Pharmacophore screening of the Protein Data Bank for specific binding site chemistry

Valérie Campagna-Slater[1], *Andrew G. Arrowsmith*[1], *Yong Zhao*[1] *and Matthieu Schapira*[1,2,*]

[1]Structural Genomics Consortium, University of Toronto, MaRS Centre, South Tower, 7th floor, 101 College Street, Toronto, Ontario, Canada, M5G 1L7

[2]Department of Pharmacology and Toxicology, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, Ontario, Canada, M5S 1A8

E-mail: matthieu.schapira@utoronto.ca

Title running head: Pharmacophore screening of the PDB

ABSTRACT

A simple computational approach was developed to screen the Protein Data Bank (PDB) for putative pockets possessing a specific binding site chemistry and geometry. The method employs two commonly used 3D screening technologies, namely identification of cavities in protein structures and pharmacophore screening of chemical libraries. For each protein structure, a pocket finding algorithm is used to extract potential binding sites containing the correct types of residues which are then stored in a large SDF-formatted virtual library; pharmacophore filters describing the desired binding site chemistry and geometry are then applied to screen this virtual library and identify pockets matching the specified structural chemistry. As an example, this approach was used to screen all human protein structures in the PDB and identify sites having similar chemistry to known methyl-lysine binding domains that recognize chromatin methylation marks. The selected genes include known readers of the histone code as well as novel binding pockets that may be involved in epigenetic signaling. Putative allosteric sites were identified on the structures of TP53BP1, L3MBTL3, CHEK1, KDM4A and CREBBP.

INTRODUCTION

With approximately 61,000 macromolecular structures available, the Protein Data Bank (PDB) is a rich source of information to understand the structural mechanism of specific biological systems, or to rationally design drug candidates for specific targets. In recent years, efforts to interrogate the protein structure space in a more systematic manner have also emerged.[1] Sophisticated computational methods have been developed to probe protein structures for potential binding pockets, analyze the properties of these sites, and even predict their druggability (see recent review by Henrich et al.[2]). Approaches for identifying putative pockets and interaction sites along the surface of proteins can be classified as either geometric or energy-based (e.g. POCKET[3], SURFNET[4], CAST[5], LIGSITE[6], LIGSITE^{cs}[7], PASS[8], PocketPicker[9], icmPocketFinder[10], Q-SiteFinder[11], etc); consensus approaches have also been proposed to combine pocket predictions arising from different methods (e.g. MetaPocket[12]). Algorithms such as these can, for instance, enable the detection of allosteric binding cavities on functionally characterized proteins, thus revealing unknown protein-ligand interaction sites that can be used as novel targets for rational drug design.

A number of techniques have also emerged to gauge pocket similarities. Assessing the similarity between putative binding cavities and pre-assembled databases of known binding sites (e.g. CASTp[13], SURFACE[14], SitesBase[15], FireDB[16], CPASS database[17]) can be used for functional annotation of uncharacterized proteins when global sequence similarity or fold recognition methods are insufficient. (e.g. work by Ferrè et al.[18] and Liu et al.[19]) Such methods rely on local *sequence* similarities (e.g. ConSurf[20]) and/or local *structural* similarities to compare sites and identify important binding cavities along protein surfaces. For instance, CPASS[17] uses an RMSD weighted BLOSUM62 scoring function to find the optimal superimposition of a site onto sites contained in a pre-compiled database of known ligand binding pockets and assess their similarities. FunClust[21] on the other hand is an algorithm that

can identify common structural motifs in sets of non-homologous proteins by finding subsets of similar residues that can be superimposed within a given RMSD threshold. Cavbase[22,23] uses physicochemical descriptors to describe the residues lining cavity surfaces, and a clique detection algorithm to identify similarities between sites. SuMo[24,25] utilizes chemical groups to represent different amino acids, triplets of chemical groups (i.e. triangles) to describe local protein regions, and finally adjacent triangles are connected to yield a graph representation of the protein; when proteins are compared, a heuristic algorithm is used to find sets of pairs of similar triangles in the two proteins. Another method, IsoCleft[26], uses an efficient graph-matching-based algorithm to detect 3D *atomic* similarities between binding cavities to discriminate between sites binding similar or different ligands. Other algorithms use energy-based functions to carry out site comparisons. For example, FLAP[27] uses GRID[28-30] molecular interaction fields to generate 4-pt pharmacophore representations of targets, and uses these fingerprints to align pairs of pockets; the GRID molecular interaction fields are then used to measure site similarity.

In this paper, we describe a computational approach that was designed to provide a simple and straightforward way to search the Protein Data Bank for sites possessing a specific chemistry and geometry, by making use of two well established technologies available in most commercial computational chemistry suites: pocket searching and pharmacophore screening. First, the pocket searching algorithm *icmPocketFinder*[10] (ICM[31], Molsoft LLC) is used to identify all pockets in human PDB structures. The coordinates of specific amino acids lining these putative binding sites are then extracted and stored as entries in a very large SDF-formatted virtual library. Finally, 3-, 4- and 5-point pharmacophores capturing the desired pocket chemistry and geometry are used as queries to screen the virtual library of protein sites with the ICM pharmacophore searching algorithm.[32]

To illustrate the potential of the proposed methodology, methyl-lysine (Me-Lys) binding domains (which recognize Me-Lys marks on histone tails) were chosen as a system of interest. The vast majority of known Me-Lys readers (PHD, Chromo, MBT, Tudor, and PWWP domains) possess an aromatic cage

(composed of two or more Phe, Tyr or Trp residues) that may additionally include an Asp or Glu residue;[33] similarly, the Me-Lys binding site in the Ankyrin repeat of EHMT1 is formed by an aromatic cage containing an acidic residue.[34] The PDB was therefore screened for pockets lined with aromatic and acidic residues, and pharmacophores based on known Me-Lys binding sites were used to identify pockets with correct relative geometry of the residues. The approach described herein is meant to be not only a simple method, but one that is general and can be adapted to search the PDB for other types of binding site chemistry and geometry.

