Integrated structural and evolutionary analysis reveals common mechanisms underlying adaptive evolution in mammals

Greg Slodkowicz[a,1,2] and Nick Goldman[a,2]

[a]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD Hinxton, United Kingdom

Edited by Eugene V. Koonin, National Institutes of Health, Bethesda, MD, and approved February 4, 2020 (received for review September 30, 2019)

Understanding the molecular basis of adaptation to the environment is a central question in evolutionary biology, yet linking detected signatures of positive selection to molecular mechanisms remains challenging. Here we demonstrate that combining sequence-based phylogenetic methods with structural information assists in making such mechanistic interpretations on a genomic scale. Our integrative analysis shows that positively selected sites tend to colocalize on protein structures and that positively selected clusters are found in functionally important regions of proteins, indicating that positive selection can contravene the well-known principle of evolutionary conservation of functionally important regions. This unexpected finding, along with our discovery that positive selection acts on structural clusters, opens previously unexplored strategies for the development of better models of protein evolution. Remarkably, proteins where we detect the strongest evidence of clustering belong to just two functional groups: Components of immune response and metabolic enzymes. This gives a coherent picture of pathogens and xenobiotics as important drivers of adaptive evolution of mammals.

protein evolution | mammals | adaptive evolution | metabolism | immunity

Over the course of evolution, the genomes of all organisms are shaped by the environment. The results of this process can be observed by comparing evolutionarily related sequences from different species: Regions that code for essential cellular functions can remain unaltered over hundreds of millions of years, while changing evolutionary pressures can lead to emergence of new functions over very short evolutionary timescales. As a result, evolutionary histories of sites in the genome hold information about their functional importance. Functionally important regions are routinely identified by taking advantage of the fact that they are highly conserved in evolution (1, 2). Similarly, methods for detecting regions harboring adaptive changes in protein-coding regions have been developed to take advantage of the fact that rapid fixation of new alleles is a hallmark of positive selection (3, 4). Analyses of patterns of evolutionary change can identify specific cases of adaptation as well as reveal general principles that guide evolution (5). Understanding evolutionary processes and distinguishing between neutral and adaptive changes is therefore one of the key aims of modern evolutionary studies.

As most proteins have to maintain a specific three-dimensional (3D) shape to perform their function, protein-coding genes exhibit particularly complex patterns of substitution. Biophysical constraints restrict the allowed amino acid substitutions and result in dependencies across the entire protein sequence. While structural features can explain a significant proportion of observed site-to-site rate variation (6), previous studies have focused on evolutionary scenarios where existing functions are maintained and little is known about the structural properties of sites evolving under positive selection.

Present lack of understanding of structural aspects of adaptive evolution is particularly surprising bearing in mind that many single-gene studies took advantage of protein structure to assess the functional significance of positively selected sites identified from sequence data. In the classic study of Hughes and Nei (7), positively selected residues in the MHC molecule were found to cluster in the groove where pathogen-derived peptides are bound, supporting the hypothesis that rapid amino acid substitutions at these sites tuned the ability to bind peptides derived from pathogens. Similarly, positively selected sites in TRIM5α, a viral restriction factor that can inhibit the cellular entry of HIV in nonhuman primates, are placed in the region that mediates binding to the virus (8). In these studies, as in others (e.g., ref. 9), proximity of positively selected residues on the protein structure was used as corroborating evidence and helped assign a molecular mechanism underlying detected adaptations.

As the amount of available genomic data increased, studies of positive selection in individual proteins were followed by genome-wide positive selection scans (10–15). Such genomic scans, using appropriately adapted statistical methodology (16, 17), can identify which cellular processes are primary targets of positive selection and generate testable hypotheses. However, structural aspects of identified examples were largely neglected

Significance

Phenotypic evolution is driven primarily by natural selection but the majority of differences between the genomes of related species are thought to be neutral. Because of this, linking differences in phenotype to underlying genetic changes is challenging. Here, we applied evolutionary sequence analysis methods to comprehensively identify sites that evolved under positive selection in mammals and used available protein structures to link them to molecular mechanisms. This allowed us to detect clusters of positively selected sites in proteins involved in immunity, as well as enzymes that detoxify foreign chemicals, and demonstrate that those sites tend to localize to functionally important regions. Our findings suggest that, in addition to functional similarities, there are common structural features and mechanisms underpinning adaptive evolution.

Author contributions: G.S. and N.G. designed research; G.S. performed research; G.S. and N.G. contributed analytic tools; G.S. analyzed data; and G.S. and N.G. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Data deposition: All results discussed in the paper (structure-mapped values of selective constraint, underlying alignments, and phylogenetic trees) are available at https://www.ebi.ac.uk/goldman-srv/sips/.

[1]Present address: Division of Structural Studies, Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 0QH, United Kingdom.

[2]To whom correspondence may be addressed. Email: gslodko@mrc-lmb.cam.ac.uk or goldman@ebi.ac.uk.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1916786117/-/DCSupplemental.

First published March 2, 2020.

and so no coherent view of how protein structure affects adaptive evolution has emerged from these investigations.

This is a significant gap in our understanding of evolution. Biophysical constraints restrict what substitutions are allowed for protein function to be maintained and are also likely to limit the emergence of adaptive changes in response to pressures from the environment, yet no evolutionary theory predicts the structural properties of sites harboring adaptive changes. It is not established whether positive selection is more likely to act on protein sites where the effect of mutations is the largest (e.g., enzyme catalytic sites or key interaction interfaces) or regions where mutations likely have a smaller effect (e.g., allosteric regulation sites). Adaptive changes are associated with rapid fixation of advantageous mutations, yet functional regions are thought to be highly conserved in evolution. Contrasting these two principles leads to an apparent paradox.

Here, we integrated structural information into evolutionary analyses in order to study the properties of positively selected sites. We demonstrate that detailed mechanistic interpretation of findings can be achieved on a genome-wide level, just as in the case of earlier studies of individual proteins. In recent years, it has become apparent that structural data can be an orthogonal source of information that can serve to validate and augment findings in different areas of genomics (18). Structural placement of sites of interest, such as those identified through genome-wide sequence analyses, can be used to strengthen the confidence in findings: Clustering of sites indicates concerted function whereas unrelated sites are expected to be more uniformly distributed in the structure. Recently developed methods based on clustering of sites on protein structures have been successful in distinguishing causal and hitchhiking mutations underlying genetic diseases (19) and for identifying mutations with a functional impact in cancer (20–23). Detailed information about the protein structure can similarly aid understanding of molecular mechanisms underlying adaptation at detected sites.

To obtain a structurally informed view of positive selection at the residue level, we developed an approach combining a genome-wide scan for positive selection with structural information (Fig. 1*A*). We applied 3D clustering to detect genes with positively selected sites in a robust manner that additionally allowed us to link identified cases to an underlying molecular mechanism. We demonstrate that positively selected sites tend to occur close to one another on protein structure and detect 20 high-confidence positively selected clusters (Table 1). Strikingly, we found that all but one of the identified cases are immune-related proteins or metabolic enzymes. In both of these functional

