GTPases. Indeed, among over 200 Ras-like GTPase structures, those that exhibit the charge-dipole pocket or nearly so consistently (e.g., see Fig. S3C) possess a switch II helix that traces out a path close to that of the homodimeric helix, but those Ras-like GTPases lacking the charge-dipole pocket (and all non-Ras-like GTPases of known structure) possess a switch II helix that traces out a different path (Fig. S3E).

The Ras-like GTPase analysis yields the following observations: **(i)** The most distinguishing feature of Ras-like GTPases at the sequence level are the five conserved residue of previously unknown function—along with five other conserved residue positions whose likely biological roles in Ras-like GTPases were noted previously. **(ii)** At the structural level, these newly identified residues come together to form either a 'charge-dipole pocket' configuration or a variety of alternative configurations. **(iii)** The specific atomic interactions forming the charge-dipole pocket configuration require the specific amino acids conserved at these positions. **(iv)** The charge-dipole pocket configuration occurs in both the GDP- and the GTP-bound states and both the charge-dipole pocket and alternative configurations can occur in the same state. **(v)** The formation of the charge-dipole pocket is highly correlated with formation within the switch II region of an unusual, outward-oriented α-helix. And **(vi)** the formation of this helix is associated with structural changes in the switch II N-terminal region, which senses the γ phosphate of GTP and which harbors three previously-noted Ras-like GTPase residues corresponding to the pattern [AG]-x-Q-[DE] (see Fig. S2C) and that are involved in GTP hydrolysis (7) and nucleotide exchange (8-10). Together, these observations suggest that—by promoting formation of the outward-directed switch II helix and, as a result, repositioning of co-conserved residues that are implicated in GTP hydrolysis and nucleotide exchange—the charge-dipole pocket facilitates bidirectional switching between on/off states.

Rho, Rab and Ran GTPases. The contrast alignment in Fig. S2D highlights six residues that most distinguish Rab, Rho/Rac, and Ran (and to some degree Ras) GTPases from other Ras-like GTPases. Four of these correspond to a

structural configuration that forms a 'glycine brace' (11) (Fig. S4): One of the aromatic residues forms a stabilizing CH-π interaction with a conserved glycine at the start of the guanine-binding loop whereas a second aromatic residue, which is nearly always a tryptophan, likewise forms stabilizing CH-π and NH-π interactions with a glycine at the start of the phosphate-binding P-loop. The two other residues (typically an aspartate and a serine or threonine), together with a conserved buried water molecule, form a network of interactions connecting the two aromatic residues. This glycine brace may influence guanine nucleotide binding or release by stabilizing two glycine hinge residues at the ends of the β-strands directly preceding the P-loop and the guanine binding loop. The presence of glycine intrinsically destabilizes β sheets, but this sort of aromatic-glycine interaction has been proposed to counteract this effect (12). In this case, this seems likely to stabilize the two nucleotide-binding loops. Moreover, co-conservation of the four glycine brace residues with a threonine and alanine (the T-A pattern in Fig. S2D) is consistent with a role in nucleotide exchange inasmuch as repositioning of this alanine residue is believed to facilitate nucleotide exchange by occluding the Mg++ binding site, leading to expulsion of the phosphate-associated Mg++ ion (7).

Analysis of Eukaryotic DNA Clamp Loaders. The eukaryotic DNA clamp loader complex, termed Replication Factor C (RFC), forms a stable complex with a DNA clamp protein in the presence of ATP. Upon recognition of a 3'-recessed single-stranded/double-stranded junction (the start of an Okazaki fragment) the RFC/ATP/clamp complex undergoes ATP hydrolysis resulting in dissociation of RFC and loading of the clamp onto DNA (Fig. S5A). The eukaryotic RFC complex is composed of 5 evolutionarily-related AAA+ subunits (13)—denoted A, B, C, D, and E—that form a semi-circular arrangement (Fig. S5B). In archaea, 2 related AAA+ subunits, denoted RFC-L and -S, form a similar arrangement (not shown) where RFC-L corresponds to RFC-A and 4 copies of RFC-S correspond to RFC-B to -E. In eubacteria, 3 subunits, denoted δ, γ, and δ', form a similar arrangement (see Supplement 4) where δ corresponds to RFC-A, 3 copies of γ correspond to RFC-B to -D and δ' corresponds to RFC-E. The RFC-A,-B,-C,-D, -L, -S and γ subunits are active ATPases, whereas RFC-E, δ and δ' are inactive. The mcBPPS-generated contrast alignments (see Fig. S6) identify six categories of constraints imposed on RFC subunits (14, 15). When considering the nature of these constraints, it is helpful to first consider their biochemical relevance:

AAA+ ATPases vs all other proteins. The contrast alignment in Fig. S6A identifies residues that are characteristic of all AAA+ ATPases and that correspond to the Walker A and B motifs, which are involved in ATP hydrolysis, and to a conserved arginine finger, which activates ATP-hydrolysis in trans. The functions of these residues are well known.

RFC-ABCD vs RFC-E. The contrast alignment in Fig. S6D identifies four residues that are both characteristic of active RFC ATPases (subunits A-D) and uncharacteristic of catalytically-impaired RFC-E subunits. These residues are likely to play key roles associated with ATP hydrolysis and, as suggested by this analysis, with coupling of ATP hydrolysis to the recognition of cognate DNA. Within the RFC–ATP–clamp complex (16) (Fig. S5C) one of these residues (e.g., Arg84 in RFC-B) forms a hydrogen bond with the main chain oxygen of a co-conserved residue, the Walker B aspartate (Asp114 in RFC-B), which plays a key catalytic role by coordinating with the ATP-bound Mg++ ion. In other P loop NTPases, this main chain oxygen normally hydrogen bonds to an adjacent β-strand, but its interaction with this arginine (termed here the Walker B-arginine) disrupts this standard geometry and thus may inhibit catalysis. This arginine also forms a hydrogen bond with the main chain oxygen of a glutamate (e.g., Glu115 in RFC-B; Fig. S5D) that serves as the catalytic base facilitating nucleophilic attack by a water molecule on the γ-phosphate of ATP (17) and that is a characteristic feature of all active AAA+ ATPases (Fig. S6A). Thus, by forming main chain hydrogen bonds on either side of this glutamate, the Walker B-arginine may influence the conformations of two key catalytic residues and, as a result, catalytic activity. In the structure of the corresponding archaeal subunit (RFC-S) bound to ADP (18), however, this (positively charged) arginine side-chain is dramatically repositioned to where it could interact with (negatively charge) DNA and thereby also allow the Walker B region to hydrogen bond in the usual way with the adjacent β-strand (Fig. S5E). This suggests a mechanism to couple recognition of DNA with ATP hydrolysis. Also distinctive of catalytically active RFC subunits (Fig. S6D) is an alanine residue (e.g., Ala80 in RFC-B in Fig. S5D) which packs against the arginine and that may assist in this putative mechanism by facilitating conformational flexibility. (A larger side-chain at this position could sterically hinder a dramatic conformational change involving the arginine.) Moreover, upon movement of the Walker B-arginine, the Walker B aspartate's main chain oxygen hydrogen bonds with the alanine's main chain nitrogen (Ala84 in RFCS in Fig. S5E)—which is the standard configuration for P loop NTPases. Together this suggests a mechanism where, upon association of the RFC–ATP–clamp complex with DNA, the positively charged Walker B-arginine moves into contact with negatively charged DNA thereby perhaps facilitating ATP hydrolysis. Finally, also showing up in the mcBPPS analysis is a threonine residue that directly follows the Walker A catalytic lysine (both residues not shown in Fig. S5); this

threonine forms a hydrogen bond with the co-conserved Walker B aspartate; thus together these two residues appear to play an important role in positioning the ATP substrate for catalysis.

RFC vs bacterial γ subunits. The contrast alignment in Fig. S6C identifies those co-conserved residues that most distinguish RFC subunits from corresponding bacterial clamp loader γ subunits, which lack the Walker B-arginine and thus presumably any associated mechanisms. This reveals an NxSD motif surrounding the above mentioned alanine residue (e.g., Ala80 in Fig. S5D) and two other residues near the active site (e.g., Asp117 and Asn145 in RFC-B). The NxSD motif is located two positions before the Walker B-arginine and corresponds both to the β-strand that hydrogen bonds to the Walker B region and to the loop following this strand. In the structure of the archaeal small subunit (RFC-S) co-crystallized with ADP (18) (Fig. S5E), one of the Walker B main chain oxygens hydrogen bonds to the main chain -NH directly following the NxSD-asparagine, but in the eukaryotic RFC /ATP/clamp complex this oxygen hydrogen bonds to the Walker B-arginine instead. Furthermore, in the eukaryotic complex, the NxSD-asparagine (N79B) establishes characteristic interactions with the NxSD-serine and NxSD-aspartate, as well as with other nearby residues (Figs S5D), while in the archaeal RFCS-ADP structure these interactions are rearranged (Fig. S5E). Taken together these observations suggest a possible role for these residues in assisting the Walker B-arginine conformational switch and perhaps in helping reposition the catalytically critical Walker B region. Another co-conserved residue showing up in this analysis (e.g., Asn145 in RFC-B) is positioned to directly interact with the γ-phosphate group of ATP and, in turn, forms a backbone hydrogen bond with another co-conserved residue, an aspartate (e.g., Asp117 in RFC-B) that is two residue positions beyond the Walker B D-E motif (structures not shown); thus, within eukaryotes these two residues form a specific link between the Walker B region and the γ-phosophate of ATP that is absent from bacterial clamp loaders.

RFC-BCDE vs RFC-A. The contrast alignment in Fig. S6E identifies co-conserved residues within subunits that interact with the ATP-binding site of an adjacent AAA+ subunit. Within the RFC-B subunit, these trans-acting residues include Gln125, Arg128, Arg129, Arg157 (see Fig. S5F) and Glu132 (not shown)— which together form a network of hydrogen bonds that locks down the relative orientation of domain II of the adjacent RFC subunit with respect to ATP (16). A co-conserved glutamine (e.g., Gln124 of RFC-B in Fig. S5F) forms hydrogen bond interactions between backbone regions harboring these trans-interacting residues and thus may assist these interactions.

RFC-BCD vs RFC-AE. The large subunit, RFC-A, which is believed to recognize primed DNA as a signal for DNA-dependent ATP hydrolysis (19), is likely to be the first subunit undergoing ATP hydrolysis and thus may initiate ATP hydrolysis in the other subunits. Thus constraints associated with the propagation of ATP hydrolysis to adjacent subunits are likely to be imposed upon active ATPases that directly interact with an adjacent ATP site (that is closer to the RFC-A subunit)—that is upon the RFC-BCD subunits (Fig. S6F). A cluster of co-conserved residue in this category are located within the α4-helix (Ile86, Arg90 and Phe96 of RFC-B in Fig. S5C) and, notably, on either side of the Walker B-interacting arginine, within a loop that is located just before this helix (Asp83 and Gly85 of RFC-B in Figs S5C-E). The α4-arginine electrostatically interacts with main-chain oxygens of the adjacent subunit's NxSD motif (Fig. S5C) and with the co-conserved acidic residue directly preceding the adjacent subunit's Walker B-arginine (e.g., $Asp83^{RFC\text{-}B}$). Therefore an ATP hydrolysis-associated conformational change in the preceding subunit could perturb this α4-arginine, which might itself play an active role in this switch, as it is predicted to interact with DNA (20). In addition to contacting the adjacent subunit via this arginine, the N-terminal end of the α4 helix is directly connected to the Walker B-arginine of its own subunit (e.g., Arg84 of RFC-B in Fig. S5C) via two other co-conserved residues (e.g., Ile86 and Gly85 of RFC-B in Figs S5C and S6F). Notably, the proposed conformational switch both repositions the isoleucine into the pocket vacated by the Walker B-arginine and repositions the (negatively charged) acidic residue away from the central hole through which (negatively charged) DNA is thread. (Compare $Ile86^{RFC\text{-}B}$ and $Asp83^{RFC\text{-}B}$ in Fig. S5D with $Ile90^{RFCS}$ and $Glu87^{RFCS}$ in Fig. S5E.) Together these observations suggest a possible role for these residues in propagating the Walker B-arginine conformational switch to adjacent subunits.

All clamp loaders adjacent to an ATP-binding site vs other AAA+ subunits. Finally, the contrast alignment in Fig. S6B reveals those residues co-conserved within both RFC and bacterial clamp loader subunits that are adjacent to an ATP-binding site but not within non-clamp loader AAA+ subunits. By far the residue most strikingly conserved in this category is a lysine (e.g., $Lys109^{RFC\text{-}B}$ in Fig. S5C) located within the β-strand preceding the Walker B motif. This lysine electrostatically interacts with and/or hydrogen bonds to backbone oxygens at the C-terminal end of the α4 helix, which is a region that binds directly to the clamp. This lysine thus seems likely to play a critical role in clamp loading for both eukaryotes and bacteria and, given the preceding observations, might play a role in coupling ATP hydrolysis to release of the clamp.

RFC-BCD summary. Taken together, this mcBPPS analysis of RFC-B, -C and -D subunits indicates that these subunits share functionally critical features with RFC-A alone, with RFC-E alone, with both RFC-A and RFC-E, and with neither RFC-A nor RFC-B. Moreover, all RFC subunits conserve features both distinct from and in common with bacterial γ and δ' subunits. Such sequence similarities and differences reflect functional similarities and differences that, presumably, are associated with each subunit's specific role within their associated clamp loader complex. Such relationships between divergent subgroups are sometimes difficult to represent in a phylogenetic tree, which averages out the subtle distinctions associated with isolated sequence patterns. For example, RFC-B,-C and -D are, in some respects, more similar to RFC-A than to RFC-E (Fig. S6D), yet, in other respects, more similar to RFC-E than to RFC-A (Fig. S6E). The mc-BPPS analysis of these clamp loaders thus suggest plausible hypotheses that, though by no means proving any associated mechanisms, are nevertheless useful for experimental design. (See Table S2 for a list of structurally and functionally relevant features of pattern residues.)

Analysis of Bacterial DNA Clamp Loaders. The bacterial DNA clamp loader complex, like the eukaryotic RFC complex, forms a stable complex with a DNA clamp protein in the presence of ATP and upon recognition of RNA-primed DNA undergoes ATP hydrolysis resulting in dissociation of the complex and loading of the clamp (Fig. S7A). The bacterial complex consists of 3 evolutionarily-related subunits, denoted δ, γ, and δ', that form an arrangement analogous to the eukaryotic complex where δ (which is very poorly conserved) corresponds to RFC-A, 3 copies of γ correspond to RFC-B to -D and δ' corresponds to RFC-E. Only the γ subunit is an active ATPases; δ and δ', though not functional ATPases, also belong to the AAA+ class (13, 21). Members of the AAA+ class are characterized by an N-terminal domain (domain I) (shown in Fig. S7B) that, for active ATPases, conserves Walker A and B ATP-binding motifs (22), followed C-terminally by a helical bundle domain (domain II)(not shown). A third, C-terminal domain (domain III) (not shown) that is associated with all three types of subunits forms a circular collar from which the AAA+ modules hang (21). These are arranged around the circle in the order δ'-γ_B-γ_C-γ_D-δ, which is important for a proposed sequential mechanism (23) for loading of the β clamp onto DNA. The mcBPPS sampler generated contrast alignments (Fig. S8) identify four categories of functionally-divergent residues within γ subunits (14, 15). Here I provide structural and functional interpretations for these co-conserved residues.

AAA+ vs other proteins and clamp loaders vs other AAA+. The contrast alignments in Figs. S8A and S8B identify distinguishing features of all (active) AAA+ ATPases and of all (bacterial, archaeal and eukaryotic) clamp loader AAA+ subunits that are adjacent to an ATP-binding site. These features correspond to those shown for the analogous RFC-BCD subunits in Figs S6A and S6B; see above for a discussion of these pattern residues.

γ and δ' vs RFC-BCDES. The contrast alignment in Fig. S8C compares bacterial clamp loader subunits that interact with an adjacent active ATP binding site versus the corresponding subunits from eukaryotic and archaeal organisms. This highlights five co-conserved residues that (within γ and δ' subunits) interact with the ATP-binding site of an adjacent AAA+ subunit. One of these is a threonine (Thr165-γ in *E. coli*) that, based on homology modeling, is positioned to interact in trans with the adjacent subunit so as to facilitate ATP hydrolysis (Fig. S7C). Fig. S7C represents the (hypothetical) region of interaction between the γ_B and γ_C subunits captured at the point in which the putative catalytic base (Glu127, a key catalytic residue in active AAA+ ATPases) is positioned to extract a proton from and thereby activate a water molecule for nucleophilic attack upon the γ-phosphate of ATP. Thr165 is positioned to participate in this process (see the proposed mechanism for ATP hydrolysis in Fig. S7D).

The roles of two other residues showing up in Fig. S7E—namely a conserved lysine and a conserved glutamate (Lys141 and Glu145 in *E. coli*)—are suggested by a recent structure of the *E. coli* clamp loader complex co-crystallized with the ATP analog ADP•BeF_3 and primer-template DNA (24) (Fig. S7E). This complex presumably represents a step prior to nucleophilic attack by the catalytic base (there is no water molecule visible in the structure). Nevertheless, this structure suggests that the conserved lysine and glutamate residues help correctly position the interface between γ_B and γ_C : Lys141- γ_C forms a hydrogen bond with the Walker B aspartate (Asp126- γ_B , which coordinates with the ATP-associated Mg++ ion and is a key catalytic residue in active AAA+ ATPases). Glu145- γ_C forms a salt bridge and hydrogen bonds with a conserved arginine (Arg56-γ_B) that is located in the helix associated with the Walker A motif. (This arginine or a lysine, which are both basic residues, are conserved across various subgroups of AAA+ subunits that were not categorized by this analysis.) Finally, the remaining two residues showing up in Fig. S4-3C, namely a conserved Leucine (Leu140- γ_C) and a conserved proline (Pro146- γ_C) that are adjacent to the lysine and glutamate residues, respectively, may stabilize the interface interactions mediated by the latter two residues.