METHODS

The methodology can be broken down into four main steps (**Fig. 1**):

Step 1: Representing the desired site chemistry. Since the number of aromatic and acidic residues can vary between different Me-Lys reading modules, and given that some binding sites are located in surface grooves while others form slightly deeper cavities, 10 different Me-Lys binding sites were selected to represent the structural diversity of the so-called aromatic cage system.[33] (**Fig. 2**) Only proteins for which at least one Me-Lys bound co-crystal structure was available were chosen, and for each protein both ligand-bound and apo structures were used when available. The PDB structures used for representing the desired chemistry are listed in **Table 1.** For each structure selected, query residues were represented by pharmacophores generated with ICM[32] as follows (**Fig. 1**, panels **a-b**): aromatic residues are represented by an aromatic center (Qm) placed at the middle of the aromatic ring and accompanied by a direction vector (Qv) perpendicular to the plane of the ring, while a negative center (Qn) is used for the carboxylate group of acidic side-chains; other residues lining the Me-Lys binding site are omitted. For Phe, Tyr and Trp residues to be treated as interchangeable, it is necessary that only one aromatic center be used for each of these 3 residues. Therefore, only one of the aromatic centers is kept for each Trp in a given site, and different pharmacophore representations are created for such sites

to generate all possible combinations of one aromatic center per Trp (i.e. a site containing *n* Trp residues is represented by 2^n pharmacophores).

Step 2: Generating a library of putative pockets. An SDF-formatted virtual library is assembled by searching the PDB for pockets lined by clusters of aromatic and acidic residues. For each PDB structure, the protein is first stripped of water molecules, ligands (including nucleic acids) and bound peptides (any chain shorter than 25 residues is removed), converted to an ICM object (using the *makeBioMT* and *convertObject* macros[32]), and the *icmPocketFinder* algorithm[10] is used to identify all putative binding pockets in the three-dimensional structure (**Fig. 1**, panel **c**). For the *icmPocketFinder* algorithm, a tolerance value of 4.0 is used instead of the default value which is 4.6; when a lower tolerance value is selected, protein surfaces are scanned at higher resolution and smaller or shallower pockets are identified. For each of these pockets, aromatic (Phe, Tyr, Trp) and acidic (Asp, Glu) residues within 3.0 Å of the pocket surface are selected (**Fig. 1**, panel **d**). Since some predicted pockets are elongated due to the protein's landscape, residues are removed iteratively to retain only residues forming a localized site (**Fig. 1**, panel **e**). This pruning procedure is carried out by calculating all pair-wise distances between residue side-chains and removing the residue having the highest average pair-wise distance between its side-chain and all side-chains included in the pocket selection, if this value is above 6.7 Å. This is repeated until the highest average pair-wise distance is no longer above 6.7 Å (a value of 6.7 Å was chosen after testing different values on representative structures). If the final selection contains 3 residues or more and at least 2 of these have aromatic side-chains, then the coordinates of these residues are extracted from the protein structure and stored as a new entry in an SD file (**Fig. 1**, panel **f**). This procedure is carried out for each pocket identified in every protein analyzed, yielding a virtual library containing groups of residues possessing the required chemistry, but not necessarily the required geometry.

Step 3: Screening the SDF-formatted library using the pharmacophore query. Although all the sites extracted into the virtual library in **Step 2** contain the correct *type* of residues, their relative *geometry* may not correspond to those observed in known Me-Lys binding sites. The pharmacophores generated in **Step 1** are therefore used as queries to search this virtual library for groups of residues having the correct chemistry *and* relative geometry. The b-factors associated with the pharmacophore size (Qm/n) and direction vectors (Qv) can be modified to capture more or fewer sites; the b-factor is a resolution parameter corresponding to the maximum allowed distance between centers within a match.[32] Only sites that are very close in geometry to the sites used to generate the pharmacophores are retrieved when using small b-factors, while using bigger b-factors allows for the identification of pockets diverging more substantially from the ideal site geometry.

Step 4: Filtering the results. To further refine the preliminary hit list, an additional filtering layer is used to identify the most promising sites. Residues matching the pharmacophore hypothesis are analyzed within the context of the entire protein structure to remove sites which, although matching the pharmacophore query within the allowed threshold, are not predicted to form a potential pocket. First, a probe atom is placed at the center point of the aromatic and acidic residues selected in **Step 2** (using only side-chain heavy atoms to compute the probe position), and the shortest distance between this probe and all residues in the protein is calculated. Sites for which this distance is smaller than 2.0Å are then removed (the 2.0 Å threshold value was selected based on the results of the validation study, and the observation that for all domains used to generate the pharmacophores, this value is greater than 2.0 Å for at least one structure used). Other tactics can also be used to rank the remaining hits, including using localization data to prioritize proteins located in the nucleus (since these are more likely to be biologically relevant), or retrieving Pubmed citations for the identified proteins and searching these for abstracts/titles containing relevant keywords such as “histone”, “chromatin” and “epigenetic”. The Pubmed hit count is used as an indicator, not a validation tool, and should be interpreted with caution: while a high hit count strongly suggests that the target is involved in epigenetic signaling, a low hit

count does not mean that the target is unrelated to epigenetic mechanisms. When ranking the hits, two sites from the same protein (originating either from different chains of the same PDB structure, or from different structure records of the same protein) are considered to be equivalent if they share at least 3 residues.

RESULTS

Two well established 3D screening technologies – pocket and pharmacophore searching algorithms – were used to extract sites possessing pre-defined chemistry from the PDB in a straightforward way (**Fig. 1**). First, pockets were extracted from a set of PDB structures using the *icmPocketFinder* pocket searching algorithm,[10] and coordinates of sites composed of at least 3 aromatic residues *or* 1 acidic and 2 aromatic residues were stored in an SDF-formatted virtual library. Next, 3-, 4- and 5-point pharmacophores generated for the canonical sites listed in **Table 1** (examples shown in **Fig. 2**) were used to screen this pocket library, in order to identify putative Me-Lys binding sites.