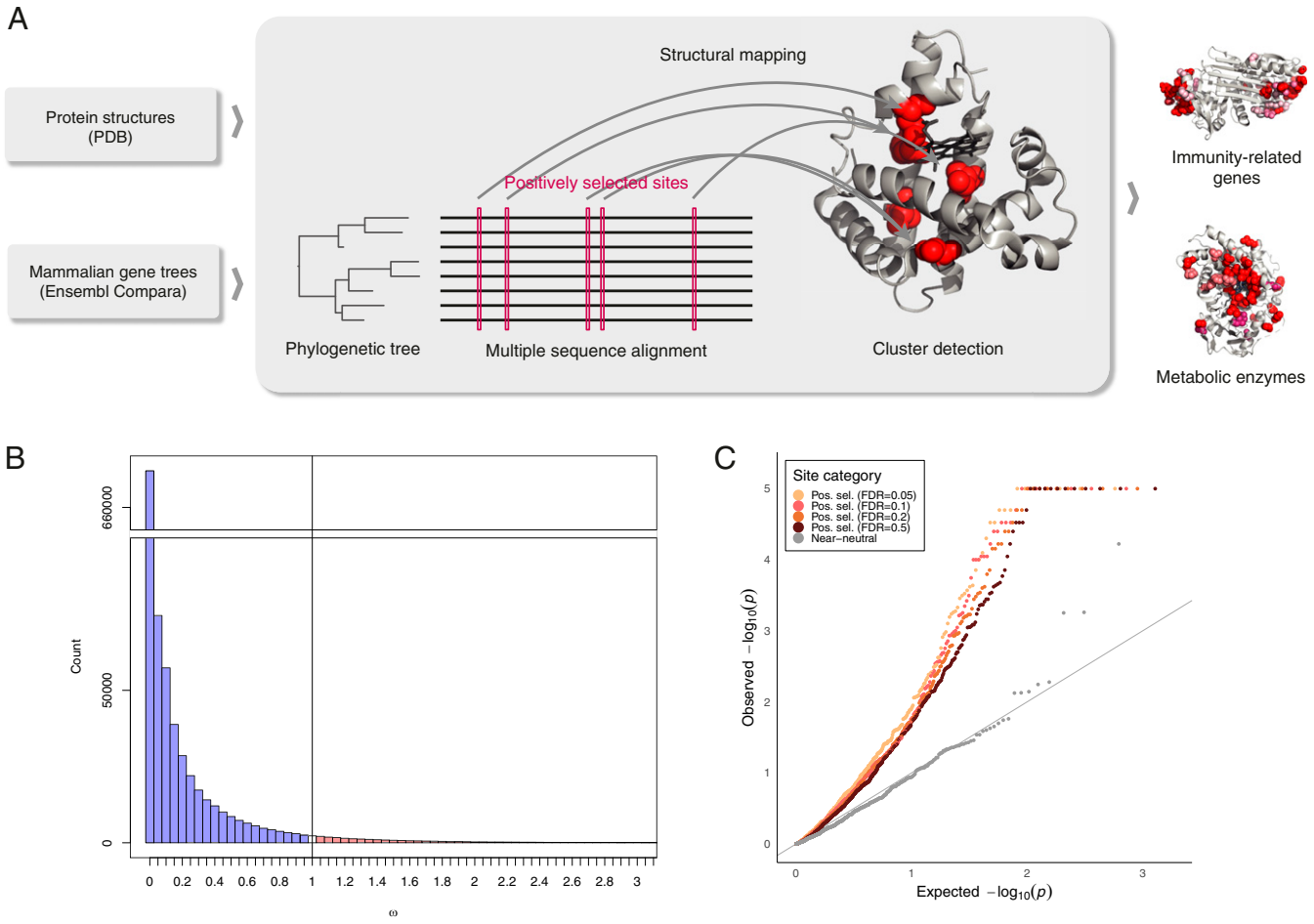

Fig. 1. Positively selected residues tend to cluster together. (*A*) Overview of the approach. (*B*) Distribution of values of selective constraint in the dataset. With 97.6% of sites having $\omega < 1$ (indicating purifying selection), and 2.4% with $\omega \geq 1$, the mean of ω across the entire dataset is 0.126. (*C*) QQ plot of *P* value distribution obtained from CLUMPS applied to positively selected sites at FDR of 0.05, 0.1, 0.2, and 0.5. If the residues under positive selection were randomly distributed on protein structures, we would expect a uniform distribution of *P* values (gray line). The observed *P* values for positively selected sites are lower than would be expected under the null hypothesis of random placement, indicating that positively selected sites tend to cluster together. In contrast, near-neutrally evolving sites (gray points) do not show a tendency to cluster.

Table 1. Proteins with clusters of positively selected sites

Gene symbol	Gene name	Protein length	PDB ID code	PDB sites	Substrate/relevant ligand in PDB	Number of possible selected sites*
Immune-related proteins						
HLA-DRBI	Major histocompatibility complex, class II, DR β1	266	1AQD	187	Endogenous peptide	16/19/21/25
FCN2	Ficolin 2	313	2J3F	217	*N*-acetyl-D-galactosamine	6/9/10/13
SERPINB3	Serpin B3	390	4ZK0	367	—	24/26/31/42
TLR4	Toll-like receptor 4	839	4G8A	601	LPS, LP4	51/61/80/119
CD1A	CD1a molecule	327	1ONQ	271	Sulfatide self-antigen	40/44/50/64
C5	Complement component 5	1676	3CU7	1625	—	28/37/53/89
C8A	Complement component 8α	584	3OJY	478	—	16/20/27/34
SIGLEC5	Sialic acid binding Ig-like lectin 5	551	2ZG1	208	Sialic acid	24/32/38/43
TFRC	Transferrin receptor 1	760	3S9L	638	—	17/19/24/40
Metabolic enzymes						
CYP2C9	Cytochrome P450, family 2, member C9	490	1R9O	453	Flurbiprofen	29/30/35/40
CYP2D6	Cytochrome P450, family 2, member D6	497	2F9Q	454	—	6/8/10/14
CYP3A4	Cytochrome P450, family 3, member A4	503	3TJS	449	Desthiazolylmethyl oxycarbonyl ritonavir	22/27/33/41
AKR1B10	Aldo-keto reductase family 1, member B10	316	1ZUA	316	Tolrestat	13/18/19/23
AKR1C4	Aldo-keto reductase family 1, member C4	323	2FVL	323	—	19/21/23/26
SULT2A1	Sulfotransferase family 2A member 1	285	3F3Y	282	Lithocholic acid	15/18/22/32
CES1	Carboxylesterase 1	568	1MX1	532	Tacrine	12/16/19/31
GSTA3	Glutathione S-transferase α3	222	1TDI	218	Glutathione	11/13/16/20
OLAH	Oleoyl-ACP hydrolase	318	4XJV	216	—	8/12/18/24
AK5	Adenylate kinase 5	562	2BWJ	195	AMP	3/3/3/3
NCSTN	Nicastrin	709	5A63	665	Phosphocholine	7/8/11/18

Summary of genes where positively selected clusters were detected. Protein length refers to human orthologs.
*The number of positively selected sites is given at FDR thresholds of 0.05, 0.1, 0.2 and 0.5, respectively.

categories, interactions with dynamic environmental parameters appear to have shaped the evolutionary histories of the genes involved. By further analyzing the placement of positively selected clusters, we found that pervasive positive selection acts on regions that are typically highly conserved in evolution, suggesting strategies for the development of more accurate models of protein evolution and methods for detecting positive selection.

Results

Identification of Positive Selection. Neutral theory (24, 25) predicts that if mutations that arise at a locus are deleterious, they will undergo purifying selection and will be purged from a population, resulting in a low observed evolutionary rate. Conversely, if mutations result in beneficial changes, they will be rapidly driven to fixation. The ratio of fixation probabilities of nonsynonymous and synonymous substitutions (ω, or dN/dS) can thus be used to directly estimate the selective constraint acting on the protein level: $\omega \approx 1$ indicates neutral evolution; $\omega < 1$ purifying selection; and $\omega > 1$ positive selection (26). In order to identify residues that were under positive selection in mammalian evolution, we estimated sitewise values of selective constraint in the mammalian proteome. To this end, we first obtained coding sequences for 39 eutherian mammals (*SI Appendix*, Fig. S1) from Ensembl and phylogenetic trees from the Ensembl Compara database (27). We then aligned coding sequences corresponding to each tree using the PRANK aligner (28) and used the SLR software (29) to detect positively selected sites. Three-dimensional structures corresponding to human proteins in our dataset were obtained from the Protein Databank (PDB) (30) and we then used the SIFTS resource (31) to map positively selected sites onto protein structures.

The resulting dataset comprises 3,347 protein alignments and covers 1,021,133 structure-mapped amino acid sites. While the majority of sites evolve under purifying selection (Fig. 1*B*), consistent with both theoretical expectations and previous empirical estimates (13), we identified 4,498 sites with strong evidence of positive selection (false-discovery rate [FDR] = 0.05). We have made these results available as an online resource which allows for displaying and downloading of the structure-mapped sitewise estimates of selective constraint, as well as the underlying alignments and phylogenetic trees (https://www.ebi.ac.uk/goldman-srv/sips/).

Detecting clustering of positively selected sites. To determine the degree of clustering of positively selected sites, we applied a modification of the CLUMPS algorithm (21) to our integrated dataset (*Methods*). As the power to detect clustering is limited if very few residues are considered, it is desirable to include as many sites with evidence of positive selection as possible. At the same time, reducing the stringency in the detection of selection by allowing a higher FDR can dilute the signal of clustering by including more false positives. As it is not clear a priori what the tradeoff between these phenomena is and at what threshold the power to detect clustering is maximized, we applied the chosen clustering detection method separately to positively selected sites detected at different stringency levels. In order to determine the degree to which positively selected residues form clusters on protein structures, we inspected the overall distribution of *P* values obtained for each protein from CLUMPS at four FDR

thresholds at which positively selected sites were detected (Fig. 1*C*). We find a significant tendency for positively selected sites to cluster together and this trend is maintained at each FDR threshold, indicating that our findings are robust to how stringently positively selected sites are identified. While structural properties of residues evolving under positive selection are underexplored, previous work suggests that neutrally evolving residues may cluster on protein structures (32). To test this in our dataset, we performed the same clustering analysis on near-neutrally evolving sites but found no overall trend of clustering (Fig. 1*C*), and only one statistically significant case of clustering.