γ vs δ'. The contrast alignment in Fig. S8D (column 17 in Table 2 of the paper) compares bacterial γ subunits, which are active ATPases, with bacterial δ' subunits, which are not. The most striking residues showing up in this analysis

are the eight highlighted in this figure. Mutation of one of these, a threonine (Thr157- γ_B in Fig. S7E) prevents ATP hydrolysis (25). Another threonine (Thr52) corresponds to the Walker A motif. Several other of these pattern residues appear to be indirectly associated with catalysis. Notably these all occur very near to the active site. Two of these, Glu92 and Asp94, contact other residues implicated either directly or indirectly in ATP hydrolysis (Fig. 7E). Thus this analysis suggests that these residues play roles associated with steps in ATP-hydrolysis (e.g., the transition state) and in loading of the clamp, for which corresponding crystal structures have not yet been determined. (See Table S2 for a list of structurally and functionally relevant features of pattern residues.)

Analysis of Superfamily 1 and 2 Helicases. Helicases are involved in various aspects of nucleic acid metabolism including DNA replication, repair, recombination, transcription, as well as ribosome biogenesis and RNA processing, translation, and decay (26). The mcBPPS sampler was applied to an input alignment containing 40,366 Helicase sequences (after removing fragments and redundant sequences) using the hyperpartition shown in Table S3. Shown in Fig. S10 below are two output contrast alignments obtained for the subgroup in row 3 of Table S3; the seed alignment consisted of a dozen eukaryotic initiation factor 4AIII (eIF4AIII) helicases, which are members of the DEAD-box family belonging to the so-called superfamily 2 (SF2) helicases. Shown in Fig. S9 below are three output contrast alignments obtained for the subgroup in row 1 of Table S3; the seed alignment consisted of ten PcrA DNA helicases, which are members of the so-called superfamily 1 (SF1) helicases. In both cases, most of the co-conserved residues that the sampler identifies either are associated with regions that bind to ATP or to an RNA or DNA substrate or that allosterically link these two regions. This is evident, for example, from the crystal structures both of the human exon junction complex containing trapped eIF4AIII bound to an ATP analog and RNA (27)(Fig. S11A) and of the PcrA DNA helicase bound to a DNA substrate (28) (Fig. S11B). For a synopsis of the functional and structural roles of pattern residues see Table S4.

Analysis of protein kinases. Eukaryotic protein kinases (EPKs) regulate signaling pathways by phosphorylating various proteins within those pathways. EPKs can be broadly classified into six major groups (AGC, CMGC, CAMK, TK, STE and CK1) based on sequence similarities within the kinase domain (29). EPK's are distantly related to eukaryotic-like kinases (ELK's) and to atypical protein kinases (APK's), both of which are present in both eukaryotes and prokaryotes (30). Although EPK's specifically phosphorylate protein substrates, ELK's and APK's are known to phosphorylate both protein and small molecule substrates (31, 32). The mcBPPS sampler was applied to an input alignment of 22,424 sequences after removing fragments, redundant sequences and EPK sequences with > 70% sequence identity; as a result, this alignment represents sequences from the EPK, ELK and APK subgroups in roughly equal proportions. The hyperpartition shown in Table S5 was used. Fig S13 gives the output contrast alignments corresponding to three categories of residues co-conserved: (i) in all protein kinases (Fig S13A), (ii) in EPK but not ELKs or APKs (Fig S13B), and (iii) in AGC kinases, but not other EPKs (Fig S13C). Residues conserved within all protein kinases (green side-chains in Fig S12A) cluster around the ATP binding site and contribute to ATP binding and phosphoryl transfer (33). Residues shared by EPKs but not by ELKs or APKs (magenta side-chains in Fig S12A) either interact with the activation loop (which contributes to protein substrate specificity and to tight regulation of EPK activity) or with the peptide substrate or are associated with active site conserved residues. AGC kinases, a major subgroup within the EPKs, co-conserved residues (yellow side-chains in Fig. S12B) that interact with the AGC C-terminal tail (the C-Tail), which allosterically couples regulatory and catalytic functions of the kinase core (34). For a synopsis of the functional and structural roles of pattern residues see Table S6.

The Sampler's Performance. ***Convergence***. Fig. S14 shows the typical convergence behavior for representative mc-BPPS runs (with P-loop GTPases). For a biologically valid input sequence alignment, the mc-LPR is negative when the sampler is arbitrarily initialized, but turns positive after converge on significant pattern-partition pairs. However, the mc-BPPS heuristic used to initialize the subgroup sequence assignments (see Methods) may cause the mc-LPR to start out positive. In any case, sequence sampling is performed until the mc-LPR turns positive, at which point column sampling is initiated, resulting in a spike in the mc-LPR (as seen in Fig. S14). As the slope of the mc-LPR curve flattens out, the sampling temperature decreases to zero, at which point sampling is replaced by a hill-climbing strategy until convergence (defined as one full cycle without increasing the mc-LPR).
Robustness. The robustness of the mcBPPS sampler was tested both by randomly removing 50% of the sequences from the main alignment and by using an independently-generated input alignment. The results are shown in Fig. S15 and consist of three pages, each of which shows five versions of one of the three contrast alignments in Fig. 3 in the paper: The first version is identical to the corresponding alignment shown in Fig. 3, the next three versions

correspond to analyses where 50% of the input sequences were randomly removed, and the last version corresponds to an input alignment (of 41,978 sequences) that was obtained using PSI-BLAST (35) initialized with the NCBI-curated CDtree alignment for P-loop GTPases (36). Notably, the PSI-BLAST-based analysis misses canonical threonine and glycine residues within the switch I region (which is due to misalignment of this region), and a canonical tryptophan (which PSI-BLAST misaligned due to an insertion before and a deletion after this residue). Nevertheless, the otherwise modest variability between these analyses demonstrates that, given an accurate input alignment, the results are reproducible. It should also be noted in this context that when a biologically unrealistic category is defined, the LPR for that category is less than zero—thereby indicating that that aspect of the hyperpartition specification is not supported by the data. No output is generated for such categories.

Tuning parameters in prior distributions. There are two user-modifiable parameters for each category: (i) the α prior, which influence the overall stringency with which the foreground pattern positions are conserved or—viewed from another perspective—the amount of contamination expected at pattern positions; and (ii) the ***S*** priors, which influence the stringency with which a sequence needs to match the associated pattern(s) in order to be included in that subgroup. Choosing these parameters is typically based on whether the foreground sequences are expected to be highly diverse or more closely related to each other. As a result, these are typically set to be more stringent for smaller, more closely-related protein subgroups (e.g., families or subfamilies) and less stringent for larger, more diverse supergroups (e.g., protein superfamilies or subclasses). Sometimes, however, additional biological information is available that influences the choice of these parameters. For example, bacterial δ' DNA clamp loader subunits, even though they constitute a distinct protein subfamily, lack ATPase activity and are poorly conserved, which thus call for less stringent parameter settings. More stringent parameter settings also favor assignment of poorly-conserved (putative pseudogene products and other aberrant) sequences to the random sequence set, which, as a result, favors better definition of the canonical patterns.

Speed. The mc-BPPS sampler took 83 minutes to run given the hyperpartition in Table 1 and 66,386 (non-redundant) P-loop GTPase sequences as input. (Representative output alignments are shown in Fig. 3.) The mc-BPPS sampler took 48 minutes to run given the hyperpartition in Table 2 and 43,298 (non-redundant) AAA+ sequences as input. (Representative output alignments are shown in Fig. 5.)

Automatically generated hyperpartition. An example of the hyperpartition and contrast alignments obtained when running the fully automated version of the mc-BPPS sampler are show below in Table S7 and Figure S16, respectively. Due to its length, Figure S16 is appended to the very end of this file (after the tables).

References:

1. Neuwald AF (2009) Rapid detection, classification and accurate alignment of up to a million or more related protein sequences *Bioinformatics* 25(15):1869-1875.
2. Neuwald AF (2007) The CHAIN program: forging evolutionary links to underlying mechanisms. *Trends Biochem Sciences* 32(00):487-493.
3. Wittinghofer A (2000) The Functioning of Molecular Switches in Three Dimensions. *GTPases*, ed Hall A (Oxford University Press, Oxford), pp 244-310.
4. Claessens CG & Stoddart JF (1998) pi-pi interactions in self-assembly. *J. Physical Organic Chem.* 10(5):254-272.
5. McGaughey GB, Gagne M, & Rappe AK (1998) pi-Stacking interactions. Alive and well in proteins. *J Biol Chem* 273(25):15458-15463.
6. Sun S & Bernstein ER (1996) Aromatic van der Waals clusters: structure and nonrigidity. *J. Phys. Chem.* 100:13348-13366.
7. Vetter IR & Wittinghofer A (2001) The guanine nucleotide-binding switch in three dimensions. *Science* 294(5545):1299-1304.
8. Gasper R, Thomas C, Ahmadian MR, & Wittinghofer A (2008) The role of the conserved switch II glutamate in guanine nucleotide exchange factor-mediated nucleotide exchange of GTP-binding proteins. *J Mol Biol* 379(1):51-63.
9. Margarit SM*, et al.* (2003) Structural evidence for feedback activation by Ras.GTP of the Ras-specific nucleotide exchange factor SOS. *Cell* 112(5):685-695.
10. Thomas C, Fricke I, Scrima A, Berken A, & Wittinghofer A (2007) Structural evidence for a common intermediate in small G protein-GEF reactions. *Mol Cell* 25(1):141-149.
11. Neuwald AF (2009) The glycine brace: a component of Rab, Rho, and Ran GTPases associated with hinge regions of guanine- and phosphate-binding loops. *BMC Struct Biol* 9:11.

12. Merkel JS & Regan L (1998) Aromatic rescue of glycine in beta sheets. *Fold Des* 3(6):449-455.
13. Neuwald AF, Aravind L, Spouge JL, & Koonin EV (1999) AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. *Genome Res* 9(1):27-43.
14. Neuwald AF (2006) Bayesian shadows of molecular mechanisms cast in the light of evolution. *Trends Biochem Sciences* 31(7):374-382.
15. Neuwald AF (2005) Evolutionary clues to eukaryotic DNA clamp-loading mechanisms: analysis of the functional constraints imposed on replication factor C AAA+ ATPases. *Nucleic Acids Res* 33(11):3614-3628.
16. Bowman GD, O'Donnell M, & Kuriyan J (2004) Structural analysis of a eukaryotic sliding DNA clamp-clamp loader complex. *Nature* 429(6993):724-730.
17. Orelle C, Dalmas O, Gros P, Di Pietro A, & Jault JM (2003) The conserved glutamate residue adjacent to the Walker-B motif is the catalytic base for ATP hydrolysis in the ATP-binding cassette transporter BmrA. *J Biol Chem.* 278(47):47002-47008. Epub 42003 Sep 47010.
18. Oyama T, Ishino Y, Cann IK, Ishino S, & Morikawa K (2001) Atomic structure of the clamp loader small subunit from Pyrococcus furiosus. *Mol Cell* 8(2):455-463.
19. Ellison V & Stillman B (2003) Biochemical characterization of DNA damage checkpoint complexes: clamp loader and clamp complexes with specificity for 5' recessed DNA. *PLoS Biol* 1(2):E33. Epub 2003 Nov 2017.
20. Goedken ER, Kazmirski SL, Bowman GD, O'Donnell M, & Kuriyan J (2005) Mapping the interaction of DNA with the Escherichia coli DNA polymerase clamp loader complex. *Nat Struct Mol Biol* 12(2):183-190. Epub 2005 Jan 2016.
21. Jeruzalmi D, O'Donnell M, & Kuriyan J (2001) Crystal structure of the processivity clamp loader gamma (gamma) complex of E. coli DNA polymerase III. *Cell* 106(4):429-441.
22. Walker JE, Saraste M, Runswick MJ, & Gay NJ (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *Embo J* 1(8):945-951.
23. Johnson A & O'Donnell M (2003) Ordered ATP hydrolysis in the gamma complex clamp loader AAA+ machine. *J Biol Chem* 278(16):14406-14413.
24. Simonetta KR*, et al.* (2009) The mechanism of ATP-dependent primer-template recognition by a clamp loader complex. *Cell* 137(4):659-671.
25. Hattendorf DA & Lindquist SL (2002) Cooperative kinetics of both Hsp104 ATPase domains and interdomain communication revealed by AAA sensor-1 mutants. *Embo J.* 21(1-2):12-21.
26. Abdelhaleem M (2010) Helicases: an overview. *Methods Mol Biol* 587:1-12.
27. Andersen CB*, et al.* (2006) Structure of the exon junction core complex with a trapped DEAD-box ATPase bound to RNA. *Science* 313(5795):1968-1972.
28. Velankar SS, Soultanas P, Dillingham MS, Subramanya HS, & Wigley DB (1999) Crystal structures of complexes of PcrA DNA helicase with a DNA substrate indicate an inchworm mechanism. *Cell* 97(1):75-84.
29. Manning G, Whyte DB, Martinez R, Hunter T, & Sudarsanam S (2002) The protein kinase complement of the human genome. *Science* 298(5600):1912-1934.
30. Kannan N, Taylor SS, Zhai Y, Venter JC, & Manning G (2007) Structural and functional diversity of the microbial kinome. *PLoS Biol* 5(3):e17.
31. Peisach D, Gee P, Kent C, & Xu Z (2003) The crystal structure of choline kinase reveals a eukaryotic protein kinase fold. *Structure* 11(6):703-713.
32. Steinbacher S*, et al.* (1999) The crystal structure of the Physarum polycephalum actin-fragmin kinase: an atypical protein kinase with a specialized substrate-binding domain. *EMBO J* 18(11):2923-2929.
33. Knighton DR*, et al.* (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. *Science* 253(5018):407-414.
34. Kannan N, Haste N, Taylor SS, & Neuwald AF (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc Natl Acad Sci U S A* 104(4):1272-1277.
35. Altschul SF*, et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.

36. Marchler-Bauer A*, et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* 37(Database issue):D205-210.
37. Neuwald AF (2009) The charge-dipole pocket: a defining feature of signaling pathway GTPase on/off switches. *J Mol Biol* 390(1):142-153.
38. Chavas LM*, et al.* (2007) Structure of the small GTPase Rab27b shows an unexpected swapped dimer. *Acta Crystallogr D Biol Crystallogr* 63(Pt 7):769-779.
39. Pasqualato S*, et al.* (2004) The structural GDP/GTP cycle of Rab11 reveals a novel interface involved in the dynamics of recycling endosomes. *J Biol Chem* 279(12):11480-11488.
40. Tarricone C*, et al.* (2001) The structural basis of Arfaptin-mediated cross-talk between Rac and Arf signalling pathways. *Nature* 411(6834):215-219.
41. Word JM, Lovell SC, Richardson JS, & Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol* 285(4):1735-1747.
42. Anonymous (Based on many published crystal structures.
43. Snyder AK, Williams CR, Johnson A, O'Donnell M, & Bloom LB (2004) Mechanism of loading the Escherichia coli DNA polymerase III sliding clamp: II. Uncoupling the beta and DNA binding activities of the gamma complex. *J Biol Chem.* 279(6):4386-4393.
44. Neuwald AF (2006) Hypothesis: bacterial clamp loader ATPase activation through DNA-dependent repositioning of the catalytic base and of a trans-acting catalytic threonine. *Nucleic Acids Res* 34(18):5280-5290.
45. Soultanas P, Dillingham MS, Velankar SS, & Wigley DB (1999) DNA binding mediates conformational changes and metal ion coordination in the active site of PcrA helicase. *J Mol Biol* 290(1):137-148.
46. Le Hir H & Andersen GR (2008) Structural insights into the exon junction complex. *Curr Opin Struct Biol* 18(1):112-119.
47. Aimes RT, Hemmer W, & Taylor SS (2000) Serine-53 at the tip of the glycine-rich loop of cAMP-dependent protein kinase: role in catalysis, P-site specificity, and interaction with inhibitors. *Biochemistry* 39(28):8325-8332.
48. Johnson DA, Akamine P, Radzio-Andzelm E, Madhusudan M, & Taylor SS (2001) Dynamics of cAMP-dependent protein kinase. *Chem Rev* 101(8):2243-2270.
49. Huse M & Kuriyan J (2002) The conformational plasticity of protein kinases. *Cell* 109(3):275-282.
50. Gould CM, Kannan N, Taylor SS, & Newton AC (2009) The chaperones Hsp90 and Cdc37 mediate the maturation and stabilization of protein kinase C through a conserved PXXP motif in the C-terminal tail. *J Biol Chem* 284(8):4921-4935.
51. Kannan N & Neuwald AF (2005) Did protein kinase regulatory mechanisms evolve through elaboration of a simple structural component? *J Mol Biol* 351(5):956-972.
52. Akamine P*, et al.* (2003) Dynamic features of cAMP-dependent protein kinase revealed by apoenzyme crystal structure. *J Mol Biol* 327(1):159-171.