Validation study. In order to determine the ability of this computational approach to identify Me-Lys binding sites not represented in the pharmacophore set (**Table 1**), a validation set was assembled by extracting from the PDB other structures containing Me-Lys binding folds with the minimum requirement of 3 aromatic residues *or* 1 acidic and 2 aromatic residues. This resulted in a validation set comprised of 46 structures covering 23 proteins and 4 different domains (6 Chromo domains [CBX1, CBX2, CBX4, CBX7, CBX8 and CDYL], 5 MBT domains [L3MBTL, L3MBTL3, SCMH1, SCML2 and SFMBT2], 4 PWWP domains [HDGF, HDGFRP3, MSH6 and WHSC1L1], and 8 Tudor domains [MTF2, PHF1, PHF19, PHF20L1, SETDB1, SND1, TDRD3 and TDRKH]). Together, these 23 selected proteins contain 25 known sites used for the validation study (L3MBTL and L3MBTL3 both had 2 sites included in the validation set: the first and third MBT domains of L3MBTL were included but not its second MBT domain which was used to generate one of the pharmacophore queries, while only the first

and third MBT domains of L3MBTL3 were included since they correspond to its only two solved domains).

An SDF-formatted virtual library containing 117 pockets representing 48 unique sites was assembled from these structures using the protocol described in **Fig. 1** (**Step 2**). Out of these 48 sites, 18 correspond to known aromatic cages, although these do not necessarily all bind methylated lysine residues (e.g. the first and third MBT domains of L3MBTL were extracted using this approach, yet so far only the second MBT domain has been shown to bind methylated lysines[36]). These 18 extracted known aromatic cages cover 16 of the 23 proteins included in the data set (the sites in CBX7, SCMH1, SFMBT2, HDGFRP3, MSH6, WHSC1L1 and MTF2 were not successfully retrieved). The remaining 30 unique aromatic sites were annotated as unexpected, since they do not correspond to the known pockets. Some of these may be of biological relevance, as discussed below.

Next, the pharmacophores generated in **Step 1** (complete list in **Table 1** and sample structures shown in **Fig. 2**) were used to screen the 117 sites for potential matches, i.e. sites matching not only the required residue types, but also the relative geometry. The number of unique sites selected using various radius (Qm/n) and direction (Qv) b-factors is shown in Fig. 3 (results from the individual pharmacophore queries were merged and redundant hits were clustered). Ultimately, the objective is not to retrieve all sites extracted from the structures (in which case it would be pointless to perform the pharmacophore query), but rather to extract exclusively aromatic cages. The results are split into known aromatic cages and unexpected aromatic sites selected in **Figs. 3a** and **3b**, respectively. Pharmacophore radius b-factors of 1.3 Å (Qm/n) and direction b-factors of 0.5 Å (Qv) were selected as the optimum values (**Fig. 3**, dashed line). Using these parameters, 15 of the 18 known aromatic cages in the SDF-formatted virtual library were retrieved (83.3%), whereas only 7 of the 30 unexpected aromatic sites were selected (23.3%). These 22 selected sites are listed in **Table 2**, and ranked according to the shortest distance between the protein and a probe located at the centroid of the site, as calculated in the filtering

procedure described in Methods, **Step 4.** From this data, a threshold of 2.0 Å was chosen for post-pharmacophore filtering, since this would allow the selection of all domains used to generate the pharmacophores (data not shown) as well as 9 of the known sites in the validation study (while only selecting one unexpected site).

In **Fig. 4,** the results from **Fig. 3** are split by pharmacophore, for b-factors (Qm/n) up to 1.3 Å. Pharmacophores were generated from sites containing 3 residues (**C, G**), 4 residues (**A, B, D, E, F, H, J**) or 5 residues (**I**), as described in **Table 1**. As can clearly be seen, pharmacophores **C** (CHD1) and **G** (PYGO1) retrieve a much larger number of sites compared to the other pharmacophores, and in particular a much larger fraction of unexpected aromatic sites (**Fig. 4b**). Indeed, queries formed by only 3 pharmacophore centers are likely to be more promiscuous, as opposed to a 3-dimensional site representation that is obtained when 4 or more pharmacophore points are used. On the other hand, pharmacophore **I** (TP53BP1) is not able to identify any aromatic cage other than itself (**Fig. 4a**), demonstrating that 5-pt pharmacophores may be too selective.

Different pocket conformations (including both ligand-bound and apo structures, **Table 1**) were used to generate pharmacophore queries in order to maximize the likeliness of extracting novel sites that may not be in an optimal conformation. Although this, together with the use of permissive pharmacophore b-factors, allows for sites diverging from an ideal ligand-bound conformation to be retrieved in the screening process by indirectly accounting for side-chain flexibility, the method is still limited by the ability of the *icmPocketFinder* algorithm[10] to identify cavities. For example, in the structure of the second MBT domain of SCMH1 (PDB: 2p0k), the Trp204 side-chain is folded into the potential site and probably involved in π-π stacking with Phe201, thus blocking the cavity and prohibiting the identification of a pocket. In other cases, a putative binding cavity may be missed by *icmPocketFinder* if the pocket is too shallow, given the chosen tolerance value (4.0 in this study). For instance, several co-crystal structures recently deposited to the PDB revealed a novel Me-Lys binding site on the WD40

domain of EED (e.g. PDB: 3ij1). Although the apo structure of the EED WD40 domain (PDB: 2qxv) was included in the list of human proteins extracted from the PDB (see below), this site was not identified because the cavity is more shallow in the unbound conformation compared to the peptide-bound conformation.