Clusters of positively selected sites. Having established that positively selected sites tend to occur close to one another on protein structures, we went on to select cases where evidence for clustering is the strongest. Depending on the FDR threshold used to identify sites as positively selected, between 35 and 52 proteins with clusters of positively selected residues were detected (FDR of clustering < 0.05), with substantial overlap between clusters detected at different thresholds (*SI Appendix*, Fig. S2). For 22 proteins, clusters were identified at all four FDR thresholds, suggesting that these constitute the most robust findings. For these proteins, we inspected the underlying alignments from which positively selected sites were identified. Correlation on the sequence level can introduce clusters on the level of structure and for this reason it is important to distinguish 3D clusters resulting purely from closeness of sites of interest in the sequence. In all but two cases, we found that positively selected sites are identified in regions of good alignment quality and that clusters of positively selected sites arise mostly from residues that are not adjacent in the sequence and become close to each other only once the protein is folded into its native conformation. The two cases where detected signature of positive selection appears to result from a stretch of contiguous residues in a region of poor alignment quality were rejected from further analysis. The remaining 20 proteins are summarized in Table 1. Remarkably, 9 of them are immune-related proteins and 10 are metabolic enzymes. The remaining protein, nicastrin, is the substrate-recruiting component of γ-secretase (33), a protein complex with catalytic activity, and we therefore considered it together with other enzymes.

To assess the impact of possible errors in the gene tree topologies on detecting positive selection, we generated 100 alternative tree topologies for each gene of interest and repeated the positive selection analysis. We found that most (84.7%) of the detected sites are supported by at least 95% of alternative topology sets, indicating that our results are not sensitive to possible small errors in the phylogeny (*SI Appendix*, Fig. S3).

Positive Selection in Proteins Involved in Immunity.

Confirmation of validity of clustering approach. Rapid evolutionary rates in genes involved in both adaptive and innate branches of the immune system are a classic example of positive selection (7, 8, 34–36). Proteins where we identified positively selected clusters (Table 1) include cases where positive selection has been documented previously, such as in HLA-DRB1 (7), CD1a (37), Toll-like receptor 4 (TLR4) (37), and transferrin receptor 1 (TfR1), a protein which is known to have been hijacked by arenaviruses for facilitating cellular entry (38). Positively selected residues are located primarily in regions involved in antigen binding, such as the structurally similar binding clefts of HLA-DRB1 and CD1a (*SI Appendix*, Figs. S4–S7). While these findings were reported previously, they give confidence in the approach we applied here.

Findings of selection clusters. We also identify cases where to our knowledge positive selection has not been previously described: Ficolin 2 (*SI Appendix*, Fig. S8) [although positive selection in the related ficolin 3 has been reported (39)], complement component 5 (*SI Appendix*, Fig. S9), complement component 8α (*SI Appendix*, Fig. S10), Siglec-5 (*SI Appendix*, Fig. S11), and serpin B3 (Fig. 2).

The placement of positively selected sites in serpin B3 is particularly interesting as this protein exhibits two clusters concentrated on the opposite poles of the protein (Fig. 2*A*). Serpin B3 belongs to the serpin superfamily of protease inhibitors, although unlike most serpins it binds cysteine rather than serine proteases. Serpins contribute to immunity by inhibiting proteases secreted by bacteria. Serpin B3 inactivates leaked lysosomal cathepsins, inactivates pathogen-derived cathepsins, and is also thought to be involved in autoimmunity (42). Comparison with other available structures of serpins reveals a remarkable correspondence of these positively selected sites to the protease binding sites before and after the conformational change that characterizes the mode of action of serpins (Fig. 2*B*). Furthermore, there is previous evidence that serpin B3 homologs have changed their substrate specificities over the course of evolution, consistent with the action of positive selection (43, 44). The presence of two positively selected residue clusters at opposite poles of the protein implies that both regions participate in the tuning of function. The importance of these regions in the proteolytic function of serpins demonstrates that the positive selection we detected is likely to have functional consequences.

Interactions with pathogens are known to be one of the dominant pressures shaping mammalian evolution (34). Our analysis adds mechanistic details to these findings: Positively selected clusters in proteins involved in host–pathogen interactions are placed in regions directly mediating binding of pathogen-derived molecules. Binding of pathogen-derived peptides by HLA and subsequent triggering of the immune response is a classic example of this (45). Here we have identified further examples of similar mechanisms in components of both innate and adaptive branches of the immune system. Interestingly, these include not only proteins or protein-derived peptides, but also lipids (CD1a) and lipopolysaccharides (TLR4). This is true both when binding is facilitating the neutralization of pathogens and, as in the case of TfR1, where host proteins are hijacked by a pathogen to facilitate cellular entry. These scenarios are examples of high evolutionary rate being the result of an "arms race" between host and pathogen. Such dynamics are predicted by the Red Queen hypothesis, which posits that evolution is driven by interspecies competition (46).

Positive Selection Acting on Metabolic Enzymes.

Cytochrome P450s. Ten of the 11 remaining positively selected clusters are found in enzymes. Three of the identified clusters of positively selected sites are in members of the cytochrome P450 (CYP) superfamily (Fig. 3). CYPs are the most important drug-metabolising enzyme class, contributing to the metabolism of 90% of drugs as well as many other xenobiotics, such as pollutants. These liver enzymes catalyze monooxygenation reactions on a wide range of small and large substrates. More than 50 CYPs have been identified in the human genome but relatively few are known to have a role in drug metabolism (47).

Strikingly, all three of the CYPs where we identified positively selected clusters of residues are known to be important for drug metabolism: CYP3A4 (Fig. 3 *A* and *B*) is the most promiscuous of all CYPs, contributing to the metabolism of ∼50% of marketed drugs, and CYP2C9 (Fig. 3 *C* and *D*) and CYP2D6 (Fig. 3 *E* and *F*) are also among the six principal CYPs thought to contribute the most to drug metabolism (48). In our dataset, alignments containing the three CYPs mentioned before also contain two further cytochrome P450 paralogs that are important for drug metabolism; in total five of six enzymes thought to be responsible for the majority of cytochrome P450 drug metabolism show evidence of positive selection.