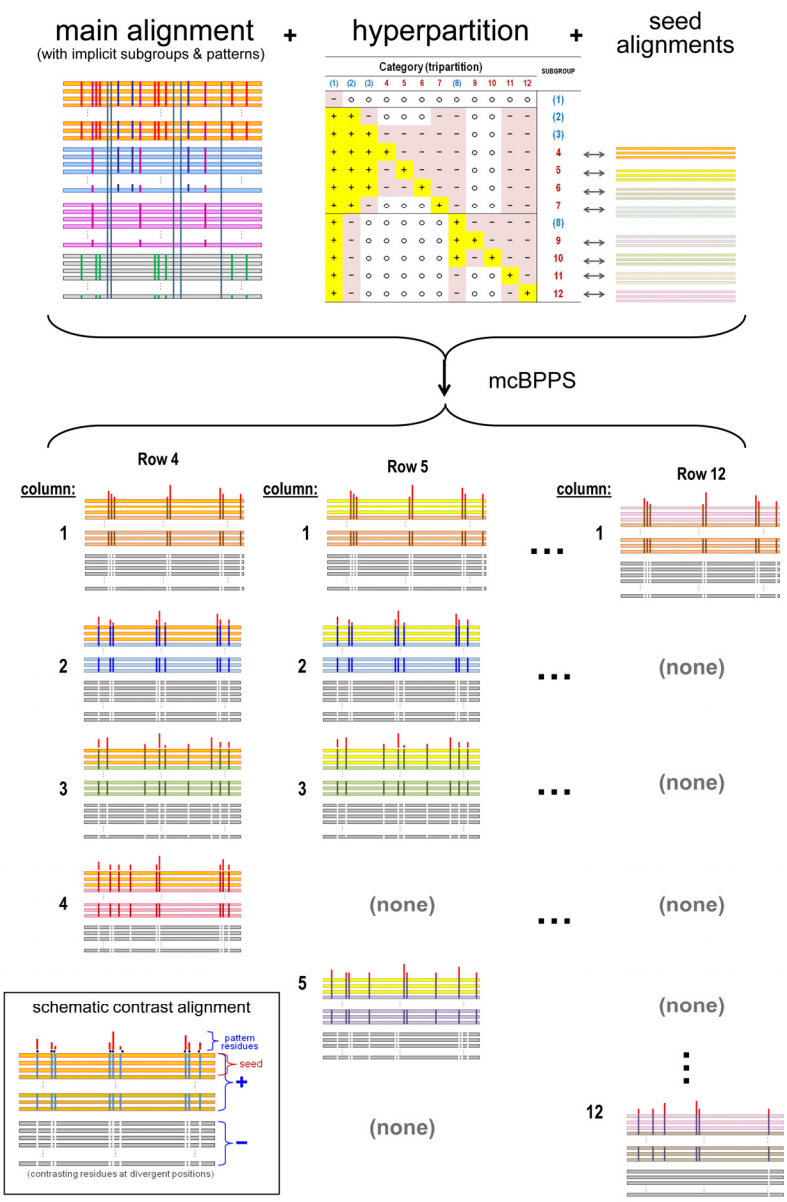


Figure S1. Flow chart for the mcBPPS sampler's input and output. **Input (top panel)**. An input sequence alignment containing functionally-divergent subgroups (represented schematically as distinctly colored horizontal bars) and associated patterns (distinctly colored vertical bars) that the sampler aims to identify. The hyperpartition corresponds to the phylogenetic tree of the subgroups (see Fig. 2 of the paper). The seed alignments also are shown schematically (as distinctly colored horizontal bars). The doubly-headed arrows show the correspondence between each seed alignment, which corresponds to a leaf of the phylogenetic tree, and each row of the hyperpartition. The remaining rows correspond to non-terminal nodes of the tree (i.e., to miscellaneous subgroups within the hyperpartition). **Output (bottom panel)**. By default, the sampler outputs one contrast alignment for each hyperpartition cell that corresponds both to a leaf-node row and to a foreground assignment (indicated by a '+' in that cell). Note that the rows and columns have been transposed (relative to the hyperpartition) in order to fit within the page margins and that, for each columns, the pattern and partition identified by the sampler is the same (only the seed alignments are different). **Schematic contrast alignment (lower left corner box in bottom panel)**. Each contrast alignment is partitioned by the sampler into a 'foreground' set (colored horizontal bars denoted by a '+') and a 'background' set (gray horizontal bars denoted by '-'). Partitioning is based on the presence of conserved foreground residues (blue vertical bars) that diverge from (or contrast with) the background residues at pattern positions (white vertical bars). For this reason, the output is termed a "contrast alignment". The sampler assigns each sequence to a partition indirectly; that is, by assigning it to one of the subgroups. The heights of the red bars above the alignment quantify the selective constraints imposed on divergent residue positions. Actual examples of contrast alignments are shown below.

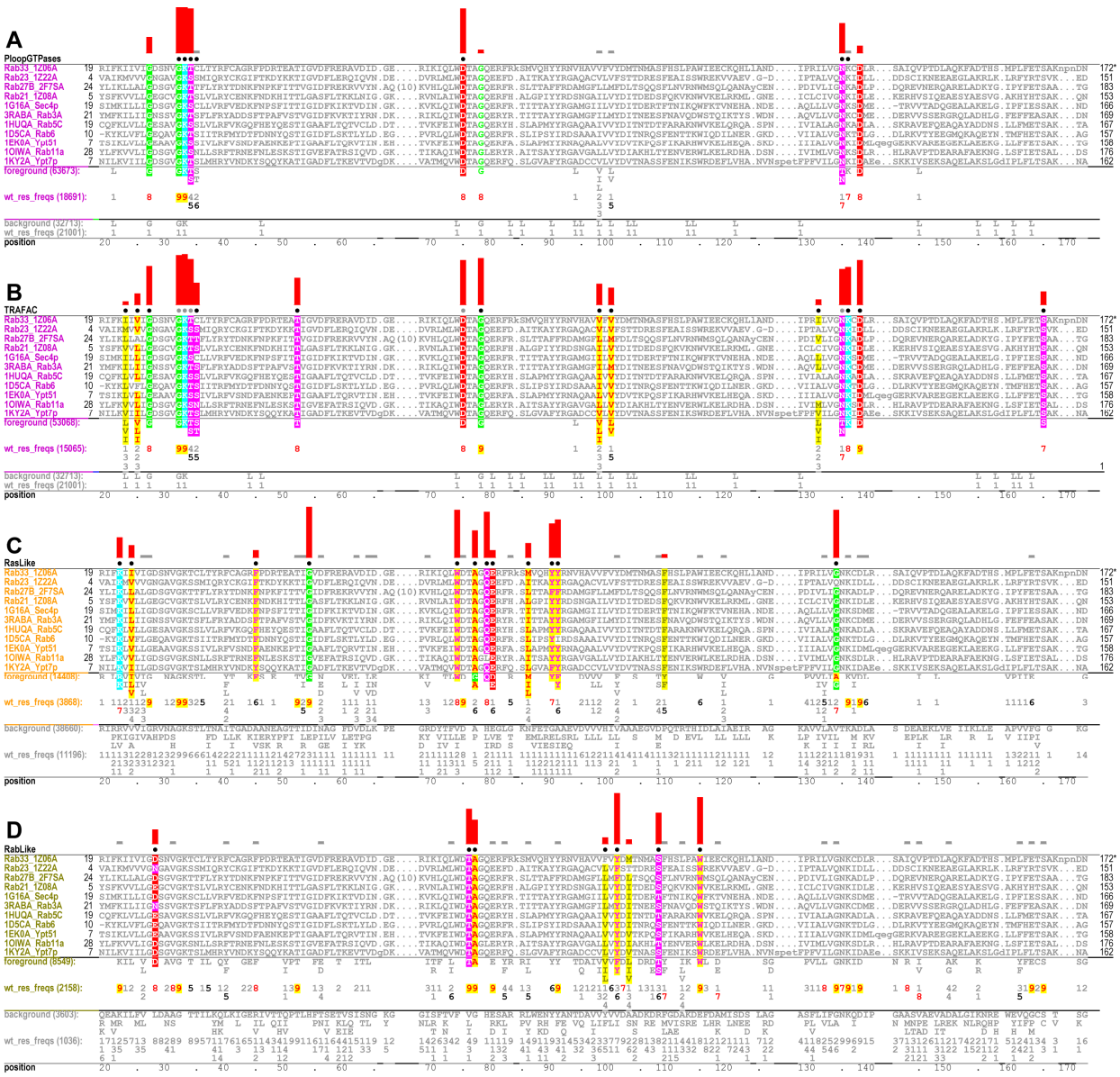


Fig. S2. Contrast alignments for Rab GTPases (corresponding to Fig. 3 in the paper). The seed alignment for the Rab subgroup (row 4 of Table 1 in the paper) is highlighted in four ways to reveal patterns characteristic of the functionally-divergent categories for which Rab is in the foreground: **(A)** Patterns distinguishing all P-loop GTPases from other proteins (column 1 in Table1). **(B)** Patterns distinguishing the TRAFAC class of GTPases from other proteins (column 3); **(C)** patterns distinguishing Ras-like from other TRAFAC GTPases (column 6); and **(D)** patterns distinguishing Rab, Ran and Rho from Arf/Arl/Gα GTPases (column 10). Characteristic foreground and background residues at each position are shown below each alignment and, directly below these, corresponding frequencies are given in integer tenths (there are too many sequences to show the actual alignments); a '7', for example, indicates that the corresponding residue occurs in 70-80% of the sequences. The histograms above highlighted columns quantify selective pressures imposed on pattern residues. Black dots below each histogram bar correspond to pattern positions selected by the sampler; gray dots (or no dot) indicates that this otherwise differentiating position lost out in the competition between categories for pattern positions that was set up by the sampler. Note that a slightly lower contrast setting was used for this output than was used for the output in Fig. 3 of the paper in order to illustrate this feature of the mcBPPS program. As a result, a few weakly-co-conserved residues are highlighted here that are not highlighted in Fig. 3.

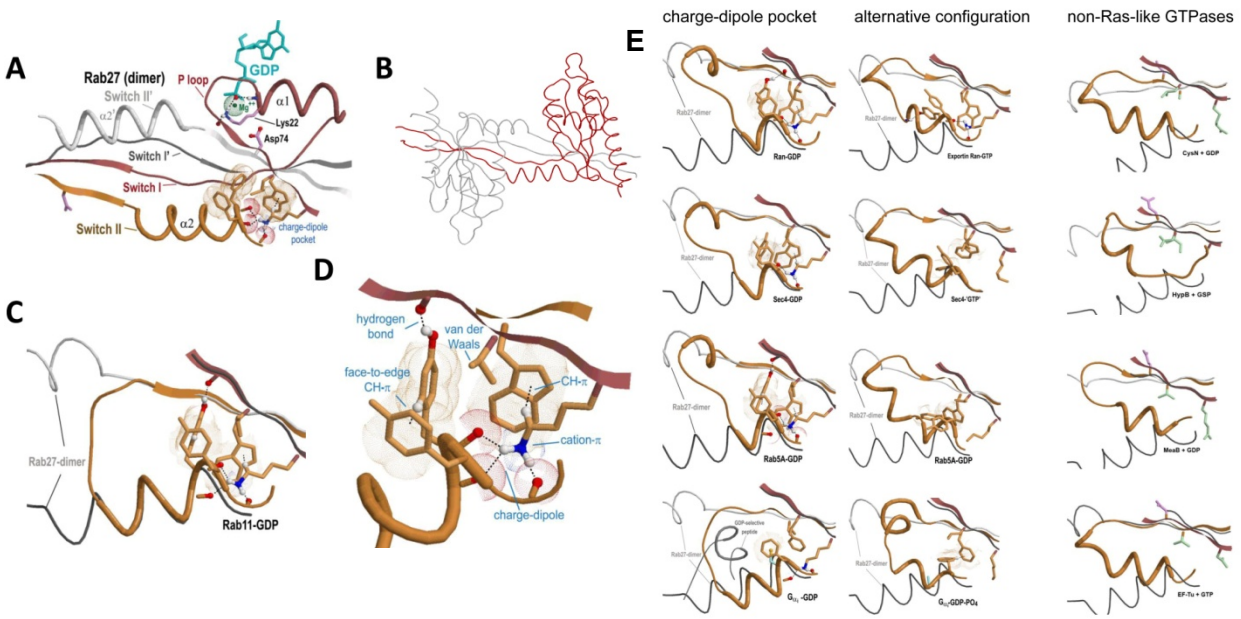


Fig. S3. The charge dipole pocket. This figure was adapted from (37). **(A)** The charge-dipole pocket within the structure of homodimeric Rab27 bound to GDP (pdb_id: 2if0 (38)). The inter-switch regions of the two subunits are exchanged and, as a result, each of the switch II regions forms an α helix that is directed away from the GTPase structural core. Color scheme: backbone of the P-loop, switch I and adjacent regions, *dark red*; backbone of the switch II region and adjacent β strand, *orange*; backbone of the adjacent subunit's switch and inter-switch regions, *gray*; side chains of the Walker A lysine and Walker B aspartate residues, *magenta*. **(B)** Backbone trace of the Rab27 homodimeric structure. **(C)** The Ras-like GTPase Rab11 both forms a charge dipole pocket and possesses an α helix that traces out the same path as is seen for the Rab27 homodimer. **(D)** Atomic interactions associated with the charge-dipole pocket (Rab11A + GDP; pdb_id: 1oiv (39)). Hydrogen bonds are indicated by dotted lines; aromatic and electrostatic interactions by dot clouds. Color scheme: backbone of the β-strand prior to the P-loop, *dark red*; backbone of the switch II region, *orange*; side chains of residues forming the charge-dipole pocket, *orange*; oxygen, nitrogen, and hydrogen atoms involved in hydrogen bonds are *red, blue* and *white*, respectively. **(E)** The switch II helices of various monomeric Ras-like GTPases that form the charge-dipole pocket configuration trace out a path that superimposes over the homodimeric form of Rab27, whereas other GTPases do not. Gα, which harbors a canonical Tyr-to-Cys substitution, may require other cellular components to form the charge-dipole pocket configuration

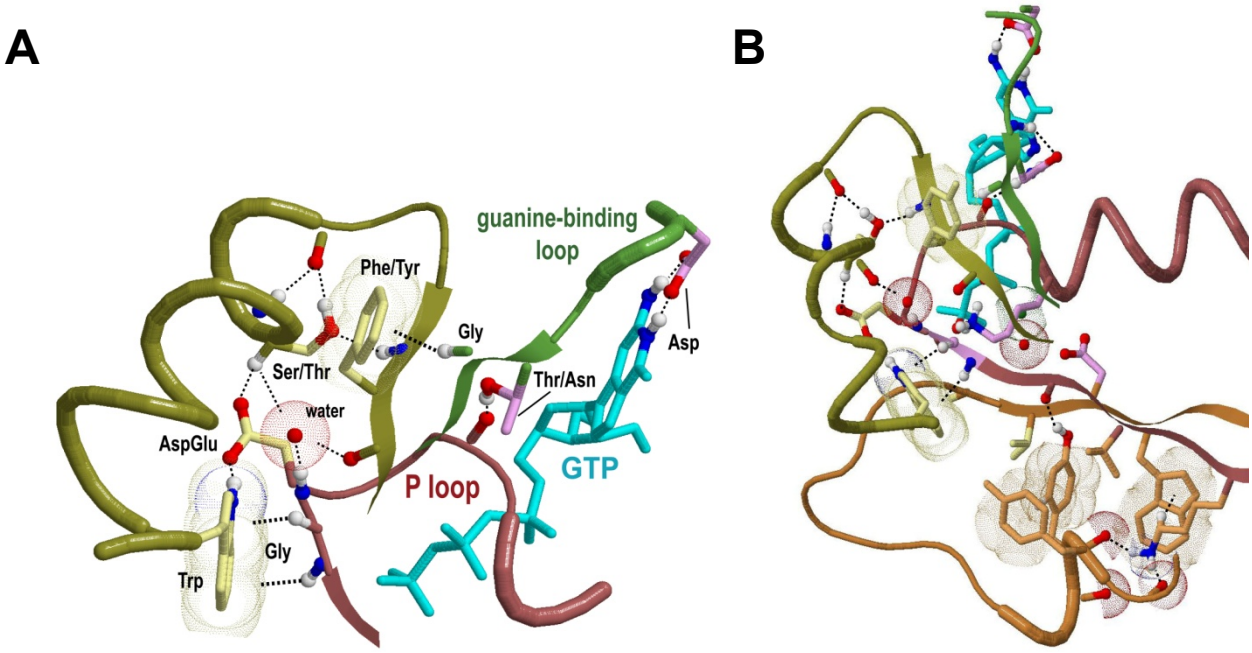
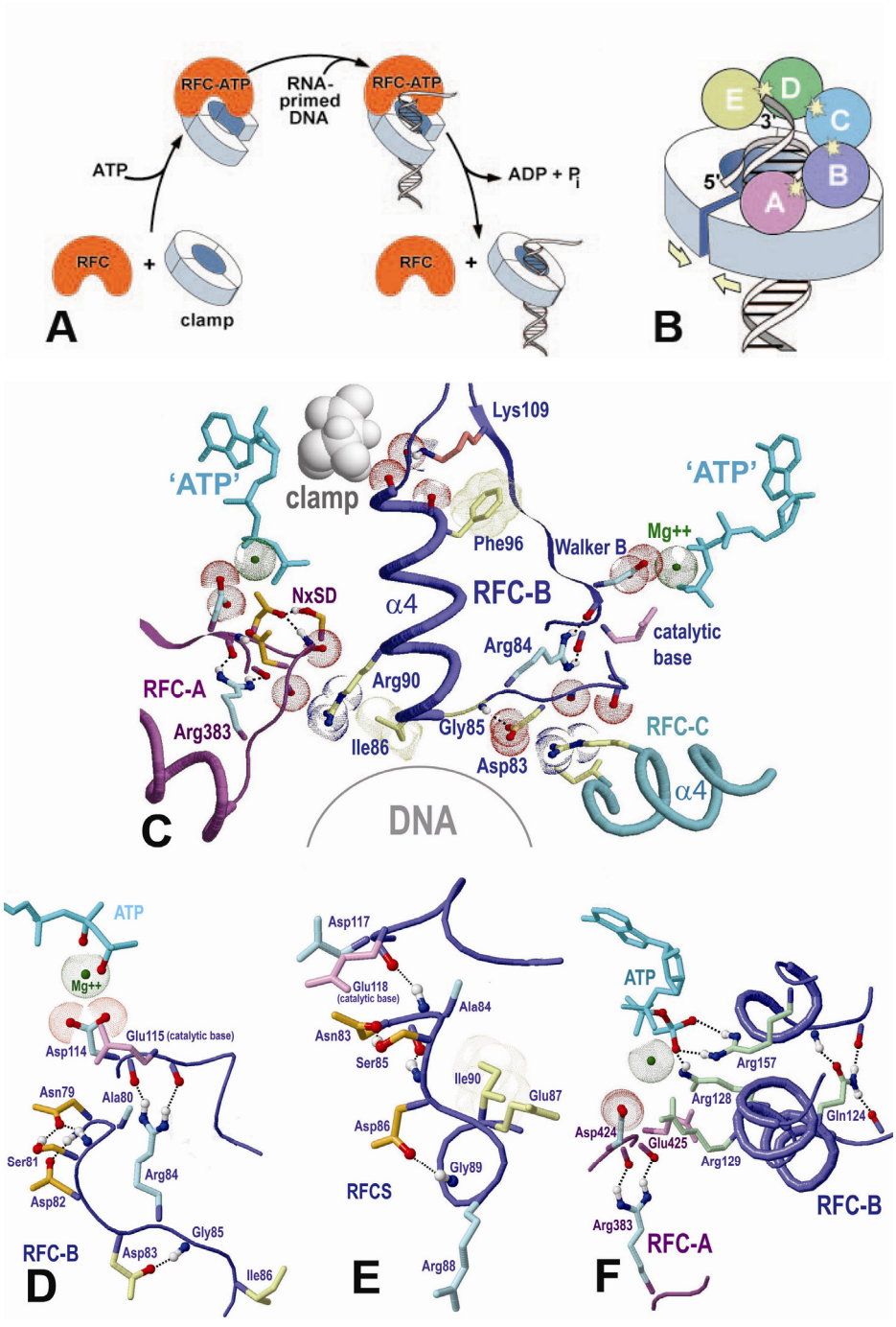


Fig. S4. The proposed glycine brace component. **(A)** The glycine brace in Rac1 GTPase (pdb_id: 1i4t) (40). This figure was adapted from (11) . Color scheme: backbone of the P-loop region, *dark red*; backbone of the glycine brace region, *dark yellow*; backbone of the guanine-binding loop region, *green*; side chains of residues characteristic of all P-loop GTPases, *magenta*; side chains of residues characteristic of Rab, Rho and Ran GTPases, *yellow*. **(B)** Location of the glycine brace relative to the charge-dipole pocket (within Rab11 bound to a GTP analog (39)).