Screening against all human proteins in the Protein Data Bank. A list of 11,199 x-ray and NMR human protein structures annotated as human in the PDB was assembled (August 2008), and **Step 2** (**Fig. 1**) of the algorithm yielded a virtual library containing 22,568 putative sites stored in a large SDF-formatted file. Next, pharmacophore screening was performed as described in **Step 3** (Qm/n = 1.3 Å, Qv = 0.5 Å), yielding a set of 6,340 non-unique sites. Pockets from PDB chains possessing less than 90% sequence identity with any gene in the human genome were then removed (1. RefSeq Release 36 was downloaded to obtain all annotated human genes and a list of human protein sequences was generated in FASTA format, 2. blastp was used to search the PDB sequences against the human proteins, and 3. Sites located in PDB chains possessing less than 90% sequence identity with human proteins according to blast were removed). This resulted in a final list of 5,883 non-unique protein sites, covering 968 different proteins (or protein complexes). Using the filter defined in **Step 4** with a threshold of 2.0 Å, this preliminary set of hits was reduced to a final list containing 236 unique sites extracted from 206 proteins, or protein complexes (i.e. sites located at the interface between different proteins). Running the pocket detection **(Step 2)** on all 11,199 structures took one day of computations using 10 CPUs (this step only needs to be carried out once), while the pharmacophore search **(Step 3)** is very fast, taking approximately 6 minutes on a single CPU to screen all 22,568 putative sites against all pharmacophore representations.

As in the validation study, the 5-pt pharmacophore (**I**) was unable to locate any site other than itself. The 3-pt pharmacophores (**C, G**) were used even though they were shown to be more promiscuous, in order to ensure that potential Me-Lys binding sites containing only 1 acidic and 2 aromatic residues

could be identified. The hits were ranked according to different criteria: 1) according to the shortest distance to a probe located at the center of the predicted aromatic site (Methods, **Step 4**), 2) according to the RMSD with the pharmacophore query, and 3) according to the total number of potentially relevant Pubmed hits (keywords used: chromatin, histone, and epigenetic) – the top 50 hits are reported in **Table 3** for each ranking criteria, and a complete list is provided in the Supporting Information, **Table S1**. **Table 4** lists 36 sites handpicked from the hit list, based on a subjective examination of the structures.

DISCUSSION

Aromatic cages are identified at non-Me-Lys binding sites. The validation study described above clearly demonstrates the ability of the method to retrieve many of the known Me-Lys binding modules from a set of protein structures. In addition, aromatic cage systems were identified at sites acting as binding platforms for non Me-Lys peptides. For instance, several Kringle domains were extracted from Plasminogen (PLG – e.g. Kringle 2 domain, PDB: 1b2i) and Lipoprotein(a) (LPA – e.g. Kringle IV-10 domain, PDB: 3kiv) structures (**Table 3**; see also Supporting Information **Fig. S1**). Although these domains bind *unmodified* lysines,[37,38] it is not surprising that such sites are retrieved, given that they possess a similar chemistry and geometry to those observed in the Me-Lys readers, i.e. an aromatic cage including acidic residues. Clearly, if theses structures are related to readers of the histone code, the biology is not in these cases, since these proteins are not localized in the nucleus.

Interestingly, some arginine binding sites were also identified, such as the WD40 repeat of WDR5[39] (PDB: 2g9a, residues: D92, F133, F219 and F263 - **Table 3**, 13th ranked by distance to pocket center), and an Arg binding site located between the tandem SH3 domains of NCF1[40] (not in the top 50 hits). This might be viewed as uncovering a chemical similarity between the binding pockets for Me-Lys and Arg residues (both of which are positively charged and of a similar size), or on the contrary as a weakness of the method, as it fails to differentiate between these binding sites. Had more restrictive b-

factors been used, these sites, for which the geometry diverges slightly from those represented by the pharmacophores, would not have been selected; this outlines the dilemma between choosing a restrictive set of parameters, which may prohibit the identification of novel sites, or choosing permissive parameters that allow such sites to be extracted, but also increase the number of false hits.

It has previously been shown that the first MBT domain of L3MBTL does not bind methylated lysines, but that it can accommodate proline residues;[36] given its high degree of similarity to other MBT domains (including those known to bind Me-Lys residues) it is not surprising that the algorithm retrieves this pocket as a putative hit (**Table 2**, rank 19). It is interesting to note that other proline binding sites were also extracted from the PDB. For example, the active site of several FKBP prolyl isomerases[41] (e.g. FKBP1A, not listed in **Table 3**), and a pocket in ERAF (**Table 3**, 16th ranked by distance to pocket center) which is located at the αHb-interaction surface[42] and accommodates Pro120 (PDB: 1y01) were selected (see Supporting Information, **Fig. S2**), outlining the similarity between Me-Lys and proline recognition domains.

Allosteric sites on known epigenetic targets. The primary goal of the computational method described here was to identify potential sites that may be involved in epigenetic recognition of Me-Lys marks on histone tails. Allosteric pockets on known epigenetic targets are of particular interest as they may reveal secondary interaction sites enhancing affinity of histone tails, and may be actual sensors of the histone code. Five interesting pockets that were identified on epigenetic target structures are shown in **Fig. 5.**

The first site corresponds to a potential binding pocket located on the side of the second tandem Tudor domain of TP53BP1 (PDB: 2ig0), and is approximately 23 Å away from the known H4K20Me$_2$ binding site[43] located in the first Tudor domain (**Fig. 5a, Table 3** – 10th ranked by number of Pubmed hits). The second is a site identified on the side of the first MBT domain of L3MBTL3 (PDB: 1wjs),

approximately 14 Å from the center of the MBT aromatic cage (**Fig. 5b, Table 2** – rank 11). Interestingly, the residues forming this allosteric site are conserved in the first and second MBT domains of L3MBTL (PDB: 2pqw). An unexpected aromatic site is also identified on the kinase domain of CHEK1 (PDB: 2e9o), a Ser/Thr protein kinase which, notably, phosphorylates H3T11[44] (**Fig. 5c, Table 3** – 4th ranked by number of Pubmed hits). On the structure of the histone lysine demethylase KDM4A (PDB: 2oq7), which is selective towards H3K9Me_3/Me_2 and H3K36Me_3/Me_2[45], an aromatic site is identified in the catalytic domain approximately 24 Å from the active site (**Fig. 5d, Table 3** – 9th ranked by number of Pubmed hits). Finally, an unexpected aromatic cage is identified at an allosteric site located on the side of the CREBBP bromodomain[46] (PDB: 1jsp), a domain known to recognize acetylated lysines (**Fig. 5e, Table 4**). Using this technology, such aromatic cages can be readily extracted from any set of protein structures, and be suggested for subsequent experimental investigation in order to confirm or disprove their capacity to bind methylated lysines.