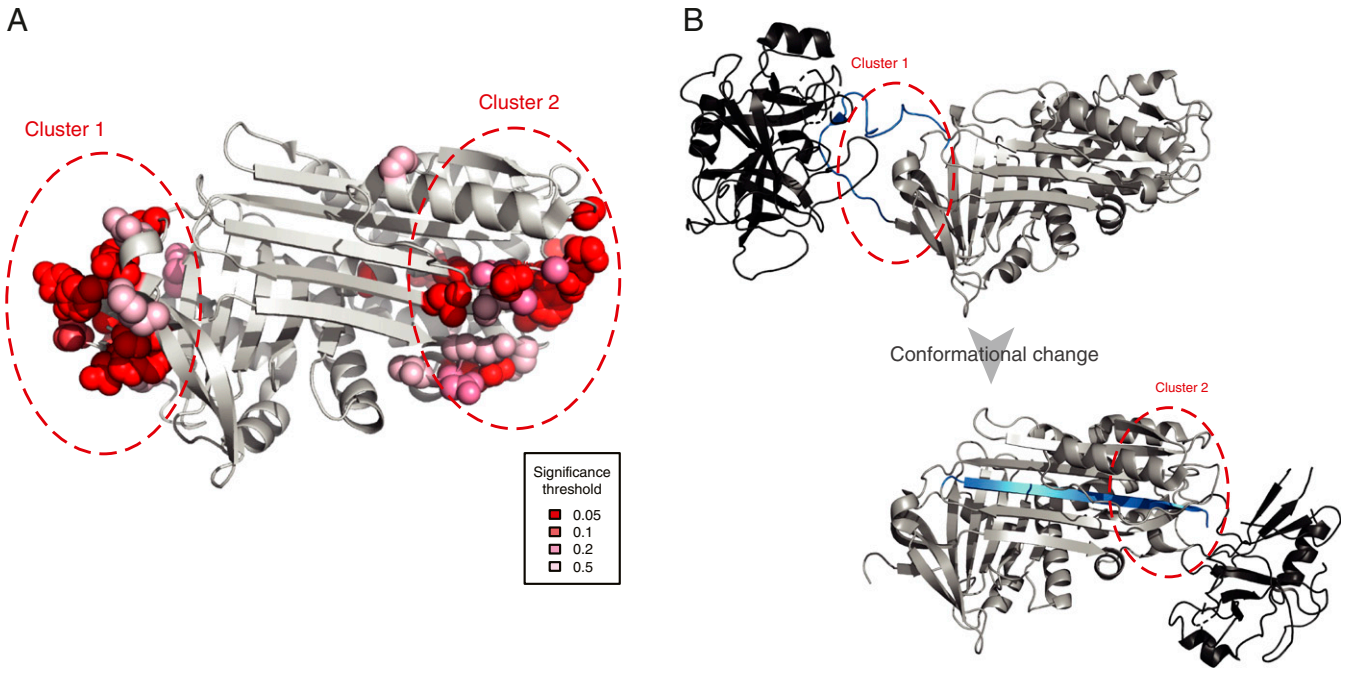

Fig. 2. Clusters of positively selected sites in serpin B3. (*A*) Placement of positively selected sites on the structure of serpin B3 (PDB ID code 4ZK0). (*B*) Mode of action of serpins shown using PDB structures 1K9O (*Upper*) and 1EZX (*Lower*) with the substrate shown in black and the reactive center loop marked in blue. Regions analogous to those where positively selected clusters were detected are marked as in *A*. Serpins function by binding their target proteases using a reactive center loop that mimics the protease substrate. They then form a covalent bond with the protease and undergo a large conformational change resulting in the protease being deformed and then acylated (40, 41). We find that positively selected residues surround the reactive center loop and are also located on the opposite side of the protein to which the bound protease is dragged.

Aldo-keto reductases. We identified positively selected clusters in two members of the 15 aldo-keto reductases (AKRs) present in human. Similar to CYPs, AKRs are a family of highly promiscuous enzymes that utilize NAD(P)(H) cofactors and can reduce a wide range of substrates (49). AKRs are part of phase II metabolism and can transform or detoxify both endogenous and environmental aldehydes and ketones (50–52). Positively selected residues in both AKRs cluster around the region where the substrate binds but not around the $NADP^+$ cofactor (Fig. 4). This suggests that evolution has tuned substrate specificity while maintaining binding to the cofactor.

Other enzymes. We also identified individual positively selected clusters in the members of three other protein families involved in detoxification: glutathione S-transferase α3 (GSTA3) (53–56),

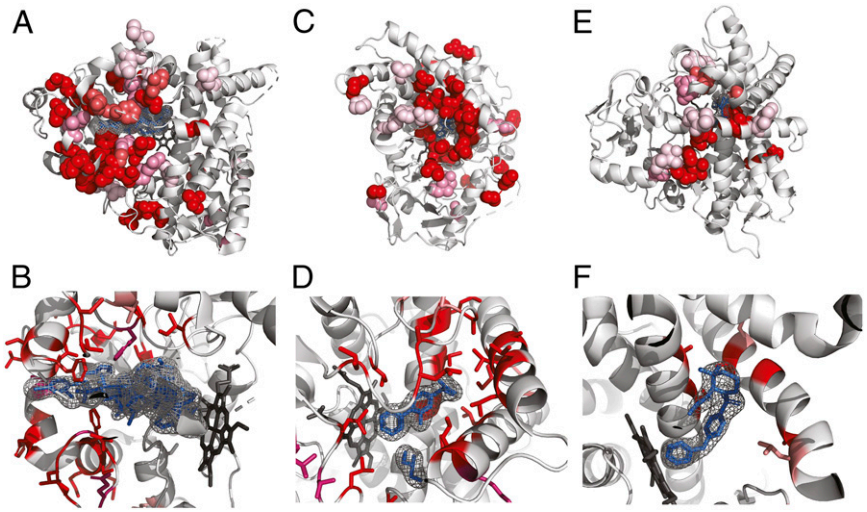

Fig. 3. Positively selected residues in CYPs cluster in the substrate entry channel and catalytic site. Positively selected residues: (*A* and *B*) CYP3A4 (PDB ID code 3TJS), (*C* and *D*) CYP2C9 (PDB ID code 1R9O), and (*E* and *F*) CYP2D6 (PDB ID code 2F9Q). Hemes are shown colored in dark gray, other ligands in blue. Additional ligands were transferred from other PDB structures by superimposition: (*A* and *B*) desthiazolylmethyloxycarbonyl ritonavir, ketoconazole (PDB ID code 2V0M), erythromycin (PDB ID code 2J0D), (*C* and *D*) flurbiprofen, (*E* and *F*) prinomastat (PDB ID code 3QM4). Specificity for the extraordinary diversity of substrates in this enzyme superfamily is facilitated by a large, flexible binding pocket at the bottom of which heme is located. In all three structures, the location of the positively selected residues tracks the binding of a ligand, and in general can be found on the sides of helices and in loops that form the binding pocket.

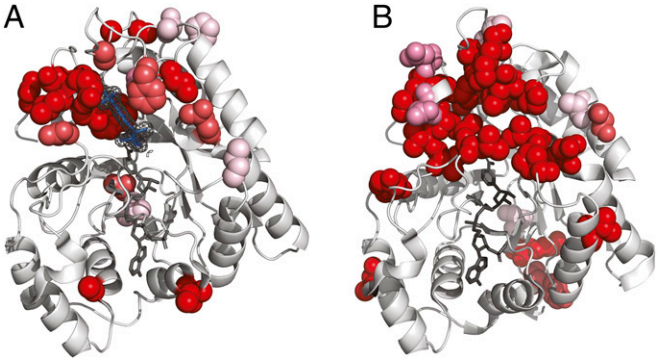

Fig. 4. Positively selected residues in AKRs surround the substrate binding site. Positively selected residues in (*A*) AKR1B10 (PDB ID code 1ZUA) and (*B*) AKR1C4 (PDB ID code 2FVL). Tolrestat marked in blue, $NADP^+$ marked in dark gray. Positively selected residues in AKR1B10 cluster around the bound ligand tolrestat, an inhibitor developed for diabetes treatment, but not around the $NADP^+$ cofactor. The structure of AKR1C4 has been solved without ligand but the positively selected residues cluster in a similar region of the structure when compared to AKR1B10. As in the case of AKR1B10, there are no positively selected residues in the neighborhood of the $NADP^+$ cofactor.

carboxylesterase 1 (57, 58), and sulfotransferase 2A1 (59, 60). In all cases, positively selected sites cluster around the active site of the enzyme where the substrate binds (Fig. 5 *A*–*C*).

In the remaining three cases, positively selected clusters are located in subdomains that interact with substrates. Adenylate kinase 5 (AK5) is a member of a family of enzymes important for maintaining the energetic balance in the cell by converting ADP into ATP (61). Positively selected residues in AK5 fall in the lid subdomain (Fig. 5*D*), which has been shown to have a role in tuning the enzyme activity (62). The three positively selected sites that constitute the positively selected cluster in AK5 flank a DD motif, which is highly conserved in AK5 and in other enzymes of the family. Experimentally mutating a residue homologous to V507, one of the sites we have predicted, has been shown to have an effect on the enzyme's kinetic parameters (62), strongly suggesting that the positively selected sites we detected contribute to enzyme specificity and kinetics.

In the case of oleoyl-ACP hydrolase (OLAH), an enzyme involved in controlling the distribution of chain lengths of fatty acids, positively selected residues are located in the capping domain that covers the substrate (Fig. 5*E*). Detailed mutational data for OLAH is lacking but enzymes of the same class have been shown to undergo changes of specificity in other species (63). Positively selected sites in nicastrin (*SI Appendix*, Fig. S12) are primarily located in the lid domain that covers the substrate (64), and changes at positively selected sites in these enzymes are therefore also consistent with positive selection acting to fine-tune enzymatic activity.

Although pervasive positive selection in metabolic enzymes, similar to that experienced by immune-related genes, may seem surprising, examples of episodic adaptation of enzymes in specific lineages exist, particularly in primates (56, 65–68). Interestingly, Monit et al. (69) recently conducted a detailed analysis of the evolutionary history of SAMHD1, a protein with both antiviral and enzymatic function. SAMHD1 is present in our dataset, although we used a different PDB structure (4MZ7) for structural mapping. In the region common to both structures, 9 of 15 sites identified by Monit et al. are also significant in our dataset (FDR = 0.05). The degree of clustering of positively selected sites in this protein does not meet all of our stringency criteria, but it is significant at two of four thresholds we considered. The function of SAMHD1 is inhibition of HIV-1 replication, and it appears that the adaptation in this protein was driven primarily by evolutionary conflict with this pathogen.