Fig. 5. Biochemical and structural features of RFC DNA clamp loaders. **(A)** Reaction catalyzed by RFC clamp loaders. In the presence of ATP, the clamp loader binds to and opens the clamp and, upon association with 3' recessed RNA-primed DNA, hydrolyzes ATP to load the clamp. **(B)** Model of the eukaryotic RFC complex (consisting of subunits A to E) loading a clamp onto DNA. Subunits A to D undergo ATP hydrolysis (yellow explosions) resulting in closure of the clamp around DNA. **(C)** Regions of interaction between RFC-A, RFC-B and RFC-C ATPases within the crystal structure of the RFC-ATP-clamp complex (16) showing the structural features associated with the categories of RFC pattern residues identified here. The central hole of the RFC complex, through which DNA presumably is thread, is located at the bottom of each subfigure. Oxygen, nitrogen and (predicted) hydrogen atoms establishing hydrogen bonds (dotted lines) or involved in ionic interactions are colored red, blue, and white, respectively. Ionic and van der Waals interactions are shown as dot clouds. Residue with cyan colored side-chains distinguish active RFC ATPases from catalytically impaired RFC-E subunits; the magenta residue corresponds to the putative catalytic base shared by all active AAA+ ATPases; orange residues distinguish all RFC subunits from bacterial clamp loader ATPases; yellow residues distinguish active RFC ATPases that interact with an adjacent ATP site from other RFC subunits; the red residue in 'A' most distinguishes all clamp loader subunits interacting with an adjacent ATP site from other AAA+ ATPases; and green residue side-chains correspond to residues that are characteristic of RFC subunits that interact with the ATP-binding site of an adjacent subunit. The most buried residue of the DNA clamp upon binding to the RFC-complex is shown as a space filling model (white spheres) located near the end of the α4 helix of RFC-B. This figure was adapted from (14). Note that residue side-chains corresponding to the NxSD motif in RFC-B have been omitted for clarity. A conserved phenylalanine in RFC-B (Phe96) appears to form a hydrophobic pocket for Lys109. **(D)** Conformation of the RFC-B Walker B region when bound to ATP and the clamp. In all four active RFC ATPases, the arginine corresponding to Arg84^{RFC-B} hydrogen bonds to Walker B main chain oxygen atoms as shown. **(E)** The conformation of the ADP-bound form of RFCS, the archaeal small subunit corresponding to RFC-B (18). The Walker B-arginine (Arg84^{RFC-B} and Arg88^{RFC-S}) is repositioned and, based on a model of the RFC–DNA–clamp complex, could interact with DNA thread through the clamp. **(F)** Region of RFC-B that interacts with the ATP-binding site of the adjacent RFC-A subunit.



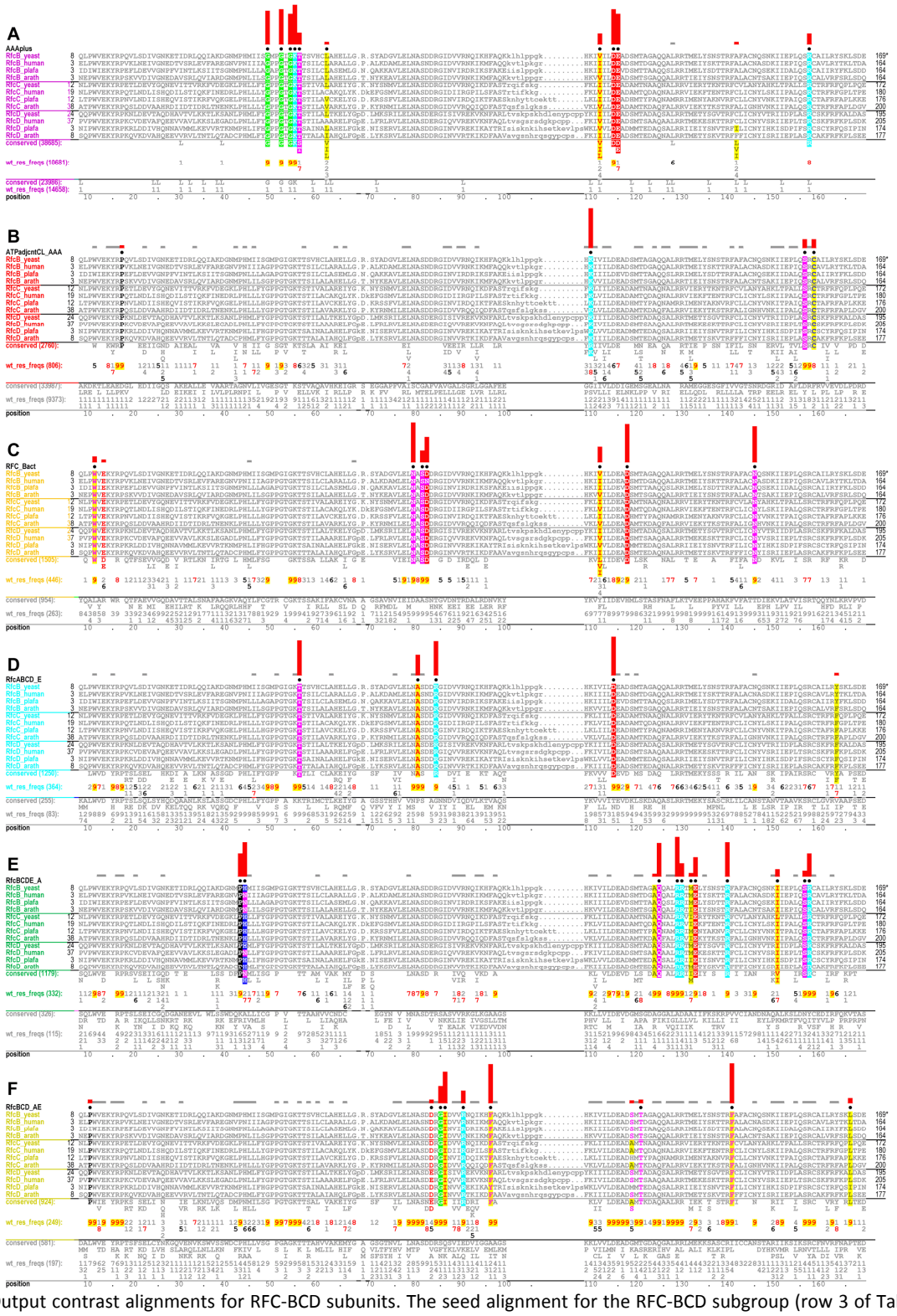
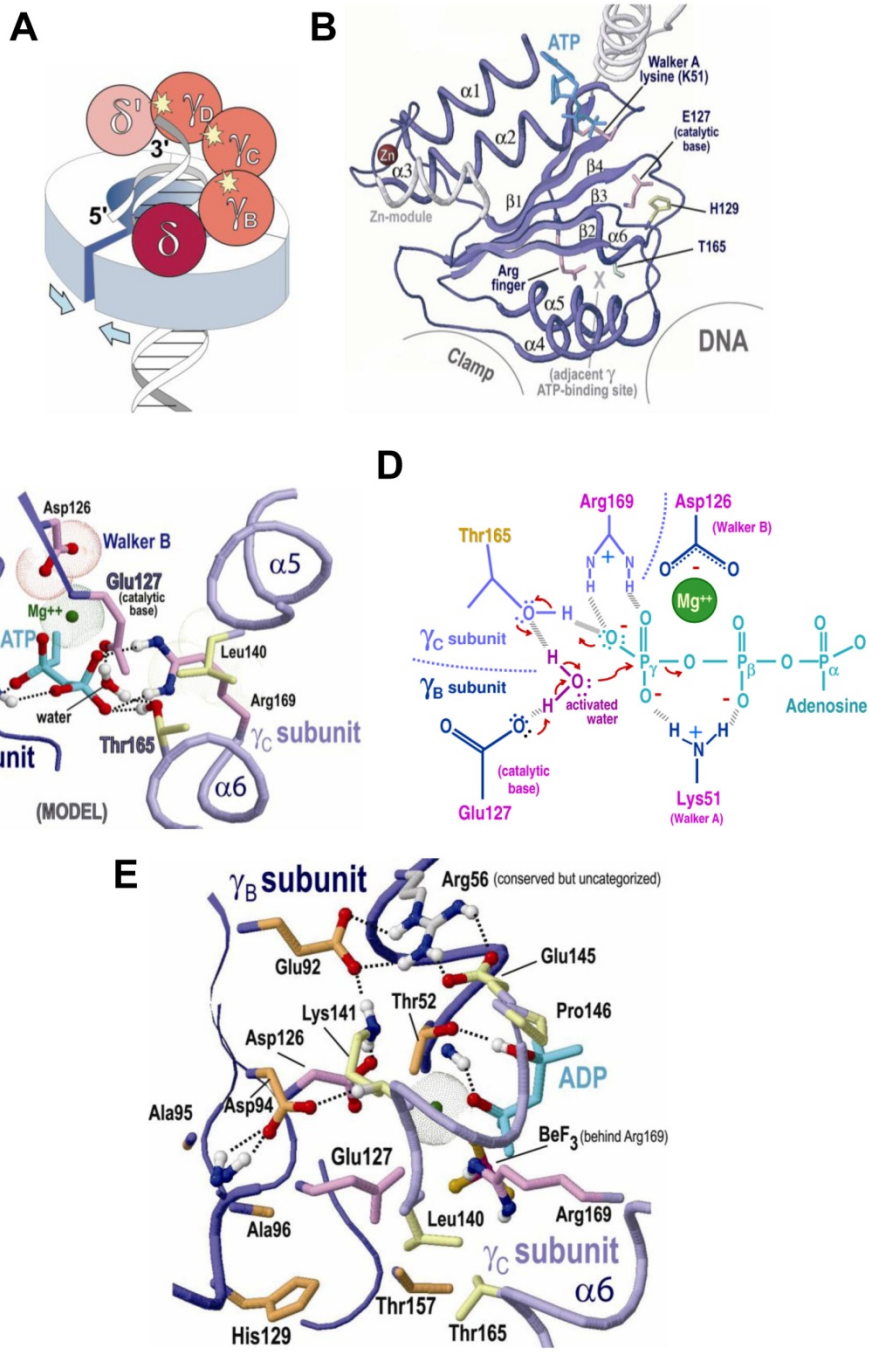


Fig. S6. Output contrast alignments for RFC-BCD subunits. The seed alignment for the RFC-BCD subgroup (row 3 of Table 2 in the paper) is highlighted in six ways to reveal patterns characteristic of the functionally-divergent categories for which RFC-BCD is in the foreground and that compare: **(A)** all AAA+ ATPases vs other proteins (column 1 in Table 2); **(B)** clamp loader subunits adjacent to ATP-binding sites vs other AAA+ ATPases (column 6); **(C)** RFC vs bacterial γ subunits (column 8); **(D)** active RFC ATPases vs inactive RFC-E subunits (column 9); **(E)** RFC subunits adjacent to ATP-binding sites vs other RFC subunits (column 10); and **(F)** internal vs 'bookended' RFC subunits (column 12). For an explanation of the format and notations see the legend to Fig. S2.

Fig. S7. Biochemical and structural features of bacterial DNA clamp loaders. **(A)** Arrangement of the five AAA+ subunits (spheres) within the γ clamp loader complex. The clamp is shown as a cylinder through which RNA-primed DNA is thread. Explosions indicate locations of bound ATP within the γ subunits. **(B)** Structure of the AAA+ module within the bacterial γ subunit. Regions of γ's ATPase domain that interact with DNA, the clamp and an adjacent γ subunit 's ATP-binding site are indicated. **(C)** Hypothetical model of a plausible interaction (in trans) between a threonine residue that is conserved within γ (or δ') subunits (Thr165-γ in *E. coli*) and an adjacent γ subunit's ATP binding site. Magenta-colored side-chains correspond to residues that are co-conserved within all AAA+ ATPases and that play critical roles in ATP hydrolysis, whereas yellow-colored side-chains correspond to residues co-conserved in γ and δ' but not within corresponding RFC subunits (i.e., RFC-BCDES). Leu140 packs against both Thr165 and the trans-activating arginine finger (Arg169) and thus may play a role in the relative positioning of these residues. **(D)** A plausible mechanism for ATP-hydrolysis where the catalytic base (Glu127) and Thr165 extract hydrogen atoms from a nearby water molecule and thereby activate it for nucleophilic attack on the γ-phosphate of ATP. **(E)** The. structural locations (pdb_id: 3glf) of residues (yellow side-chains) that are co-conserved within γ and δ', but not within corresponding RFC subunits and of residues (orange side-chains) that are co-conserved within γ subunits, which are active ATPases, but not within δ' subunits, which are inactive. Locations of hydrogen atoms (white spheres) participating in hydrogen bonds (dotted lines) were predicted using the reduce program (41). Oxygen and nitrogen atoms participating in hydrogen bonds are colored red and blue, respectively.



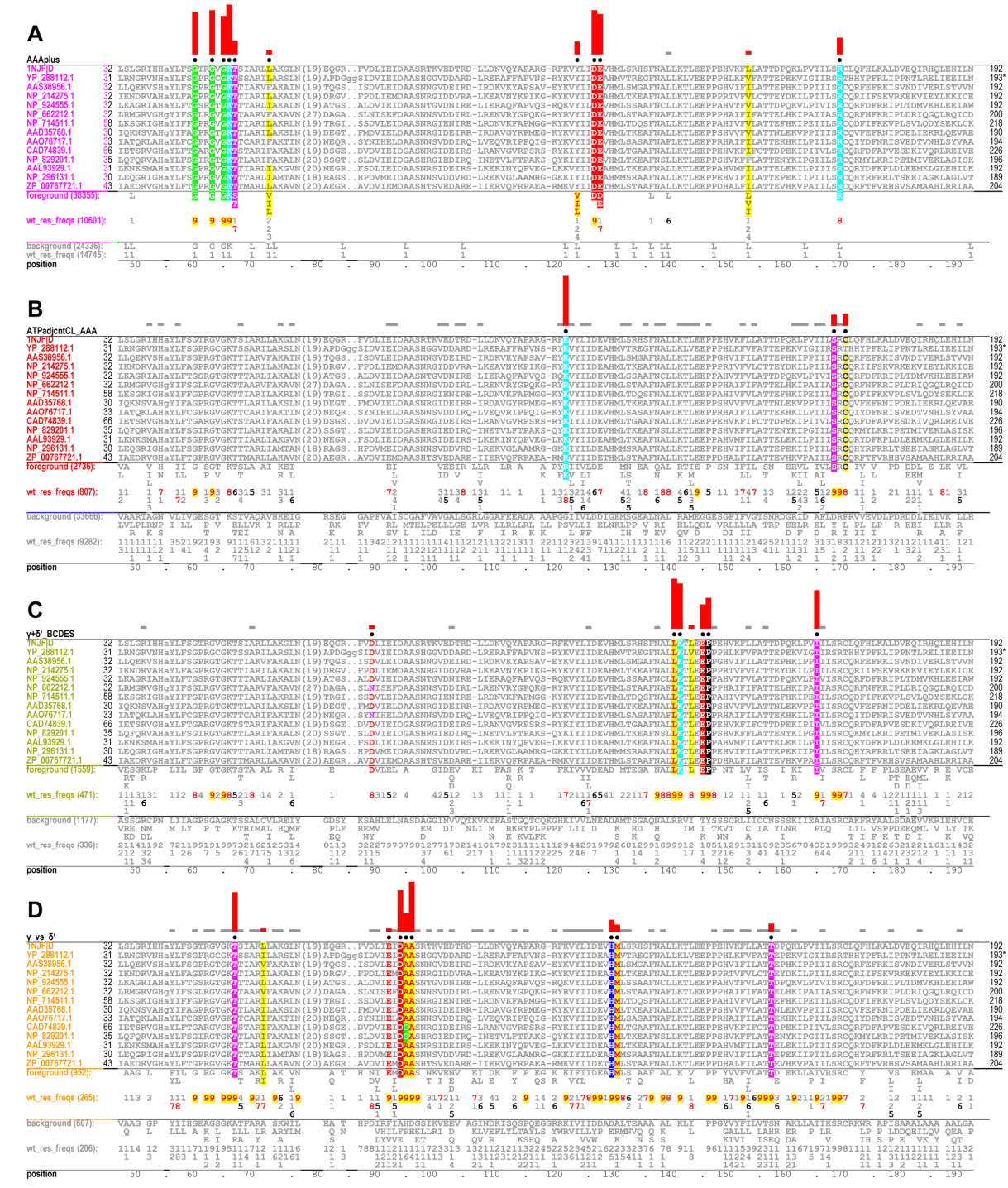


Fig. S8. Output contrast alignments for bacterial γ subunits. These contrast alignments highlight the seed alignment for the γ subgroup (row 6 of Table 2 in the paper) in four ways to reveal patterns characteristic of the functionally-divergent categories, for which the γ subunit is in the foreground. The foreground and background sets being compared are: **(A)** all AAA+ ATPases vs other proteins (column 1 in Table 2); **(B)** all clamp loader subunits adjacent to ATP-binding sites vs other AAA+ ATPases (column 6); **(C)** bacterial γ and δ' subunits vs corresponding RFC subunits (column 16); and **(D)** γ (active ATPase) subunits vs inactive δ' subunits (column 17). For an explanation of the format and notations see the legend to Fig. S2.