CONCLUSION

The approach described here to screen the PDB for specific binding site chemistry is based on tools readily available from computational chemistry suites, and simple scripting language. When applied to a specific structural system, namely Me-Lys binding sites, it could effectively retrieve known readers of the histone code, and identified novel putative sites which may be of interest to the epigenetics research community. Additional applications of this method could include screening the PDB for putative *off*-target binding sites of known drugs, based on pharmacophores extracted from the known target. Additionally, small-molecule ligands co-crystallized to aromatic cages retrieved in the PDB, irrespective of the gene's biological relevance, represent valid chemotypes that can be exploited to design antagonists of Me-Lys binding modules. While the methodology was applied to identifying putative Me-Lys binding sites in the current study, it is meant to be a general purpose screening approach which can easily be adapted to search the PDB for sites possessing various types of pre-defined binding pocket

chemistry, by applying any combination of pharmacophore filters (hydrophobic, aromatic or charged centers, hydrogen bond donors or acceptors) implemented in a number of commercial packages.

ACKNOWLEDGMENT

The authors wish to thank Alexandr Ignachenko, Sigrun Rumpel and Cheryl Arrowsmith. Valérie Campagna-Slater acknowledges the Natural Sciences and Engineering Research Council of Canada for funding. This work was supported by the Structural Genomics Consortium. The SGC is a registered charity (number 1097737) that receives funds from the Canadian Institutes for Health Research, the Canadian Foundation for Innovation, Genome Canada through the Ontario Genomics Institute, GlaxoSmithKline, Karolinska Institutet, the Knut and Alice Wallenberg Foundation, the Ontario Innovation Trust, the Ontario Ministry for Research and Innovation, Merck & Co., Inc., the Novartis Research Foundation, the Swedish Agency for Innovation Systems, the Swedish Foundation for Strategic Research and the Wellcome Trust.

Supporting Information Available. Additional figures are presented as supplementary material: Fig. S1: Kringle domains, Fig. S2: Proline recognition sites. A complete list of the 206 proteins (or protein complexes) in which putative aromatic sites were identified (after filtering, Methods – **Step 4**) is also provided. This material is available free of charge via the Internet at http://pubs.acs.org.

Figure 1. Workflow describing the methodology, using the human chromobox homolog 3 (CBX3) as an example (PDB: 3dm1). As shown by the co-crystal structure in panel **a**, CBX3 binds the trimethylated Lys9 residue of Histone 3 (H3K9). In panels **a** and **b**, aromatic centers are depicted by orange discs, while a blue sphere represents a negative charge.