Eight of 10 enzymes where we identified positively selected sites are involved in the metabolism of xenobiotics. This interaction with the environment makes them plausible targets of positive selection. Much like parts of the immune system that directly interact with pathogens, these metabolic enzymes form an interface with the environment and act as one line of defense. The diversification of mammals involved adaptation to varied

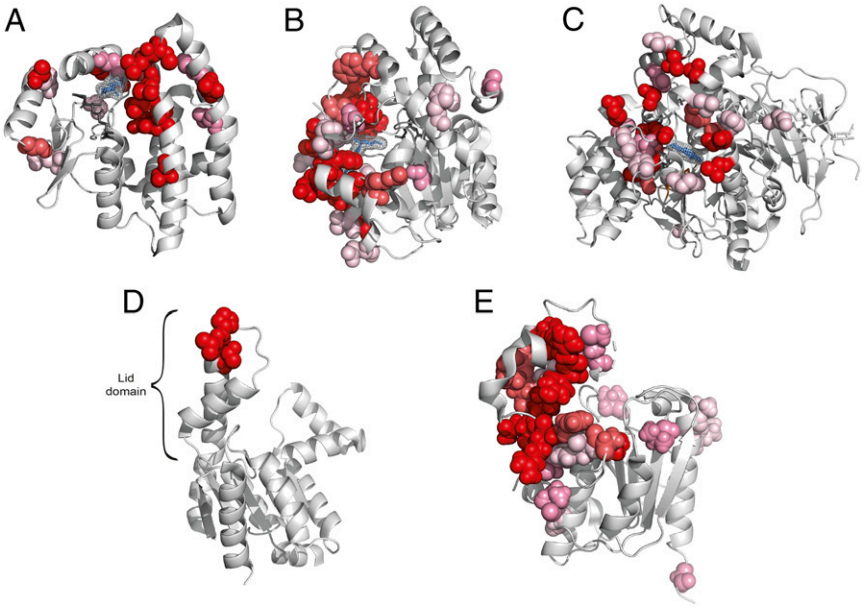

Fig. 5. Positively selected residues in other enzymes. (*A*) Positively selected sites in GSTA3 (PDB ID code 1TDI). Glutathione shown in dark gray, δ-4-androstene-3-17-dione (blue) transferred by structure superimposition from structure 2VCV. (*B*) Positively selected residues in sulfotransferase 2A1 (PDB ID code 3F3Y). Adenosine-3′-5′ diphosphate shown in dark gray, lithocholic acid shown in blue. (*C*) Positively selected sites in carboxylesterase 1 (PDB ID code 1MX1). Tacrine shown in blue. (*D*) Positively selected residues in AK5 (PDB ID code 2BWJ). (*E*) Positively selected sites in OLAH (PDB ID code 4XJV).

environments and new diets and as the environment in which they live and feed has changed, so did their exposure to toxins. This is likely to have required the widespread, repeated adaptive changes that we observed.

Placement of positively selected sites in relation to functional sites. Having observed the tendency of observed clusters to occur in the direct neighborhood of bound ligands, we sought to quantify this trend. For structures solved with exogenous ligands, we obtained the distribution of distances for positively selected residues and compared them to remaining residues (Fig. 6*A*). We found that positively selected residues are significantly closer to those ligands (mean distance 16.9 Å vs. 24.4 Å; $P < 2.2 \times 10^{-16}$; Kolmogorov–Smirnov test), confirming that positively selected clusters tend to occur closer to bound ligands than would be expected by chance and providing further evidence for positive selection acting to fine-tune ligand binding.

We then investigated the overall distribution of ω as a function of distance to catalytic sites, using annotations from the Catalytic Site Atlas (70). In proteins where we detected no evidence of positive selection, purifying selection is the strongest in the neighborhood of catalytic sites and gradually relaxes with distance from them (Fig. 6*B*). This trend is consistent with previous studies of selective constraint where positive selection was not considered (71, 72). However, in cases where we detected positively selected sites, we observed a very different distribution of ω, with a peak at 20 Å from the catalytic residues. In cases where we detected 10 or more positively selected sites, this trend is even more pronounced, with the peak of ω occurring at 14 Å from catalytic residues. The enrichment of positively selected residues and elevated mean ω in the neighborhood of catalytic sites indicates that the action of positive selection reshapes the selective constraint on the entire protein structure.

Properties of amino acids at positively selected sites. As different regions of proteins are known to have different amino acid frequencies (73, 74), we asked whether the positively selected residues we detected exhibit a distinct amino acid distribution. For each protein class, we calculated the change in amino acid frequency at positively selected sites compared to the background frequencies (Fig. 6*C*). While the overall distributions of amino acids are very similar in the different protein classes (*SI Appendix*, Fig. S13), we observed differences in the distribution of amino acids at positively selected sites compared to the background distribution (Fig. 6*C*). We correlated these enrichment scores with common amino acid physicochemical properties (size, hydrophobicity, net charge, and polarity) but found no significant correlations (*SI Appendix*, Table S1), indicating that, while certain amino acids are preferred or avoided at positively selected sites, these trends bear no straightforward relationship to amino acid properties.

The role of gene-duplication events in adaptive evolution. Gene duplications are thought to be one of the main forces driving evolution, providing "raw material" for evolutionary innovations (75). While gene-duplication events in themselves are frequently assumed to have no effect on fitness, their retention can be evidence of adaptation (76). In order to quantify the effect of duplication events in our dataset, we calculated the fraction of gene duplications (i.e., the number of duplication nodes divided by the total number of nodes) for each phylogenetic tree. We found that both in enzymes and in immune-related genes, the mean paralog fraction is significantly larger than in other genes (0.342 and 0.276, respectively, compared to 0.0397 in the remaining trees) (Fig. 6*D*). This trend is significant both in the case of immune proteins and metabolic enzymes ($P = 0.015$ and $P = 3.1 \times 10^{-5}$, respectively; Kolmogorov–Smirnov test). This elevated duplication rate in genes where we detected positively selected clusters is consistent with positive selection acting not only on point mutations, but also driving gene-duplication events to fixation. At the same time, some genes where we detected strong evidence of adaptation (complement component 5, transferrin

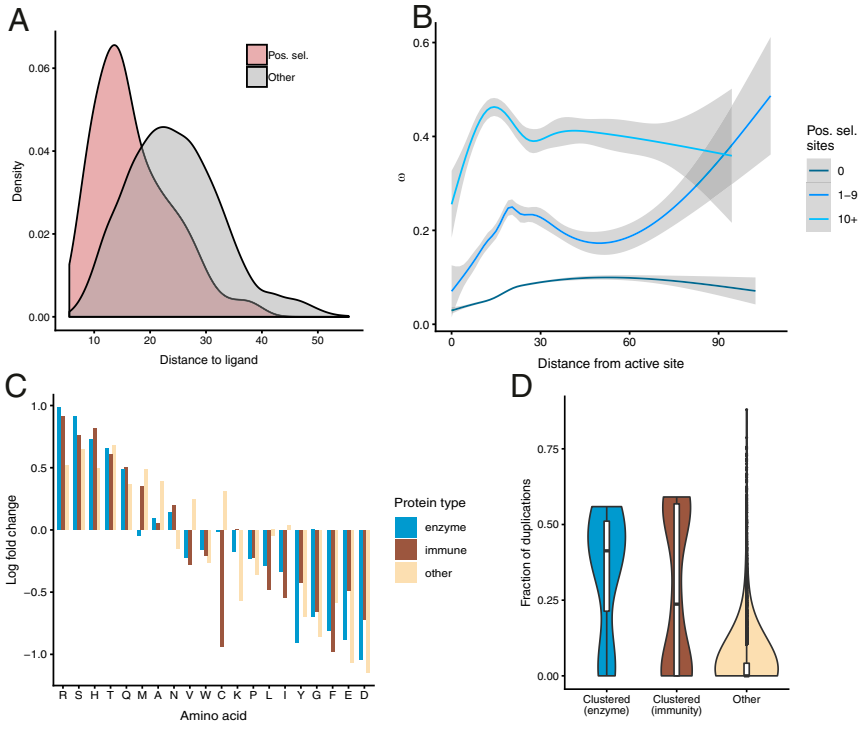

Fig. 6. Properties of positively selected sites. (*A*) Distance of positively selected residues from bound exogenous ligands. (*B*) The distribution of ω as a function of distance from catalytic residues. (*C*) Departures from the background amino acid frequencies in positively selected residues. (*D*) Distribution of fraction of gene duplications in proteins with positively selected clusters.

receptor 1, complement component 8α, AK5, and nicastrin) have not undergone any gene duplications, proving that rapid sitewise evolutionary rate and gene duplications can occur independently.