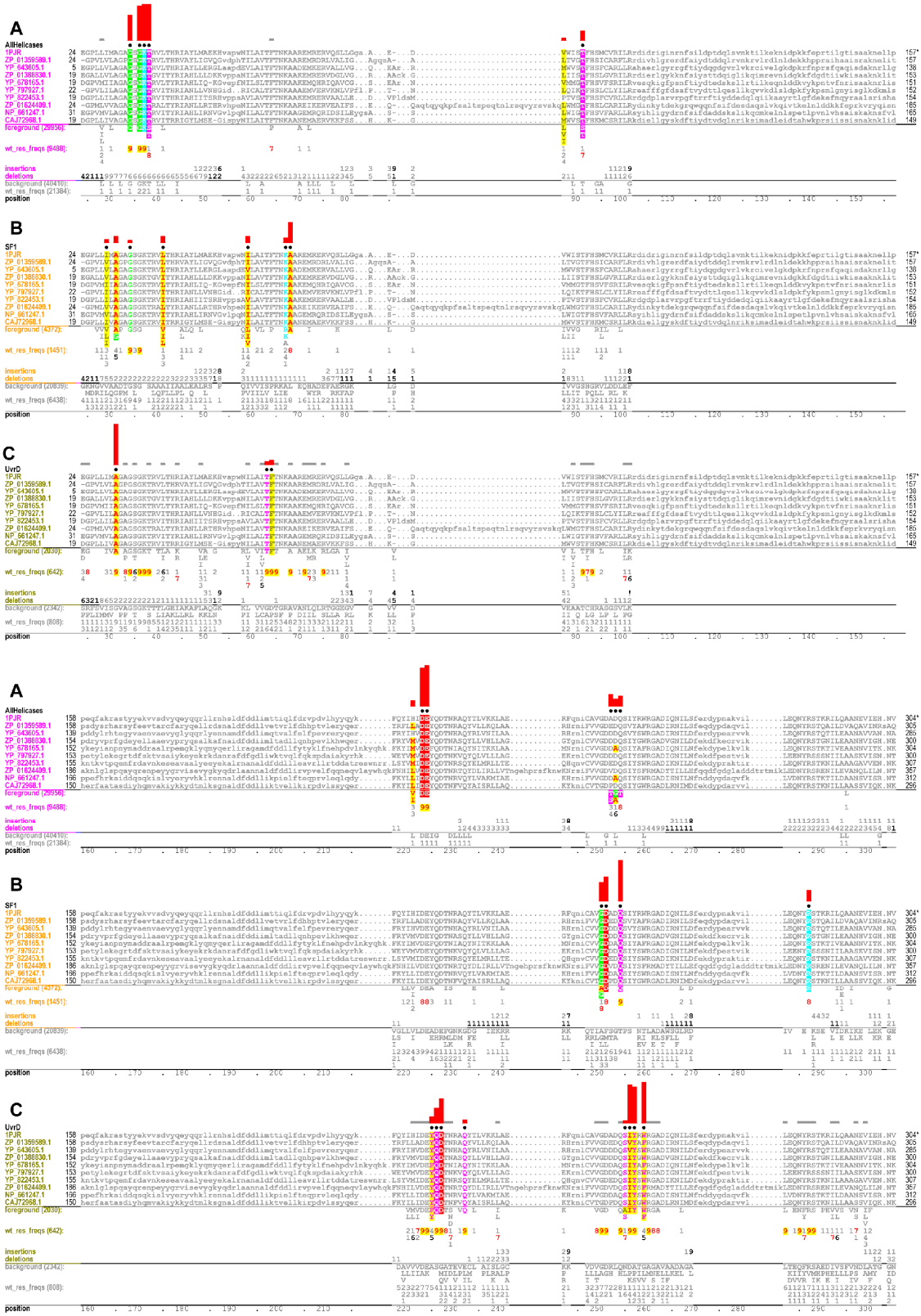


Fig. S9. Output contrast alignments for the subgroup in row 1 of Table S3, which was seeded with PcrA DNA helicases. **(A)** Contrast alignment of all helicases versus other proteins (column 1 in Table S3); **(B)** Contrast alignment of SF1 helicases vs SF2 helicases (column 3); **(C)** Contrast alignment of SF1 helicases closely related to PcrA vs other SF1 helicases (column 14). For an explanation of the format and notation see the legend to Fig. S2.

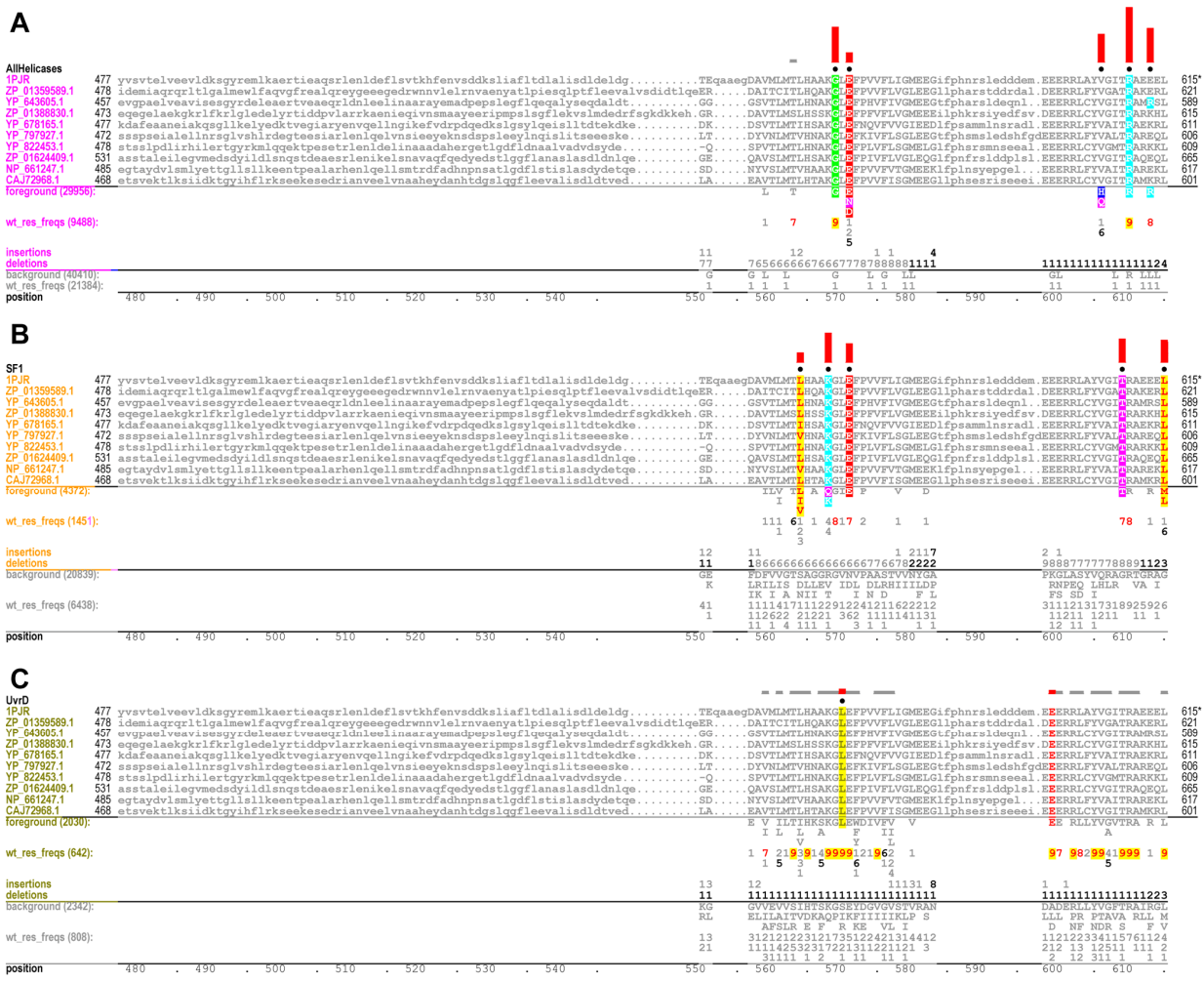


Fig. S9. (continued).

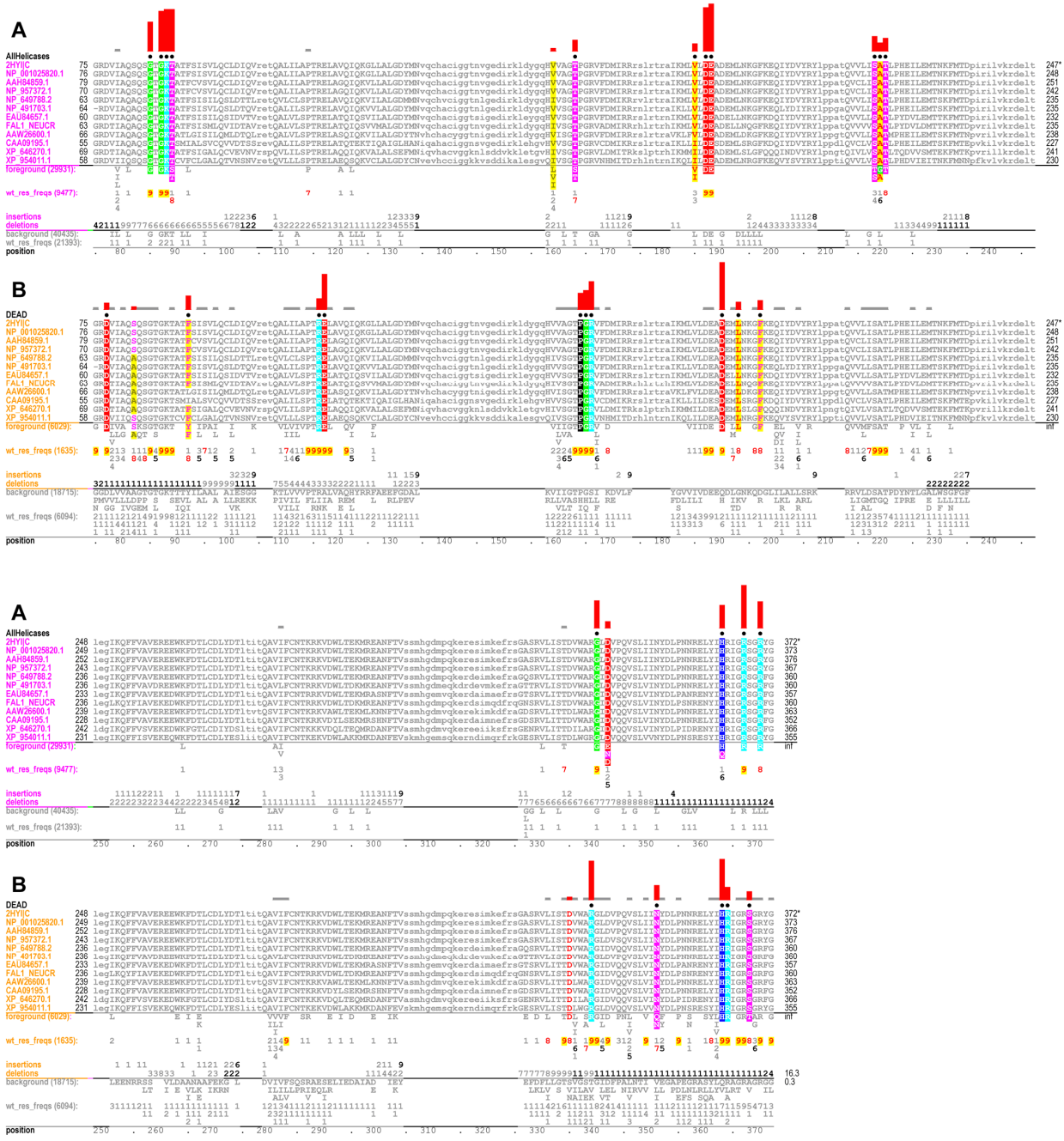


Fig. S10. Output contrast alignments for the subgroup in row 3 of Table S3 which was seeded with EIF4A3 Helicases. **(A)** Contrast alignment of all Helicases vs other proteins (column 1 in Table S3); **(B)** Contrast alignment of DEAD box helicases vs other helicases (column 3). For an explanation of the format and notations see the legend to Fig. S2.

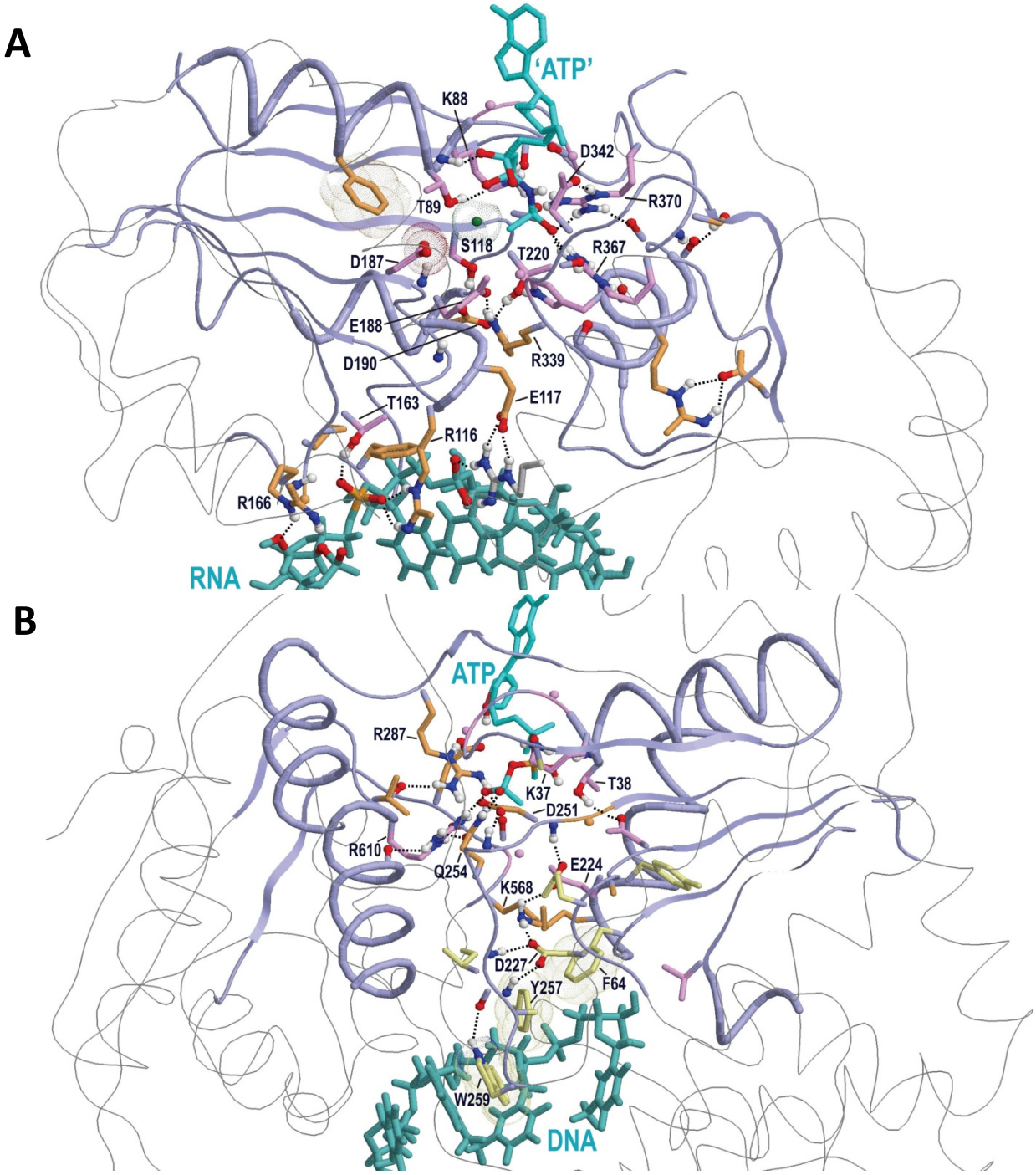


Fig. S11. Structural locations of several categories of co-conserved residues within SF1 and SF2 helicases. Labeled sidechains correspond to pattern residues for which functional and/or structural roles are listed in Supplement 9. **(A)** Structure of a SF2 helicase: human exon junction complex containing trapped eIF4AIII bound to an ATP analog and to RNA (pdb_id: 2hyi). Only eIF4AIII, an ATP analog and the RNA are shown. Side-chains of residues co-conserved in all helicases and in DEAD-box helicases are colored magenta and orange, respectively. **(B)** Structure of a SF1 DNA helicase: PcrA bound to DNA (pdb_id: 3pjr). Side-chains of residues co-conserved in all helicases, in SF1 helicases and in SF1 helicases closely related to PcrA- are colored magenta, orange and yellow, respectively.