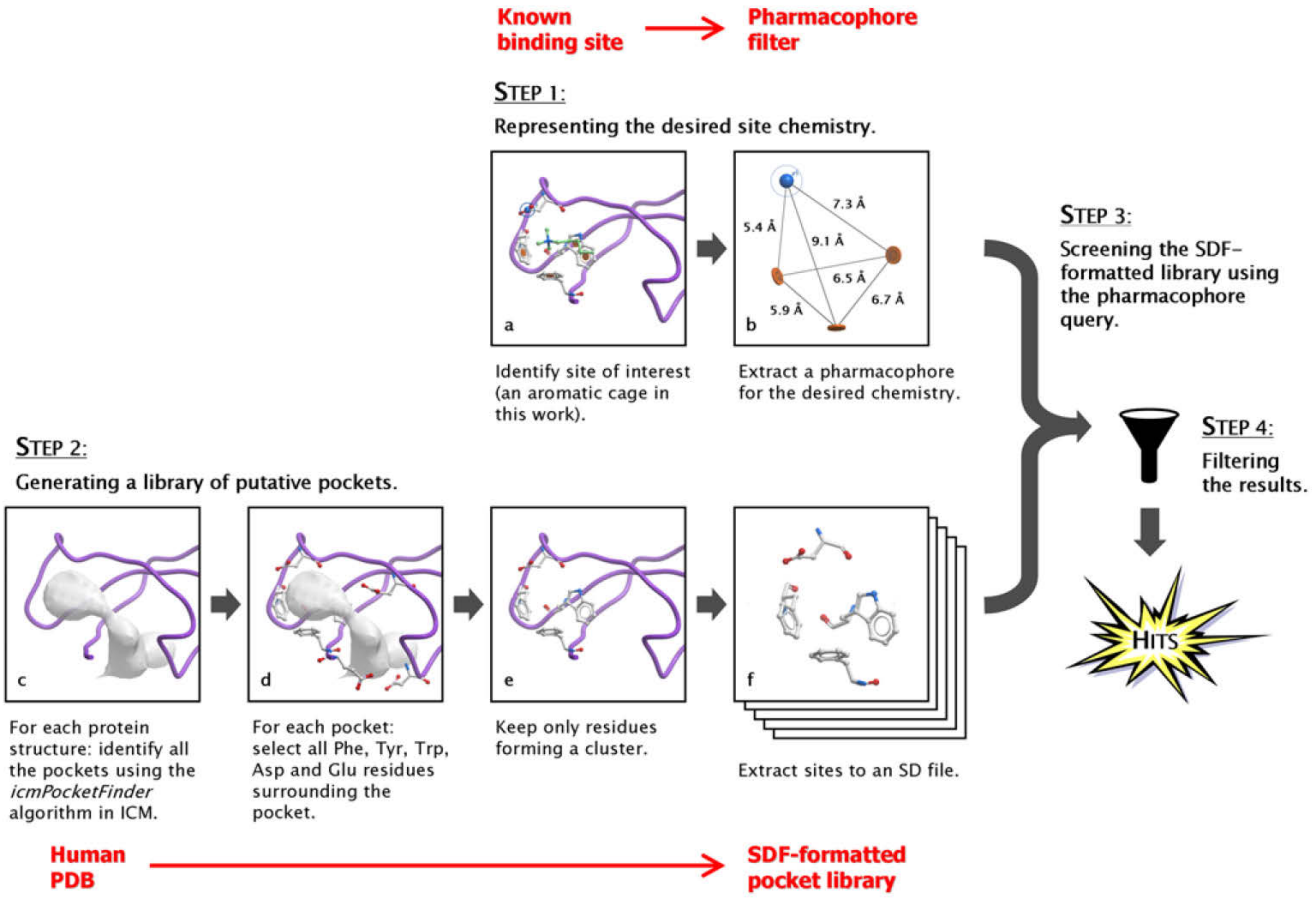


Figure 2. Representative Me-Lys binding sites possessing the desired chemistry: **A.** EHMT1 Ankyrin repeat (PDB: 3b95 – grey), **B.** CBX3 Chromo domain (PDB: 3dm1 – purple), **C.** CHD1 Chromo domain (PDB: 2b2w – pink), **D.** L3MBTL 2nd MBT domain (PDB: 2rje – red), **E.** L3MBTL2 4th MBT domain (PDB: 3dbb – orange), **F.** BPTF PHD domain (PDB: 2fsa – dark green), **G.** PYGO1 PHD domain (PDB: 2vpe – bright green), **H.** KDM4A Tudor domain (PDB: 2qqs – navy blue), **I.** TP53BP1 1st Tudor domain (PDB: 2ig0 – royal blue), **J.** UHRF1 1st Tudor domain (PDB: 3db3 – cyan).

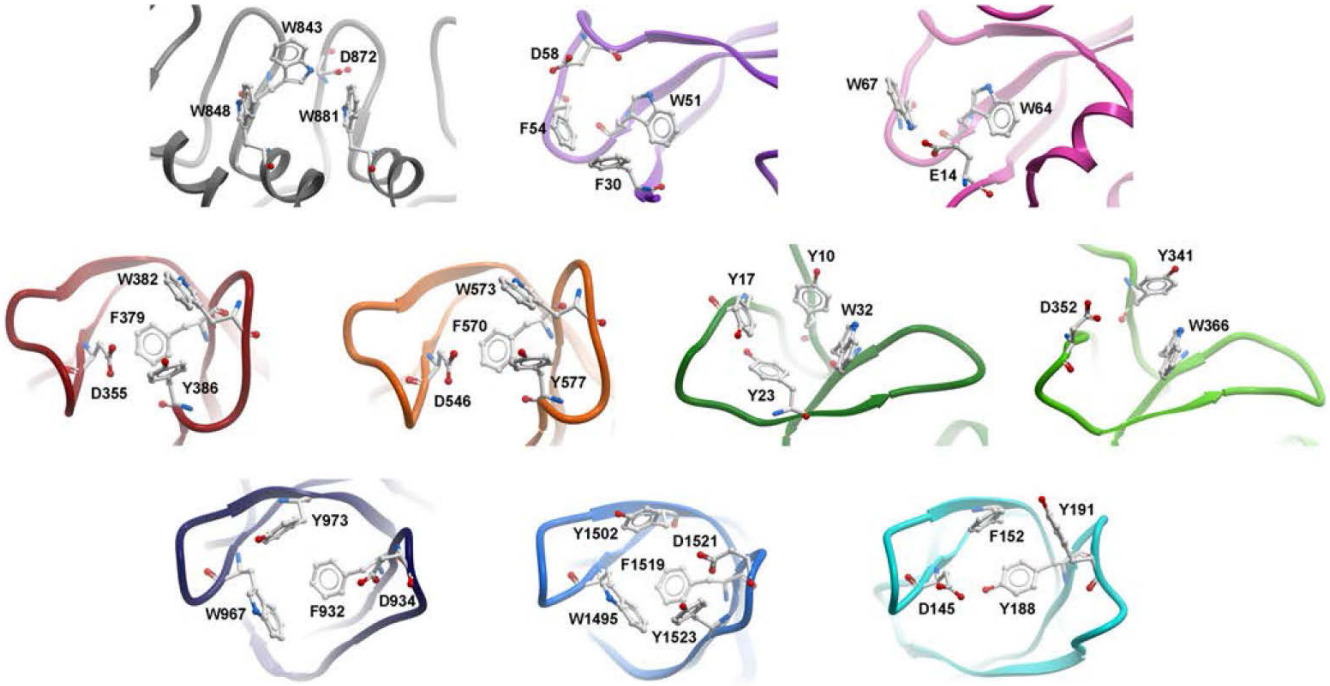


Figure 3. Effect of varying the pharmacophore b-factors on the number of sites retrieved: a) Number of previously known sites selected. b) Number of unexpected sites selected. The pharmacophore size b-factors (Qm/n) were tested at values ranging from 0.1 Å to 4.0 Å. In a first set of calculations, the direction b-factors (Qv) were modified to be the same value as the size b-factors, i.e. Qv=Qm/n (blue triangles), while in the second set they were kept at a constant value of 0.5 Å (red squares). A dashed line is placed at b-factor = 1.3 Å.