Discussion

In this study, we curated a dataset covering over one million structurally mapped sites in 3,347 mammalian proteins and assessed the placement of positively selected residues on their 3D structures in an unbiased, genome-wide manner. We find that positively selected sites tend to occur closer to each other in protein structures than is expected by chance and to form clusters in the neighborhood of functionally important regions. Strikingly, proteins where we found the strongest evidence for clustering of positively selected sites are primarily involved in two major types of environmental responses: Host–pathogen interactions and metabolism of xenobiotic compounds. The fact that we observed the strongest evidence of positive selection in these types of proteins gives a coherent view of mammalian evolution being shaped by these two major influences from the environment. Clusters of positively selected sites we identified share both functional and structural similarities and allow us to infer more general principles underlying adaptive evolution.

Xenobiotic-metabolising enzymes are typically able to process a wide range of substrates. Indeed, CYPs and AKRs, where we identified three and two positively selected clusters, respectively, are among the most promiscuous known protein superfamilies. Promiscuous enzymes are thought to be malleable in evolution, as they can maintain their original function as well as acquire specificity for new substrates by going through a promiscuous intermediate, which can bind multiple substrates (77). The mechanisms by which enzymes acquire new substrates has to date been primarily studied by directed evolution (78–81). The examples we have highlighted here provide direct evidence that similar scenarios are also common in natural evolution.

Enzymes involved in xenobiotic metabolism are of great medical relevance, as in humans they are responsible for metabolism of prescribed drugs. Traditional analyses of protein conservation are frequently not suitable for the analysis of genes involved in xenobiotic metabolism, as these tend to evolve rapidly and the analyses used do not explicitly distinguish between neutral evolution and positive selection (82). Specific examples we have identified here could be investigated further, for example by detailed mutational studies that have been shown to augment statistical modeling of adaptive evolution (83).

Our study highlights the power of incorporating independent sources of information to understand principles governing evolution. The clusters we detected consist of residues that are distributed along the linear sequence of proteins and could not be found without considering protein structure. Consideration of structural information has also allowed us to better understand the mechanistic details of processes underlying adaptation in terms of specific structural and functional features. Information about structural placement of residues can also help to address technical issues that have hindered methods for detecting positive selection. Criticisms leveled at methods for detecting positive selection have revolved around the nonneutrality of synonymous substitutions, local variation in synonymous substitution rate (84–87), and the influence of errors in alignment (88, 89). These phenomena may cause false positives in parts of a protein sequence, but none will result in clustering on protein structure. Structural information can thus serve as an independent validation and a means of demonstrating that observed patterns of positive selection are not a product of confounding factors. Structural clusters can additionally be inspected post hoc for proximity to functional features to assess their plausibility and aid interpretation.

The structural and functional similarities we identified here point toward common rules governing the occurrence of pervasive positive selection. Positively selected metabolic enzymes we describe here share many structural and functional similarities: Positively selected clusters lie in close proximity to bound ligands, indicating that the primary mode in which these enzymes adapt is by affecting residues in the direct neighborhood of active sites. This finding may seem to contradict the common assumption that functionally important residues are conserved in evolution: For example, the finding that average evolutionary rate is lowest in the neighborhood of catalytic sites (71, 72). However, this is only a superficial disagreement: While functional regions evolve more slowly on average, this does not mean they cannot harbor rapidly evolving, positively selected sites. Indeed, nonfunctional regions cannot, by definition, undergo adaptive evolution.

As we demonstrate here, while functional regions of proteins are typically more conserved, they can also exhibit a high evolutionary rate that is a hallmark of adaptive evolution. This strongly suggests that instances where positive selection is operating can contradict overall trends of protein evolution. For this reason, it may be counterproductive to incorporate known correlates of evolutionary rate into statistical models for detecting positive selection. In contrast, the fact that positively selected residues can form clusters on protein structures could inform the development of better methods for detecting positive selection. One of the ultimate goals of evolutionary research is integrating evolution of sequence with structure in a general model of protein evolution (90, 91). Such a universal model of protein evolution has been elusive so far, primarily because the most general approaches require an intractable number of parameters. We would suggest that one way forward is to identify further universal evolutionary trends and gradually incorporate them into mathematical models of protein evolution. Structural approaches are powerful tools for interpreting observed patterns of sequence divergence but, as regulatory and other noncoding regions also contribute to adaptive evolution, structure-based analyses cannot explain all instances of adaptation. Our understanding of protein evolution should ultimately be integrated with understanding of the evolution of other determinants of cellular function. The development of new methods for identifying adaptation in noncoding regions is an important future direction for evolutionary studies.

We have demonstrated that analyzing selective constraint in the context of structure can help interpret findings and increase their robustness, but all approaches reliant on detailed structural information are limited by the availability and coverage of crystal structures. Similarly, the analysis performed here focused on mammals but could be extended to other clades. We hope that the results highlighted here and others we have made available online in our web server will assist experimental validation and further understanding of protein function and adaptation. We aimed to establish the relationship between protein structure and the occurrence of positive selection and this proof-of-principle study called for the highest-possible quality data, but incorporating homology-based structural models would be a direct extension to our approach. Protein structures for the majority of human proteins are still not known and the PDB database is biased toward certain protein families. This suggests that there may be yet unknown adaptively evolving functions, and that new examples of adaptation will be identified in protein families where there is currently little or no structural information available.

Methods

Genomic Data. Coding sequences for mammalian genomes were downloaded from Ensembl (92), v78. Noneutherian genomes (platypus, gray short-tailed opossum, wallaby, and Tasmanian devil) were excluded. Coding sequences

for principal isoforms were used. Incomplete and stop codons at ends of sequences were removed.

Phylogenetic Data. The Compara database (93) provides gene trees for species stored in Ensembl. Compara gene trees are reconstructed from nucleotide and amino acid alignments augmented with information about the species tree, which ensures overall agreement with the species phylogeny while accounting for gene duplication events and also allowing for variations in the tree topology if they are supported by sequence data. The Compara pipeline generates trees containing up to 750 related genes, which frequently results in multiple paralogs being included in the same tree. Bearing in mind that selective constraint can be estimated more accurately if more sequences are included, but that including more paralogs can result in averaging over genes, which may be under different constraints, we designed a tree-splitting scheme to enable single-gene analysis. As we aimed to maximize the number of orthologous sequences included in each alignment while minimizing the number of paralogous sequences, we quantified these criteria in different possible subtrees by calculating the percentage of all species included (taxonomic coverage) and the total number of additional genes for each species beyond the first gene per species (permitting calculation of the paralog fraction). We required a taxonomic coverage of at least 60% and wished to minimize the paralog fraction. To achieve this, starting from each human protein, the tree is traversed toward the root until the desired taxonomic coverage was achieved. Then, the tree is traversed further but only if this does not increase the paralog fraction. The final node of this traversal process and all its descendant nodes then become a tree used for further analysis.

Sequence Alignment. Compara gene trees are reconstructed using principal isoforms and the same sequences were used for alignment. The PRANK aligner (28) has been shown to limit the number of false-positive identifications of positive selection compared to other commonly used aligners (89, 94, 95). PRANK was run in codon mode on sets of sequences corresponding to each Compara-derived tree and with these trees used as guide trees.

Detecting Positive Selection. SLR (29) was used to obtain sitewise estimates of ω within each alignment, using tree topologies from Ensembl Compara and allowing branch lengths to be optimized by SLR. SLR implements the Goldman–Yang codon site model (96) similar to that in PAML (3). The main difference between SLR and PAML is that SLR makes no assumption about the distribution of ω values over the sites of the alignment. SLR first estimates parameters of the phylogenetic model for the entire alignment and then performs a likelihood ratio test between the optimal ω and $\omega = 1$ for each site. *P* values reported by SLR associated with each structure-mapped site (see below) were then corrected for multiple testing using the Benjamini–Hochberg FDR method (97).

Robustness of Identified Positively Selected Sites. We created 100 bootstrap replicates for each alignment (98) and used them to generate alternative phylogenetic trees using the Ensembl Compara methodology. We then repeated the positive selection analysis in SLR using those trees and calculated a measure of support for each site detected as positively selected in the original analysis by tallying the number of replicates of 100 where that site was detected as positively selected using the same stringency criteria (with FDR = 0.05) as in the original analysis.