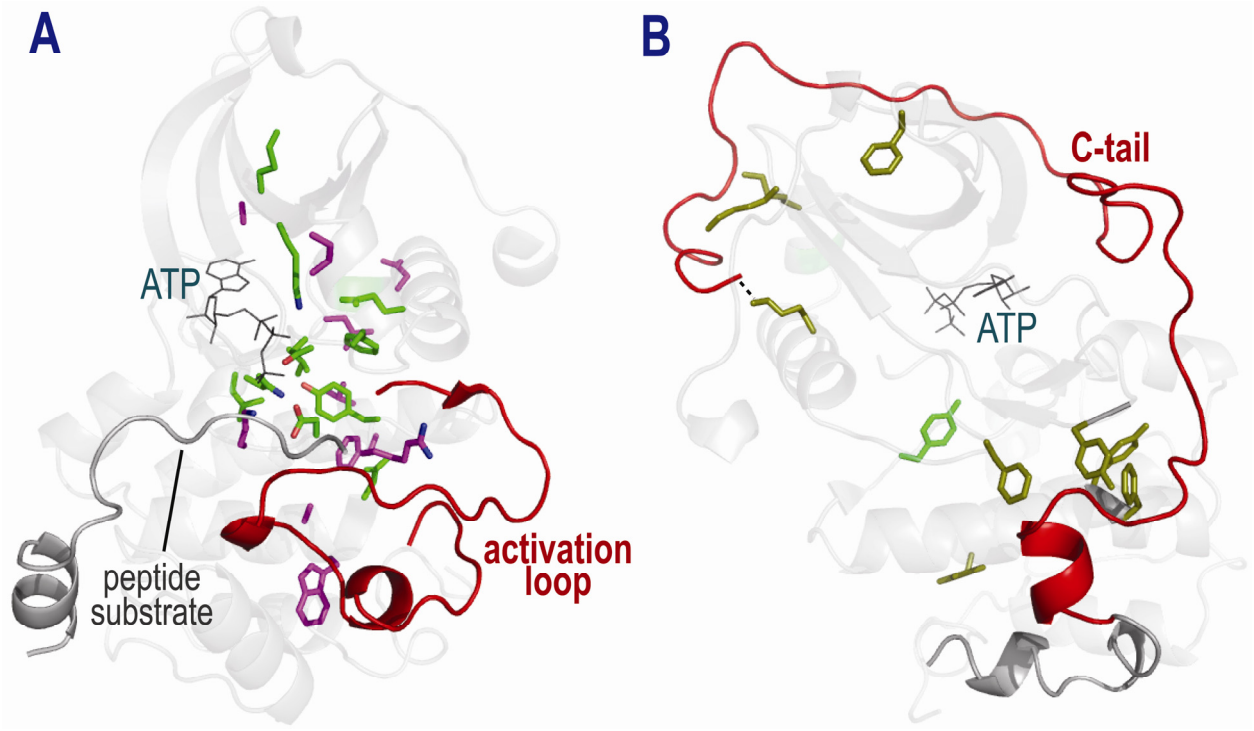


Fig. S12. Structural locations of several categories of co-conserved residues within an AGC protein kinase. **(A)** Structure of the cyclic adenosine monophosphate-dependent protein kinase bound to an ATP analog and to a peptide substrate (33). Side-chains of residues co-conserved in all protein kinases or specifically within EPKs are colored green and magenta, respectively. **(B)** Structure of the same kinase with AGC-specific co-conserved residues highlighted in yellow. The isolated tyrosine residue is highlighted in green because it corresponds to a catalytic histidine residue that is conserved in all protein kinases, but that is often a tyrosine specifically within AGC kinases.

Fig. S13. Output contrast alignments for protein kinases. The seed alignment for the AGC subgroup (row 1 of Table S5) is highlighted in three ways to reveal patterns characteristic of the three functionally-divergent categories for which AGC is in the foreground. These categories compare: **(A)** all protein kinases vs other proteins (column 1 in Table S5); **(B)** EPKs vs ELKs and APKs (column 2); **(C)** AGC kinases vs other EPKs (column 8). For an explanation of the notation and formatting see the legend to Fig. S2-4.

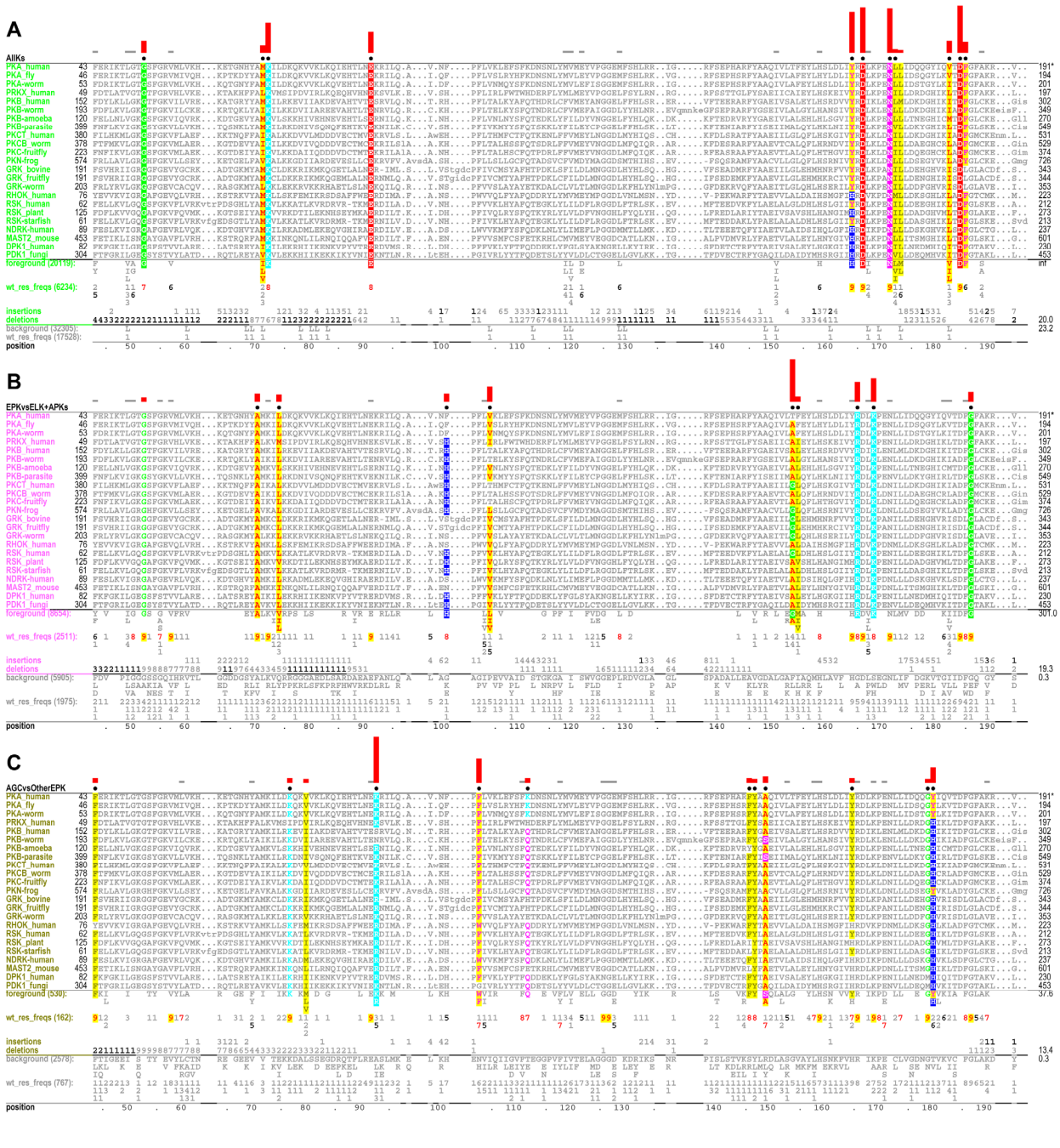
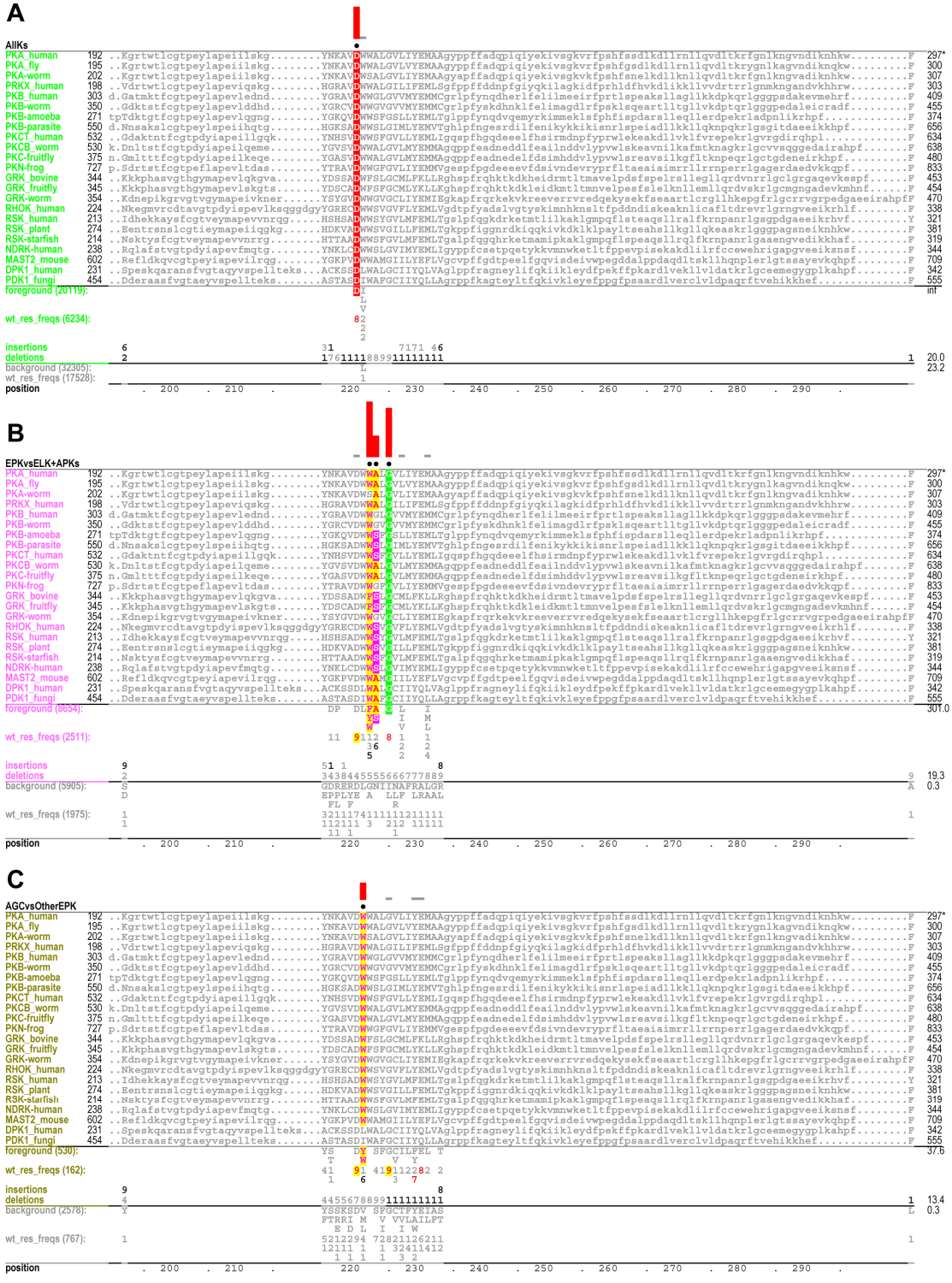


Fig. S13. (Continued)



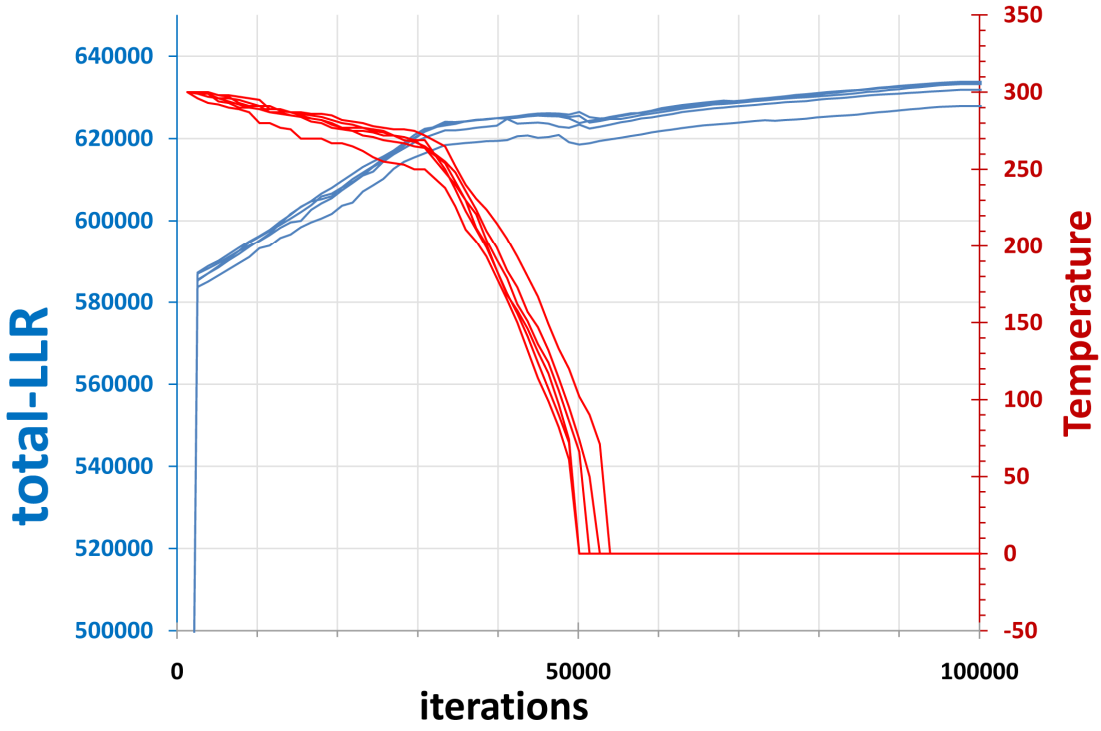
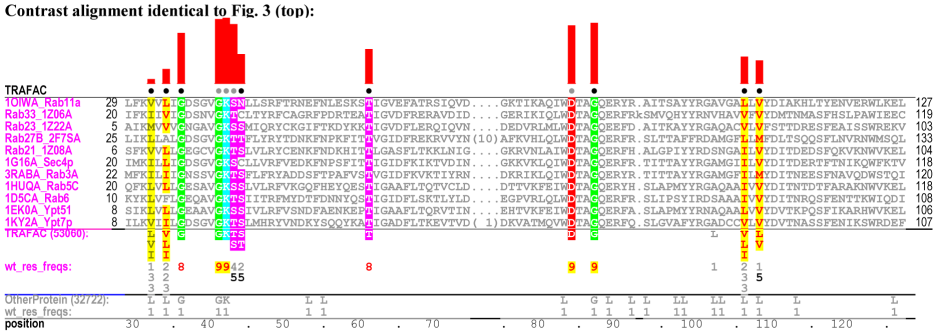
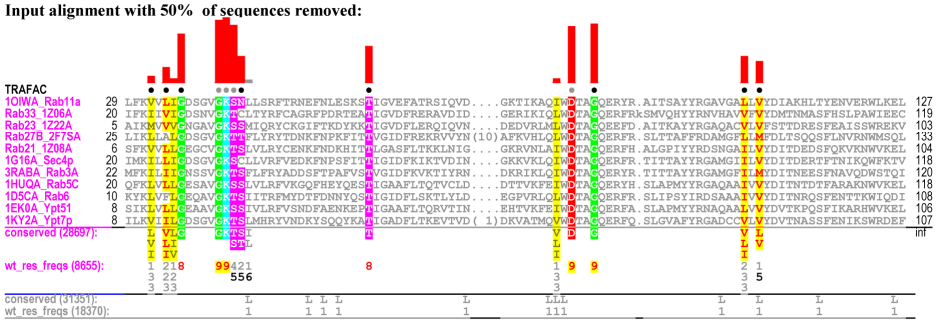


Fig. S14. Convergence behavior of the mc-BPPS sampler. Plots correspond to five analyses using the hyperpartition in Table 1 of the paper and—to better assess inherent variability—where, in each case, 50% of input aligned sequences were randomly removed. The total (multiple category) log likelihood ratio (LLR) and the (simulated annealing) sampling temperature (which is arbitrarily set to 300° for the true distribution) are plotted as functions of the number of sampling steps (iterations). The annealing scheme is as follows: Whenever the slope of the total-LLR function over a 50 iteration window remains fairly flat (defined as a change in the LLR of less than one nat) the temperature is dropped by one degree; this continues until a temperature of zero is reached.