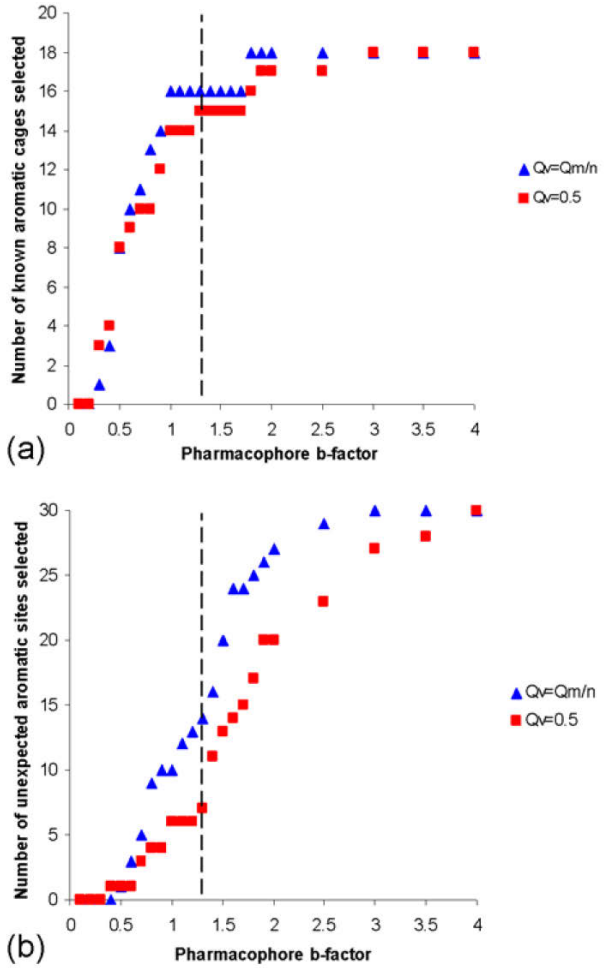


Figure 4. Number of sites selected by each pharmacophore, using a constant direction b-factor (Qv) of 0.5 Å: a) Previously known aromatic cages selected. b) Unexpected aromatic sites selected. (Pharmacophore filters are extracted from **A:** EHMT1 Ankyrin repeat, **B:** CBX3 Chromo domain, **C:** CHD1 Chromo domain, **D:** L3MBTL 2nd MBT domain, **E:** L3MBTL2 4th MBT domain, **F:** BPTF PHD domain, **G:** PYGO1 PHD domain, **H:** KDM4A Tudor domain, **I:** TP53BP1 1st Tudor domain, **J:** UHRF1 1st Tudor domain)

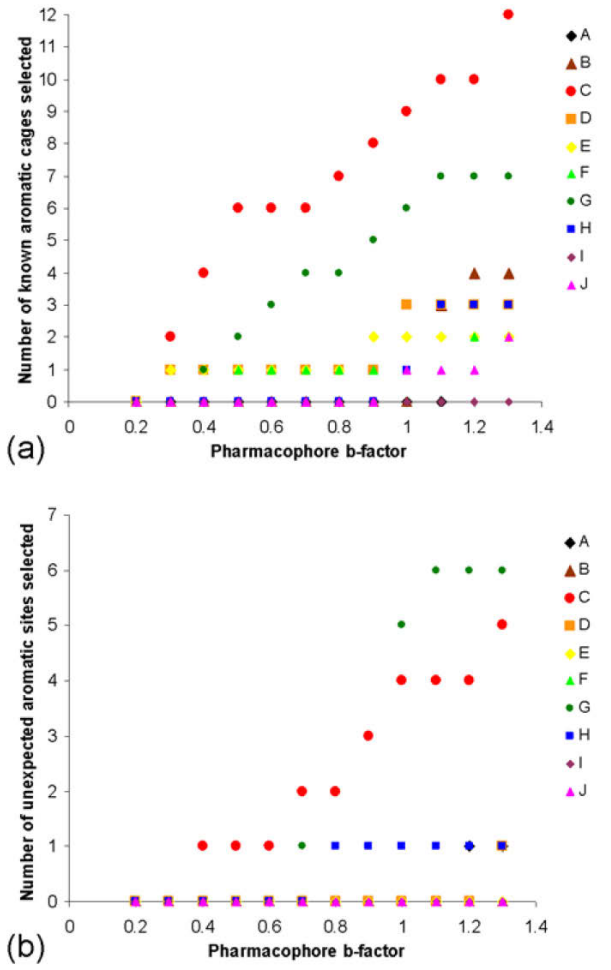


Figure 5. Unexpected aromatic cages identified in various epigenetic targets: a) TP53BP1 tandem Tudor domain (PDB: 2ig0, orange: allosteric site [Glu1564, Tyr1569, and Trp1580], grey: known aromatic cage, light yellow: H4K20Me$_2$ lysine side-chain), b) first MBT domain of L3MBTL3 (PDB: 1wjs, orange: allosteric site [Asp49, Tyr85, Phe90 and Tyr95], grey: known aromatic cage), c) CHEK1 kinase domain (PDB: 2e9o, orange: allosteric site [Trp9, Glu33, Tyr71 and Phe83]), d) KDM4A catalytic domain (PDB: 2oq7, orange: allosteric site [Glu23, Tyr30, Tyr33 and Phe353]), and e) CREBBP bromodomain (PDB: 1jsp, orange: allosteric site [W1158, F1161, W1165, F1185, E1186], light yellow: acetylated K382 of the co-crystallized p53 peptide).