Structural Data. PDB structures matching human proteins in the sequence dataset were downloaded from PDBe (30). Structures covering fewer than 100 residues were excluded, and in cases where more than one structure was available, the one with the highest sequence similarity to the protein sequence was chosen. In rare cases where more than one human protein with a structure was present for an alignment, one was retained at random. Individual residues were then mapped using the SIFTS database (31). SIFTS provides a mapping between PDB (30) and UniProt (99) sequences and, as the UniProt protein sequences can vary from those in Ensembl, we performed an additional mapping step by constructing pairwise alignments between UniProt and Ensembl sequences, resulting in a sitewise mapping between Ensembl and PDB residues. The pairwise alignments were calculated using the Biopython (100) implementation of the Smith–Waterman algorithm (101), using the scoring of 1 for matching characters and 0 otherwise, and gap opening and extension penalties of −10 and −0.5 respectively.

Clustering of Positively Selected Sites. The degree of clustering of the positively selected and near-neutrally evolving sites (defined as those where the 95% confidence interval for ω as reported by SLR includes 1) within each protein structure was assessed using the CLUMPS algorithm (21). In CLUMPS, the degree of clustering for a set of residues of interest is quantified by the sum of pairwise distances in 3D space. In contrast to the original implementation, we used equal weights for all sites when calculating the pairwise distances. For each set of residues, we then performed 100,000 Monte Carlo simulations permuting the placement of sites by randomly selecting positions from the PDB chain, in order to determine statistical significance of observed patterns. *P* values resulting from this analysis were then corrected for multiple comparisons using the Benjamini–Hochberg FDR method (97).

Statistical analyses were performed in the R environment (102).

Data Availability Statement. All results discussed in the paper (structure-mapped values of selective constraint, underlying alignments, and phylogenetic trees) are available at https://www.ebi.ac.uk/goldman-srv/sips/.

ACKNOWLEDGMENTS. We thank Dr. Leo C. James, Dr. M. Madan Babu, Dr. Patrycja Kozik, Dr. Maria Martí-Solano, and members of the Goldman Group at the European Molecular Biology Laboratory, European Bioinformatics Institute for helpful discussions and comments on the manuscript.

1. J. M. Havrilla, B. S. Pedersen, R. M. Layer, A. R. Quinlan, A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
2. Z. L. Fuller, J. J. Berg, H. Mostafavi, G. Sella, M. Przeworski, Measuring intolerance to mutation in human genetics. *Nat. Genet.* **51**, 772–776 (2019).
3. Z. Yang, PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
4. S. Weaver *et al.*, Datamonkey 2.0: A modern web application for characterizing selective and other evolutionary processes. *Mol. Biol. Evol.* **35**, 773–777 (2018).
5. S. A. Benner, Natural progression. *Nature* **409**, 459 (2001).
6. J. Echave, S. J. Spielman, C. O. Wilke, Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17**, 109–121 (2016).
7. A. L. Hughes, M. Nei, Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170 (1988).
8. S. L. Sawyer, L. I. Wu, M. Emerman, H. S. Malik, Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2832–2837 (2005).
9. R. K. Schott, S. P. Refvik, F. E. Hauser, H. López-Fernández, B. S. Chang, Divergent positive selection in rhodopsin from lake and riverine cichlid fishes. *Mol. Biol. Evol.* **31**, 1149–1165 (2014).
10. T. Endo, K. Ikeo, T. Gojobori, Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* **13**, 685–690 (1996).
11. C. Kosiol *et al.*, Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).
12. L. Eory, D. L. Halligan, P. D. Keightley, Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. *Mol. Biol. Evol.* **27**, 177–192 (2010).
13. K. Lindblad-Toh *et al.*; Broad Institute Sequencing Platform and Whole Genome Assembly Team; Baylor College of Medicine Human Genome Sequencing Center Sequencing Team; Genome Institute at Washington University, A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
14. J. Roux *et al.*, Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* **31**, 1661–1685 (2014).
15. F. Cicconardi, P. Marcatili, W. Arthofer, B. C. Schlick-Steiner, F. M. Steiner, Positive diversifying selection is a pervasive adaptive force throughout the *Drosophila* radiation. *Mol. Phylogenet. Evol.* **112**, 230–243 (2017).
16. Z. Yang, R. Nielsen, N. Goldman, In defense of statistical methods for detecting positive selection. *Proc. Natl. Acad. Sci. U.S.A.* **106**, E95, author reply E96 (2009).
17. W. Zhai, R. Nielsen, N. Goldman, Z. Yang, Looking for Darwin in genomic sequences—Validity and success of statistical methods. *Mol. Biol. Evol.* **29**, 2889–2893 (2012).
18. R. A. Laskowski, J. M. Thornton, Understanding the molecular machinery of genetics through 3D structures. *Nat. Rev. Genet.* **9**, 141–151 (2008).
19. J. R. Homburger *et al.*, Multidimensional structure-function relationships in human β-cardiac myosin from population-scale genetic variation. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6701–6706 (2016).
20. M. L. Miller *et al.*, Pan-cancer analysis of mutation hotspots in protein domains. *Cell Syst.* **1**, 197–209 (2015).
21. A. Kamburov *et al.*, Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5486–E5495 (2015).
22. B. Niu *et al.*, Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).
23. C. L. Araya *et al.*, Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat. Genet.* **48**, 117–125 (2016).
24. M. Kimura, Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
25. M. Kimura, On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719 (1962).
26. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**, 418–426 (1986).