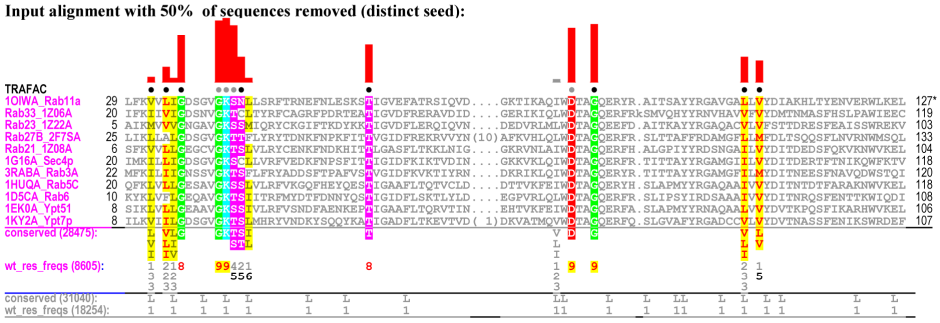
Contrast alignment identical to Fig. 3 (top):



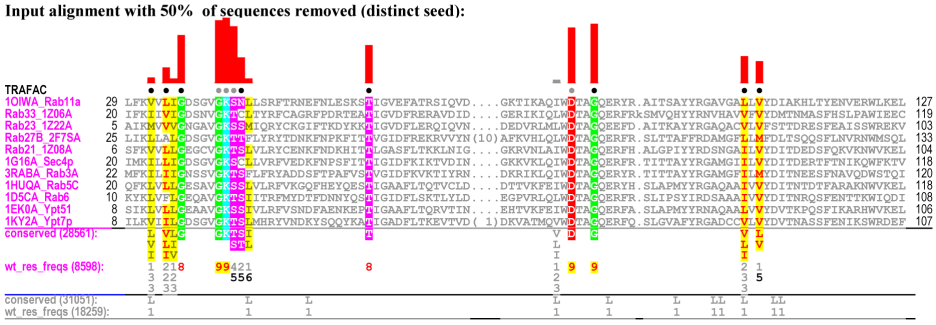
Input alignment with 50% of sequences removed:



Input alignment with 50% of sequences removed (distinct seed):



Input alignment with 50% of sequences removed (distinct seed):



Independently-obtained (PSI-BLAST) input alignment:

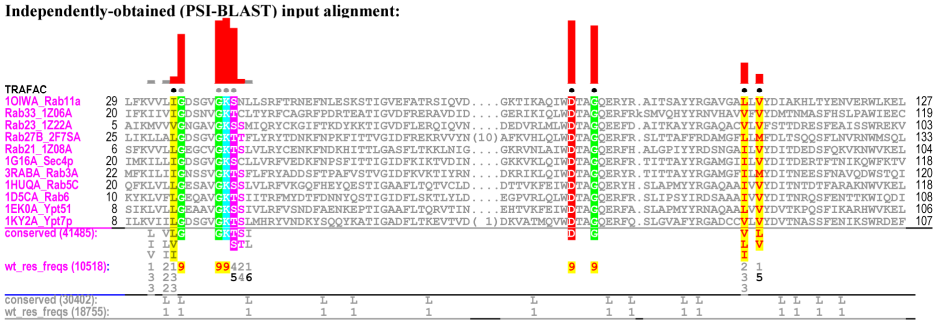
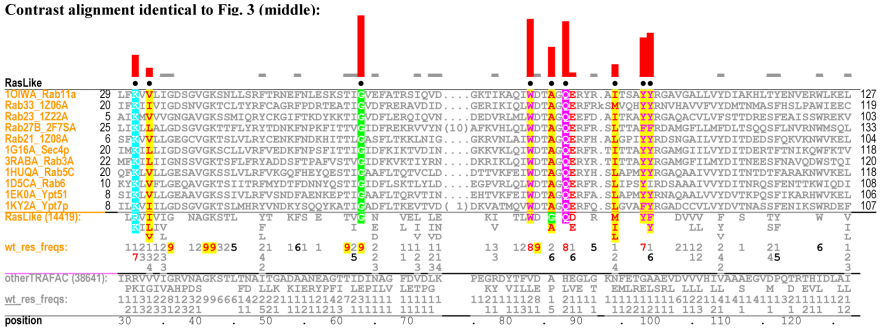
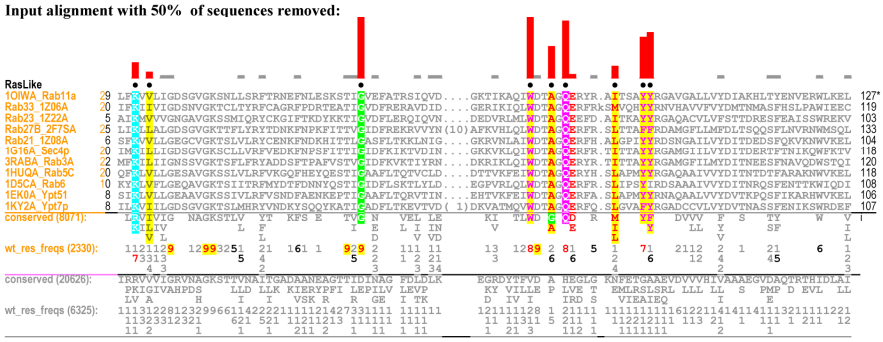


Fig. S15. The robustness of the mcBPPS sampler (see text).

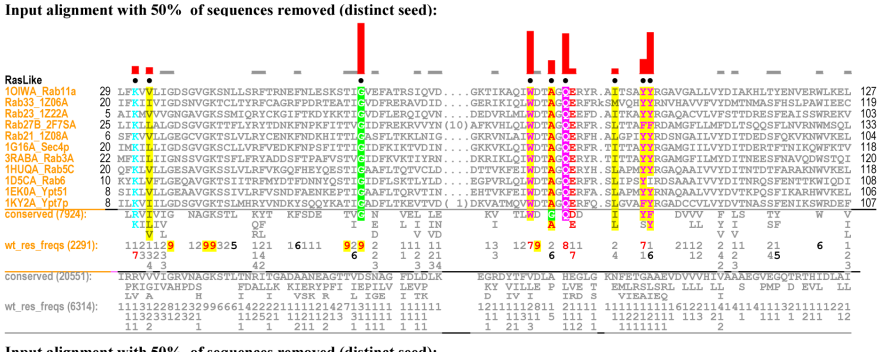
Contrast alignment identical to Fig. 3 (middle):



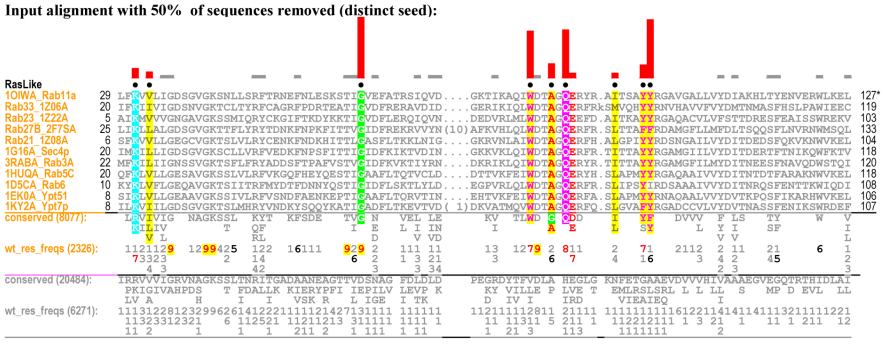
Input alignment with 50% of sequences removed:



Input alignment with 50% of sequences removed (distinct seed):



Input alignment with 50% of sequences removed (distinct seed):



Independently obtained (PSI-BLAST) input alignment:

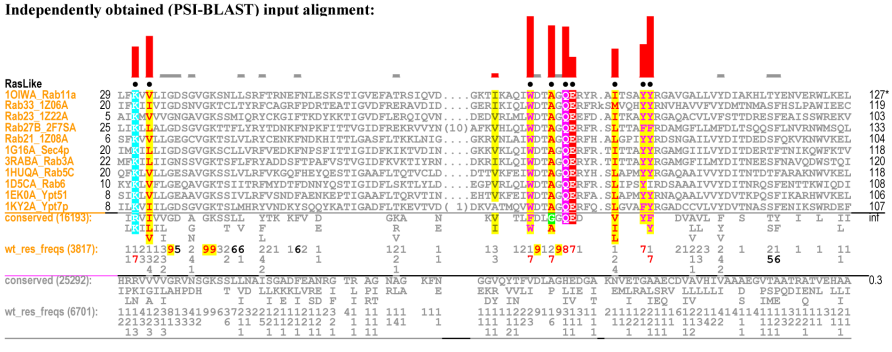
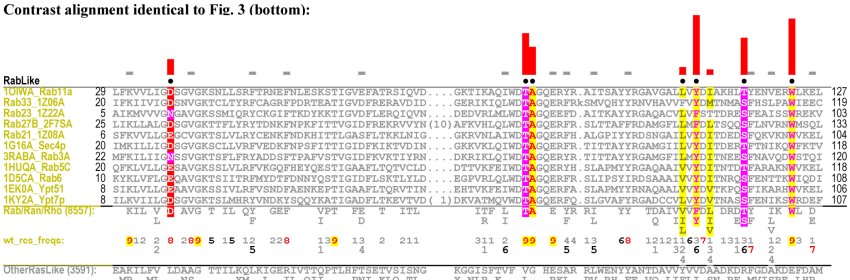
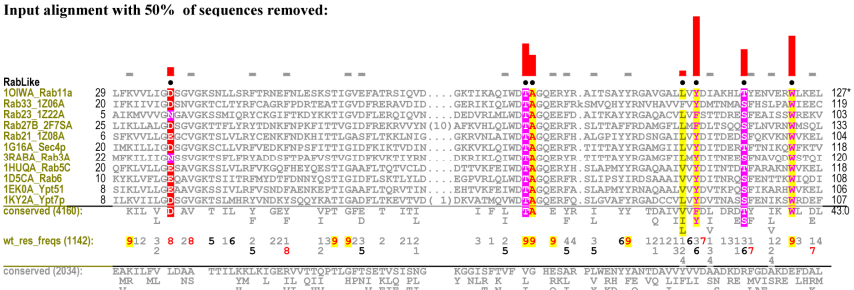


Fig. S15. (continued).

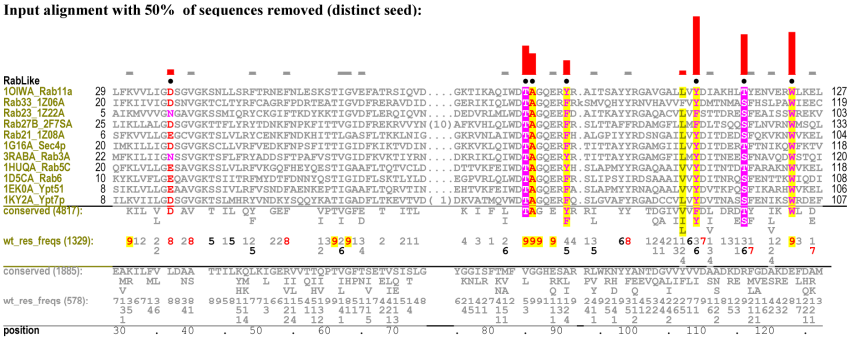
Contrast alignment identical to Fig. 3 (bottom):



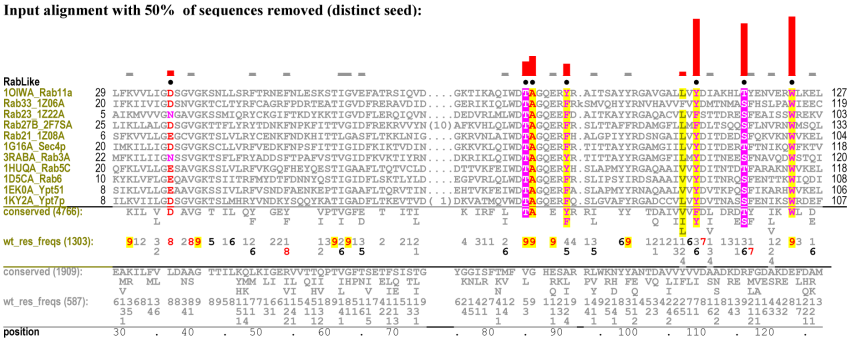
Input alignment with 50% of sequences removed:



Input alignment with 50% of sequences removed (distinct seed):



Input alignment with 50% of sequences removed (distinct seed):



Independently obtained (PSI-BLAST) input alignment:

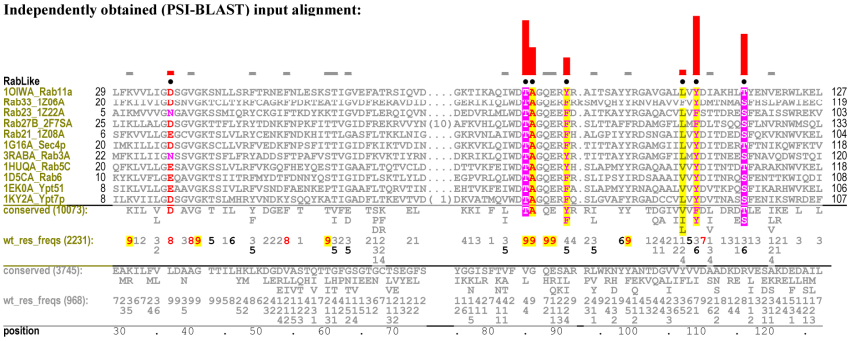


Fig. S15. (continued).

Table S1. Functional/structural relevance of pattern residues within Ras-like GTPases.

Residue*	Functional/structural relevance	References
Lys13	Forms charge-dipole pocket associated with atypical switch II α helix	(37)
Val15	Forms charge-dipole pocket associated with atypical switch II α helix	(37)
Gly18	Hinge point for the P-loop; part of Walker A motif	(3)
Gly23	Main chain hydrogen bond with GTP-phosphate group	(3)
Lys24	Side-chain hydrogen bond with nucleotide phosphate groups	(3)
Ser25	hydrogen bonds with phosphate groups and Mg^{++} ion	(3)
Phe36	Forms an aromatic-aromatic interaction with the guanine base	(3)
Thr43	hydrogen bonds with phosphate groups and Mg^{++} ion	(3)
Gly45	Switch I region hinge point	(42)
Trp65	Forms charge-dipole pocket associated with atypical switch II α helix	(37)
Asp66	Walker B residue that coordinates with Mg++ ion via water	(7)
Ala68	Blocks the Mg2+ binding site to promote nucleotide exchange	(7, 9)
Gly69	Switch II backbone hydrogen bond with γ-phosphate	(3, 7)
Gln70	Key catalytic residue for GTP hydrolysis	(3, 7)
Glu71	Interacts with the Walker A lysine whenever Mg2+ or guanine nucleotide is absent; has a proposed role in Mg2+ release and nucleotide exchange	(7, 8)
Tyr80	Forms charge-dipole pocket associated with atypical switch II α helix	(37)
Tyr81	Forms charge-dipole pocket associated with atypical switch II α helix	(37)
Tyr91	Aromatic stabilizing interaction (12) with guanine-binding loop hinge point	(11, 42)
Trp105	Aromatic stabilizing interaction (12) with P-loop hinge point	(11, 42)
Gly123	Hinge point for the guanine base binding loop	(37, 42)
Asn124	Links together various subregions of the nucleotide-binding site	(3)
Lys125	Links together various subregions of the nucleotide-binding site	(3)
Asp127	Binding to the guanine base of GTP and GDP	(3)
Ser154	Assists binding to the guanine base	(3)

* Residues correspond to Rab11A and are color coded to correspond to the categories in Fig. S2.

Table S2. Functional/structural relevance of pattern residues within DNA clamp loader subunits.

A. Eukaryotic RFC subunits		
Residue*	**Functional/structural relevance**	**Reference**
Gly49	Walker A motif	(22, 42)
Gly54	Main chain hydrogen bond with GTP-phosphate group	(22, 42)
Lys55	Side-chain hydrogen bond with nucleotide phosphate groups	(22, 42)
Thr56	hydrogen bonds with phosphate groups and Mg^{++} ion	(22, 42)
Arg84	Forms hydrogen bonds with backbone oxygen atoms on either side of the catalytic base	(16)
Arg90	Clamp and DNA binding in corresponding bacterial arginine	(20)
Lys109	Interacts with the C-terminal end of the α4 helix, which binds directly to the clamp	(16)
Asp114	Walker B aspartate: co-ordinates with ATP-associated Mg++	(22, 42)
Glu115	Catalytic base	(17)
Gln124	Forms hydrogen bonds between backbone regions harboring trans-interacting residues	(16)
Arg128	Part of a network of hydrogen bonds that locks down the relative orientation of domain II of the adjacent RFC subunit with respect to ATP	(16)
Arg129	Part of a network of hydrogen bonds that locks down the relative orientation of domain II of the adjacent RFC subunit with respect to ATP	(16)
Glu132	Hydrogen bonds in trans with the "sensor 2" conserved argining	(16)
Arg157	Arginine finger that senses bound ATP and facilitates ATP hydrolysis in trans	(23, 43)
B. Bacterial clamp loader subunits		
Residue*	**Functional/structural relevance**	**Reference**
Gly60	Walker A motif	(22, 42)
Gly65	Walker A: main chain hydrogen bond with GTP-phosphate group	(22, 42)
Lys66	Walker A: Side-chain hydrogen bond with nucleotide phosphate groups	(22, 42)
Thr67	Walker A: hydrogen bonds with phosphate groups and Mg^{++} ion	(22, 42)
Lys121	Corresponds to Lys109 in RFC subunits	(16)
Asp126	Walker B: co-ordinates with ATP-associated Mg++	(22, 42)
Glu127	Catalytic base	(17)
Lys141	Forms a hydrogen bond in trans with the Walker B aspartate	(24)
Thr157	Required for ATP hydrolysis	(25)
Thr165	May facilitate ATP hydrolysis in trans	(44)
Arg169	Arginine finger that senses bound ATP and facilitates ATP hydrolysis	(23, 43)

* Eukaryotic residues correspond to yeast RfcB and are color coded as for the categories in Fig. S6; bacterial clamp loader subunits correspond to the *E. coli* γ subunit and are color coded as for the categories in Fig. S8.

Table S3. Hyperpartition for Helicases.

Category	Subgroup:
+ + - - - - o o o o o o o +	PcrA
+ + - - - - o o o o o o o -	*MiscSF1*
+ - + - - - o o o o - - + o	eIF4AIII
+ - + - - - o o o o - + - o	Uap56
+ - + - - - o o o o - - - o	*MiscDEAD-box*
+ - - + - - o o o o + o o o	Snf2
+ - - + - - o o o o - o o o	*MiscSwiSnf*
+ - - - + - - - - + o o o o	HelS
+ - - - + - - - + - o o o o	RecG
+ - - - + - - + - - o o o o	HsdR
+ - - - + - + - - - o o o o	Prp16
+ - - - + - - - - - o o o o	*MiscDEAH*
+ - - - - + o o o o o o o o	OtherHelicases
+ - o o o o o o o o o o o o	*MiscHelicases*
- o o o o o o o o o o o o o	Random

Table S4. Functional/structural relevance of pattern residues within SF1 Helicases.

A. Superfamily 1 Helicases		
Residue*	**Functional/structural relevance**	**Reference**
Gly36	Walker A glycine	(28)
Lys37	Walker A: ATP-binding site	(28, 45)
Thr38	Walker A: ATP-binding site	(28, 45)
Phe64	Single-stranded DNA binding site	(28)
Asp223	Walker B: Mg++ ion binding	(28, 45)
Glu224	Walker B: catalytic base	(28, 45)
Asp227	Links ATP-binding and DNA-binding sites	(28)
Gln254	Directly contacts the γ phosphate of ATP	(28)
Tyr257	Single-stranded DNA binding site	(28)
Trp259	Single-stranded DNA binding site	(28)
Arg260	Single-stranded DNA binding site	(28)
Arg287	Directly contacts the γ phosphate of ATP	(28, 45)
Lys568	Links ATP-binding and DNA-binding sites	(28, 45)
Gly569	Links ATP-binding and DNA-binding sites; backbone H-bond to Arg287	(28)
Glu571	Forms a hydrogen bond with ribose moiety of ATP	(28)
Arg610	Directly contacts the γ phosphate of ATP	(28, 45)
B. Superfamily 2 Helicases		
Residue*	**Functional/structural relevance**	**Reference**
Gly87	Walker A glycine	(27)
Lys88	Walker A: ATP-binding site	(27, 46)
Thr89	Walker A: ATP-binding site	(27, 46)
Arg116	Forms H-bonds with RNA	(27)
Glu117	Positions two RNA-binding residues, Arg116 and Arg316	(27)
Thr163	Forms H-bond with RNA	(27, 46)
Arg166	Forms H-bonds with RNA	(27, 46)
Asp187	Walker B: Mg++ ion binding	(27, 46)
Glu188	Walker B: catalytic base	(27, 46)
Asp190	Links ATP hydrolysis to RNA unwinding	(27, 46)
Phe197	Contacts RNA substrate	(27)
Ser218	Forms H-bond with Asp190	(27, 46)
Ala219	Backbone forms an H-bond to γ phosphate-associated water molecule	(27, 46)
Thr220	Forms H-bond with Asp190	(27, 46)
Arg339	Was proposed to help link ATP-hydrolysis to RNA unwinding	(27, 46)
Gly340	Contacts the γ phosphate of ATP.	(27, 46)
Asp342	Forms a hydrogen bond with ribose moiety of ATP	(27, 46)
His363	Forms an H-bond to γ phosphate-associated water molecule	(27, 46)
Arg367	Forms an H-bond to γ phosphate-associated water molecule	(27, 46)
Arg370	Helps position the γ phosphate binding region	(27, 46)

*SF 1 residues correspond to PcrA DNA helicase and are color coded as for the categories in Fig. S5-3; SF-2 residues correspond to eIF4AIII RNA helicase and are color coded as for the categories in Fig. S5-2.

Table S5. Hyperpartition for protein kinases.

Category	Subgroup:
+ + - - - - - + - - - - - -	AGC
+ + - - - - - + - - - - - -	CMGC
+ + - - - - + - - - - - - -	CAMK
+ + - - - + - - - - - - - -	TK
+ + - - + - - - - - - - - -	STE
+ + + - - - - - - - - - - -	CK1
+ + - - - - - - - - - - - -	*MiscEPKs*
+ - o o o o o o + - - - - -	ELK1-CAK
+ - o o o o o o - + - - - -	ELK2-Rio
+ o o o o o o o - - + - - -	ELK3-KdoK
+ - o o o o o o - - - - - -	*MiscELKs*
+ - o o o o o o - - - - + -	APK1-PI3K
+ - o o o o o o - - - - - +	APK2-ubiB
+ - o o o o o o o o o o - -	*MiscAPKs*
- o o o o o o o o o o o o o	Random

Table S6. Functional/structural relevance of pattern residues within protein kinases.

Residue	Functional/structural relevance	Reference
Gly52	Coordinates the γ-phosphate of ATP for hydrolysis	(47)
Lys72	anchors α- and β-phosphates of ATP	(48)
Glu91	Forms a salt bridge with Lys72 and functions as a regulatory switch	(49)
Lys92	Anchors the C-terminal carboxyl group	(48)
Phe102	Hydrophobic anchor to large lobe	(48)
Phe145	Functionally interacts with a PxxP motif that is implicated in localization and processing by phosphorylation of AGC kinases; see (50).	(34, 50)
Tyr146	Functionally interacts with the AGC kinase PxxP motif	(50)
Tyr164	Integrates substrate and ATP binding regions (often a histidine)	(51)
Arg165	Bridge to activation and Mg++ positioning loops and to Asp220	(48)
Asp166	The catalytic base	(48)
Lys168	Bridge to γ-phosphate of ATP	(48)
Asn171	Coordinates secondary Mg++ ion; involved in phosphoryl transfer	(30)
Tyr179	Interacts with Phe145 (often a histidine)	(34)
Asp184	Coordinates primary Mg++ ion	(30, 48)
Phe185	Shielding active site from solvent	(48)
Asp220	Bridge to active site	(30, 48)
Trp221	Anchor to core	(48)
Trp222	Proposed role in communicating ATP binding from the active site to a distal peptide-binding ledge	(52)

* Residues correspond to PKA_AGC and are color coded to correspond to the categories in Fig. S13.

Table S7. A hyperpartition generated automatically by the mc-BPPS sampler given as input an alignment of 78,143 Rossmann fold protein sequences. The numbers in parentheses at the end of each row correspond to the number of sequences assigned to each subgroup by the sampler.

	Category											subgroup
Set:	1	2	3	4	5	6	7	8	9	10	11	
1:	+	-	-	-	-	-	-	-	-	-	-	MainSet1 (51853)
2:	+	+	-	-	-	-	-	-	-	-	-	Set1_0 (855)
3:	+	-	+	-	-	-	-	-	-	-	-	Set1_1 (660)
4:	+	-	-	+	-	-	-	-	-	-	-	Set1_2 (1735)
5:	+	-	-	-	+	-	-	-	-	-	-	Set1_3 (1913)
6:	+	-	-	-	-	+	-	-	-	-	-	Set1_4 (2669)
7:	+	-	-	-	-	-	+	-	-	-	-	Set1_5 (731)
8:	+	-	-	-	-	-	-	+	-	-	-	Set1_6 (761)
9:	+	-	-	-	-	-	-	-	+	-	-	Set1_7 (266)
10:	+	-	-	-	-	-	-	-	-	+	-	Set1_8 (4695)
11:	+	-	-	-	-	-	-	-	-	-	+	Set1_9 (145)
12:	-	o	o	o	o	o	o	o	o	o	o	Rejected (11860)

Figure S16. Output contrast alignments generated by a simple automated version of the mc-BPPS sampler given as input an alignment of 78,143 Rossmann fold protein sequences.

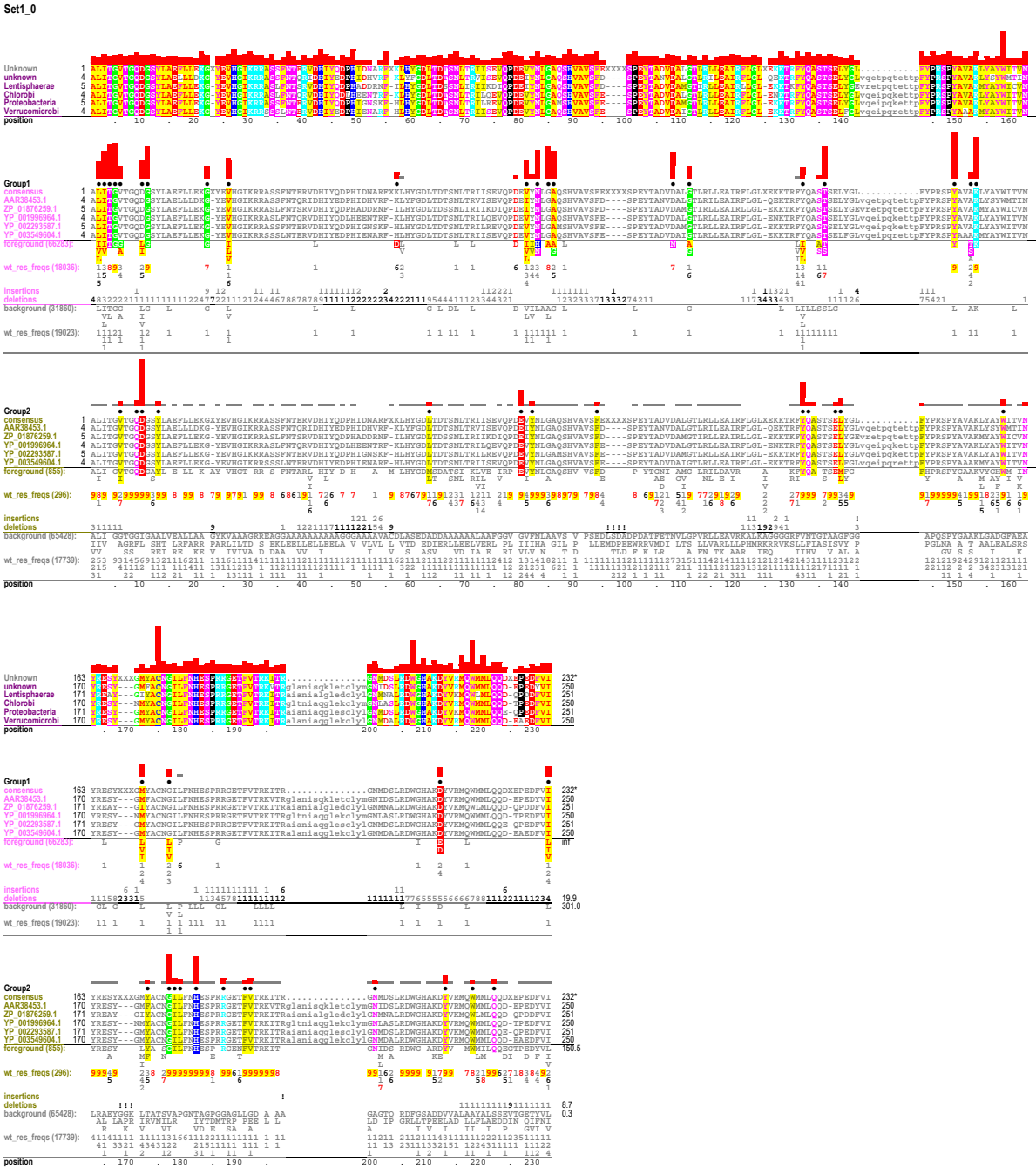


Figure S16. (continued).

Set1_1

Figure S16. (continued).

Set1_2

Figure S16. (continued).

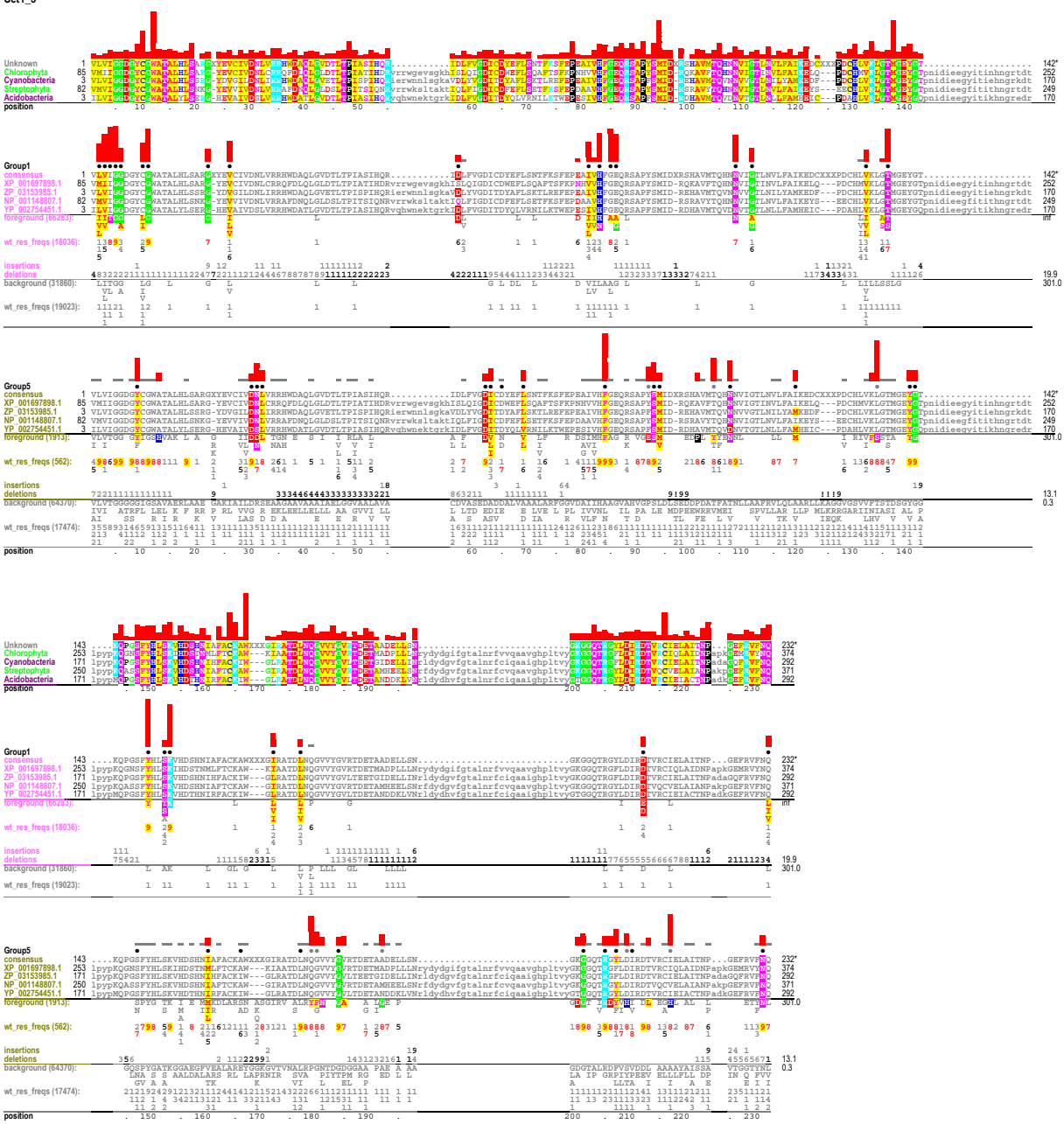


Figure S16. (continued).

Figure S16. (continued).

Set1_5

Figure S16. (continued).

Set1_6

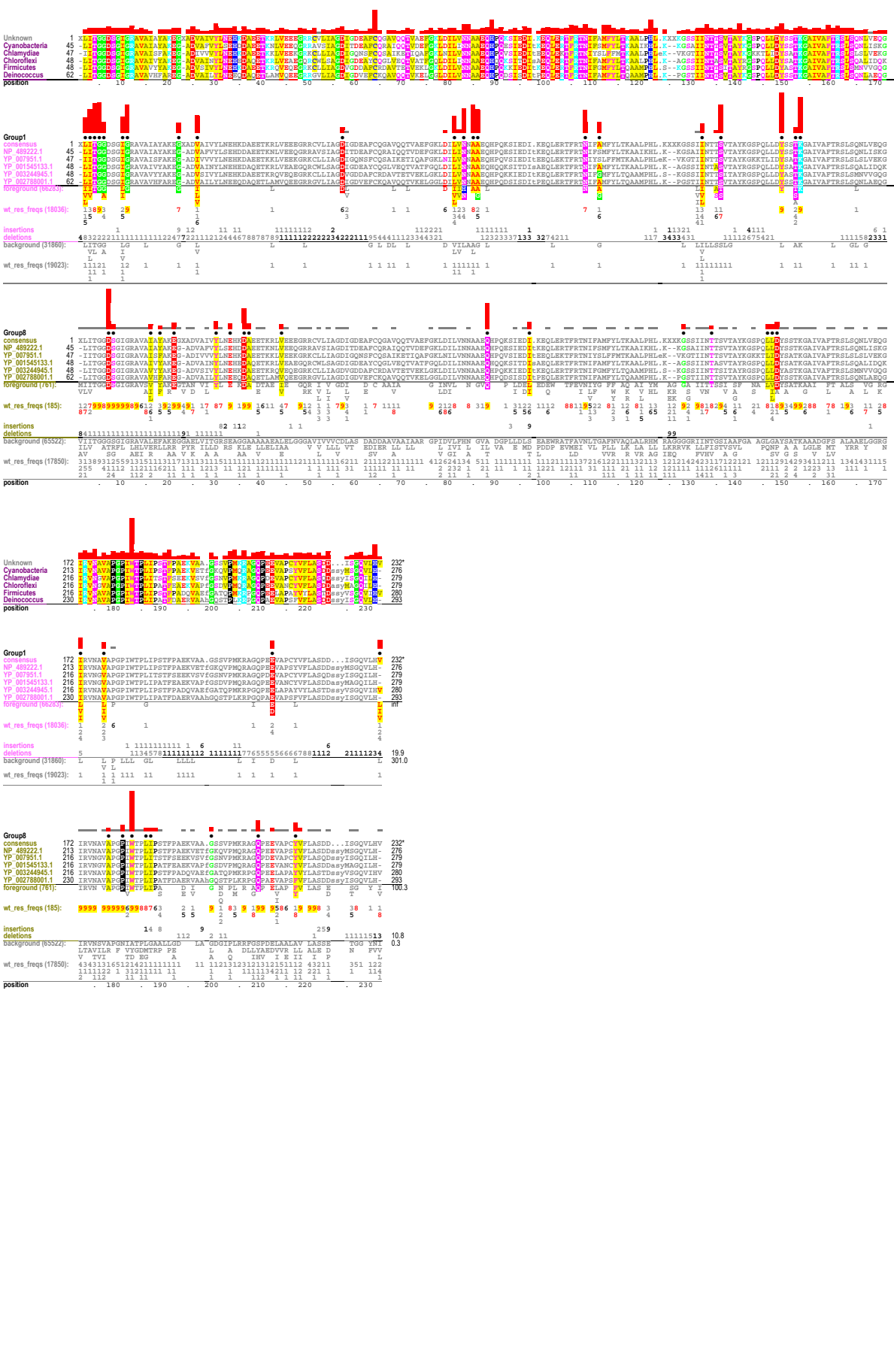


Figure S16. (continued).

Set1_7

Figure S16. (continued).

Set1_8

Figure S16. (continued).