27. A. J. Vilella *et al*., EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
28. A. Löytynoja, N. Goldman, Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635 (2008).
29. T. Massingham, N. Goldman, Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753–1762 (2005).
30. S. Mir *et al*., PDBe: Towards reusable data delivery infrastructure at protein data bank in Europe. *Nucleic Acids Res.* **46**, D486–D492 (2018).
31. J. M. Dana *et al*., SIFTS: Updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489 (2019).
32. A. Tóth-Petróczy, D. S. Tawfik, Slow protein evolutionary rates are dictated by surface-core association. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 11151–11156 (2011).
33. T. Xie *et al*., Crystal structure of the γ-secretase component nicastrin. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 13349–13354 (2014).
34. D. Enard, L. Cai, C. Gwennap, D. A. Petrov, Viruses are a dominant driver of protein adaptation in mammals. *eLife* **5**, e12469 (2016).
35. A. E. Webb *et al*., Adaptive evolution as a predictor of species-specific innate immune response. *Mol. Biol. Evol.* **32**, 1717–1729 (2015).
36. E. R. Ebel, N. Telis, S. Venkataram, D. A. Petrov, D. Enard, High rate of adaptation of mammalian proteins that interact with Plasmodium and related parasites. *PLoS Genet.* **13**, e1007023 (2017).
37. M. Sironi, R. Cagliani, D. Forni, M. Clerici, Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **16**, 224–236 (2015).
38. A. Demogines, J. Abraham, H. Choe, M. Farzan, S. L. Sawyer, Dual host-virus arms races shape an essential housekeeping protein. *PLoS Biol.* **11**, e1001571 (2013).
39. H. M. Kim *et al*., Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. *Cell* **130**, 906–917 (2007).
40. J. A. Huntington, R. J. Read, R. W. Carrell, Structure of a serpin-protease complex shows inhibition by deformation. *Nature* **407**, 923–926 (2000).
41. S. Ye *et al*., The structure of a Michaelis serpin-protease complex. *Nat. Struct. Biol.* **8**, 979–983 (2001).
42. L. Vidalino *et al*., SERPINB3, apoptosis and autoimmunity. *Autoimmun. Rev.* **9**, 108–112 (2009).
43. C. Heit *et al*., Update of the human and mouse SERPIN gene superfamily. *Hum. Genomics* **7**, 22 (2013).
44. K. Izuhara, S. Ohta, S. Kanaji, H. Shiraishi, K. Arima, Recent progress in understanding the diversity of the human ov-serpin/clade B serpin family. *Cell. Mol. Life Sci.* **65**, 2541–2553 (2008).
45. A. L. Hughes, T. Ota, M. Nei, Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. *Mol. Biol. Evol.* **7**, 515–524 (1990).
46. L. Van Valen, A new evolutionary law. *Evol. Theory* **1**, 1–30 (1973).
47. T. Lynch, A. Price, The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *Am. Fam. Physician* **76**, 391–396 (2007).
48. G. R. Wilkinson, Drug metabolism and variability among patients in drug response. *N. Engl. J. Med.* **352**, 2211–2221 (2005).
49. T. M. Penning, The aldo-keto reductases (AKRs): Overview. *Chem. Biol. Interact.* **234**, 236–246 (2015).
50. Y. Jin, T. M. Penning, Aldo-keto reductases and bioactivation/detoxication. *Annu. Rev. Pharmacol. Toxicol.* **47**, 263–292 (2007).
51. N. R. Bachur, Cytoplasmic aldo-keto reductases: A class of drug metabolizing enzymes. *Science* **193**, 595–597 (1976).
52. O. A. Barski, S. M. Tipparaju, A. Bhatnagar, The aldo-keto reductase superfamily and its role in drug metabolism and detoxification. *Drug Metab. Rev.* **40**, 553–624 (2008).
53. A. D. Gloss *et al*., Evolution in an ancient detoxification pathway is coupled with a transition to herbivory in the drosophilidae. *Mol. Biol. Evol.* **31**, 2441–2456 (2014).
54. T. Lan, X.-R. Wang, Q.-Y. Zeng, Structural and functional evolution of positively selected sites in pine glutathione S-transferase enzyme family. *J. Biol. Chem.* **288**, 24441–24451 (2013).
55. R. R. da Fonseca, W. E. Johnson, S. J. O'Brien, V. Vasconcelos, A. Antunes, Molecular evolution and the role of oxidative stress in the expansion and functional diversification of cytosolic glutathione transferases. *BMC Evol. Biol.* **10**, 281 (2010).
56. Y. Ivarsson, A. J. Mackey, M. Edalat, W. R. Pearson, B. Mannervik, Identification of residues in glutathione transferase capable of driving functional diversification in evolution. A novel approach to protein redesign. *J. Biol. Chem.* **278**, 8733–8738 (2003).
57. D. Wang *et al*., Human carboxylesterases: A comprehensive review. *Acta Pharm. Sin. B* **8**, 699–712 (2018).
58. S. Bencharit, C. L. Morton, Y. Xue, P. M. Potter, M. R. Redinbo, Structural basis of heroin and cocaine metabolism by a promiscuous human drug-processing enzyme. *Nat. Struct. Biol.* **10**, 349–356 (2003).
59. A. Allali-Hassani *et al*., Structural and chemical profiling of the human cytosolic sulfotransferases. *PLoS Biol.* **5**, e97 (2007).
60. N. Gamage *et al*., Human sulfotransferases and their role in chemical metabolism. *Toxicol. Sci.* **90**, 5–22 (2006).
61. S. J. Kerns *et al*., The energy landscape of adenylate kinase during catalysis. *Nat. Struct. Mol. Biol.* **22**, 124–131 (2015).
62. T. P. Schrank, J. O. Wrabl, V. J. Hilser, Conformational heterogeneity within the LID domain mediates substrate binding to *Escherichia coli* adenylate kinase: Function follows fluctuations. *Top. Curr. Chem.* **337**, 95–121 (2013).
63. F. Jing *et al*., Phylogenetic and experimental characterization of an acyl-ACP thioesterase family reveals significant diversity in enzymatic specificity and activity. *BMC Biochem.* **12**, 44 (2011).
64. X. C. Bai *et al*., An atomic structure of human γ-secretase. *Nature* **525**, 212–217 (2015).
65. W. Messier, C. B. Stewart, Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154 (1997).
66. J. Zhang, Y. P. Zhang, H. F. Rosenberg, Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.* **30**, 411–415 (2002).
67. F. Rodríguez-Trelles, R. Tarrío, F. J. Ayala, Convergent neofunctionalization by positive Darwinian selection after ancient recurrent duplications of the xanthine dehydrogenase gene. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13413–13417 (2003).
68. L. Yu *et al*., Adaptive evolution of digestive RNASE1 genes in leaf-eating monkeys revisited: New insights from ten additional colobines. *Mol. Biol. Evol.* **27**, 121–131 (2010).
69. C. Monit *et al*., Positive selection in dNTPase SAMHD1 throughout mammalian evolution. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 18647–18654 (2019).
70. N. Furnham *et al*., The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42**, D485–D489 (2014).
71. B. R. Jack, A. G. Meyer, J. Echave, C. O. Wilke, Functional sites induce long-range evolutionary constraints in enzymes. *PLoS Biol.* **14**, e1002452 (2016).
72. L. Rockah-Shmuel, Á. Tóth-Petróczy, D. S. Tawfik, Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. *PLOS Comput. Biol.* **11**, e1004421 (2015).
73. N. Goldman, J. L. Thorne, D. T. Jones, Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**, 445–458 (1998).
74. G. J. Bartlett, C. T. Porter, N. Borkakoti, J. M. Thornton, Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121 (2002).
75. S. Ohno, *Evolution by Gene Duplication* (Springer-Verlag, London, 1970).
76. M. P. Francino, An adaptive radiation model for the origin of new gene functions. *Nat. Genet.* **37**, 573–577 (2005).
77. O. Khersonsky, C. Roodveldt, D. S. Tawfik, Enzyme promiscuity: Evolutionary and mechanistic aspects. *Curr. Opin. Chem. Biol.* **10**, 498–508 (2006).
78. D. M. Schmidt *et al*., Evolutionary potential of $(\beta/\alpha)_8$-barrels: Functional promiscuity produced by single substitutions in the enolase superfamily. *Biochemistry* **42**, 8387–8393 (2003).
79. S. C. Rothman, J. F. Kirsch, How does an enzyme evolved in vitro compare to naturally occurring homologs possessing the targeted function? Tyrosine aminotransferase from aspartate aminotransferase. *J. Mol. Biol.* **327**, 593–608 (2003).
80. D. Hoffmeister, J. Yang, L. Liu, J. S. Thorson, Creation of the first anomeric D/L-sugar kinase by means of directed evolution. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 13184–13189 (2003).
81. A. Aharoni *et al*., The 'evolvability' of promiscuous protein functions. *Nat. Genet.* **37**, 73–76 (2005).
82. Y. Zhou, S. Mkrtchian, M. Kumondai, M. Hiratsuka, V. M. Lauschke, An optimized prediction framework to assess the functional impact of pharmacogenetic variants. *Pharmacogenomics J.* **19**, 115–126 (2019).
83. J. D. Bloom, Identification of positive selection in genes is greatly improved by using experimentally informed site-specific models. *Biol. Direct* **12**, 1 (2017).
84. J. L. Parmley, J. V. Chamary, L. D. Hurst, Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol. Biol. Evol.* **23**, 301–309 (2006).
85. M. Macossay-Castillo, S. Kosol, P. Tompa, R. Pancsa, Synonymous constraint elements show a tendency to encode intrinsically disordered protein segments. *PLOS Comput. Biol.* **10**, e1003607 (2014).
86. R. Savisaar, L. D. Hurst, Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution. *Mol. Biol. Evol.* **34**, 1110–1126 (2017).
87. I. I. Davydov, N. Salamin, M. Robinson-Rechavi, Large-scale comparative analysis of codon models accounting for protein and nucleotide selection. *Mol. Biol. Evol.* **36**, 1316–1332 (2019).
88. A. Schneider *et al*., Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.* **1**, 114–118 (2009).
89. G. Jordan, N. Goldman, The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.* **29**, 1125–1139 (2012).
90. U. Perron, A. M. Kozlov, A. Stamatakis, N. Goldman, I. H. Moal, Modeling structural constraints on protein evolution via side-chain conformational states. *Mol. Biol. Evol.* **36**, 2086–2103 (2019).
91. U. Perron, I. Moal, J. Thorne, N. Goldman, Eds., *Probabilistic Models for the Study of Protein Evolution*, D. J. Balding, I. Moltke, J. Marioni, Eds. (Wiley-Interscience, ed. 4, 2019).
92. F. Cunningham *et al*., Ensembl 2019. *Nucleic Acids Res.* **47**, D745–D751 (2019).
93. J. Herrero *et al*., Ensembl comparative genomics resources. *Database* **2016**, baw053 (2016).
94. W. Fletcher, Z. Yang, The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267 (2010).
95. P. Markova-Raina, D. Petrov, High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res.* **21**, 863–874 (2011).
96. N. Goldman, Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**, 725–736 (1994).
97. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
98. J. Felsenstein, Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
99. The UniProt Consortium, UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
100. P. J. Cock *et al*., Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
101. T. F. Smith, M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).
102. R Core Team, *R: A Language and Environment for Statistical Computing* (Version 3.5.0, R Foundation for Statistical Computing, Vienna, 2018). https://www.R-project.org/. Accessed 1 May 2018.