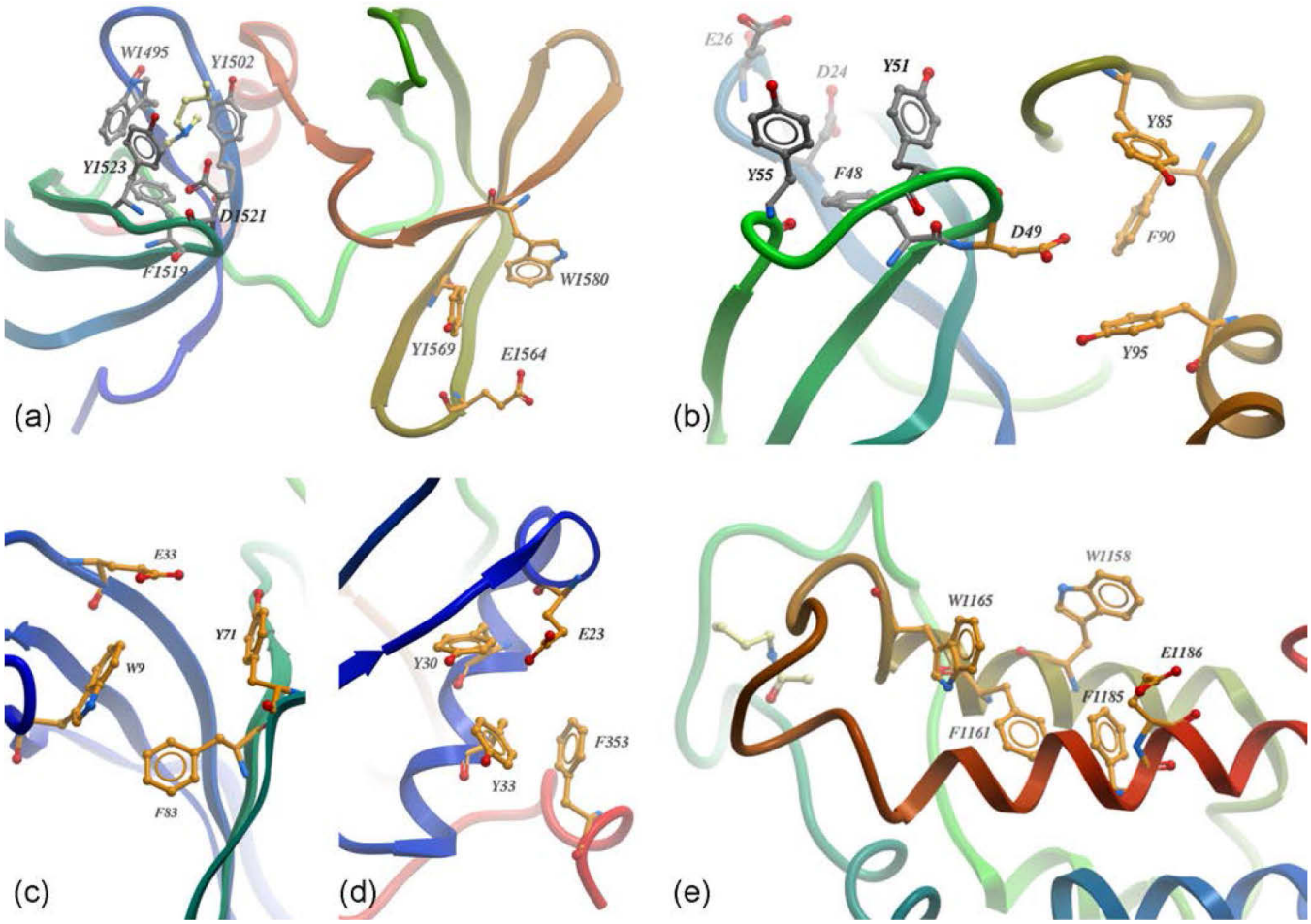


Table 1. List of proteins selected to generate pharmacophores representing the desired receptor chemistry.

Pharmacophore ID	Protein	Domain[a]	PDB codes[b]	Number of aromatic residues	Number of acidic residues
A	EHMT1	Ankyrin	3b95, 3b7b	3	1
B	CBX3	Chromo	3dm1	3	1
C	CHD1	Chromo	2b2t, 2b2u, 2b2v, 2b2w, 2b2y	2	1
D	L3MBTL	MBT	1oyx, 1oz2, 1oz3, 2rhu, 2rhx, 2ri3, 2ri5, 2rjc, 2rjd, 2rje, 2pqw	3	1
E	L3MBTL2	MBT	3dbb, 3cey	3	1
F	BPTF	PHD	2f6j, 2fsa, 2fui, 2fuu	4	0
G	PYGO1	PHD	2vpe	2	1
H	KDM4A	Tudor	2gf7, 2gfa, 2qqr, 2qqs	3	1
I	TP53BP1	Tudor	1xni, 2g3r, 2ig0	4	1
J	UHRF1	Tudor	3db3	3	1

[a] For protein structures possessing multiple repeats of the domain, only the Me-Lys binding one is used to generate a pharmacophore. No PWWP domains were used to generate pharmacophores of the desired site chemistry since none has been co-crystallized with a methylated lysine to date.

[b] For some PDB structures, more than one conformation is used if it contains multiple chains, yielding separate pharmacophores for each conformation, and generating each combination of selected aromatic centers for sites containing tryptophan residues.

Table 2. List of sites extracted from the validation set, ranked according to the calculated distance between the protein and a probe located at the center of the selected residues (from furthest to closest).

Rank	Distance to pocket center (Å)	Protein[a]	PDB code[b]
1	3.10	L3MBTL *(3rd MBT)*	2rjf
2	2.99	TDRD3 *(Tudor)*	2d9t
3	2.85	L3MBTL3 *(3rd MBT)*	1wjq
4	2.84	SCML2 *(2nd MBT)*	2biv
5	2.81	SND1 *(Tudor)*	2hqe
6	2.62	HDGF *(PWWP)*	2b8a
7	2.61	CDYL *(Chromo)*	2dnt
8	2.23	L3MBTL *(1st MBT)*	2rhu
9	2.20	PHF19 *(Tudor)*	2e5q
10	2.06	L3MBTL	1oyx
11	1.89	L3MBTL3	1wjs
12	1.83	L3MBTL	2rjd
13	1.76	CBX1 *(Chromo)*	1ap0
14	1.67	SCML2	2biv
15	1.60	PHF20L1 *(Tudor)*	2eqm
16	1.45	L3MBTL	2rjd
17	1.35	CBX2 *(Chromo)*	2d9u
18	1.35	TDRKH *(Tudor)*	2diq
19	1.31	SETDB1 *(Tudor)*	3dlm
20	1.16	L3MBTL	1oyx
21	0.94	SFMBT2	1wjr
22	0.66	CBX8 *(Chromo)*	2dnv

[a] The domain name is given between brackets when the site corresponds to the conserved aromatic cage. Other sites are unexpected hits: the most interesting ones are indicated with an asterisk.

[b] Some sites were identified in more than one PDB structure.

Table 3. Top 50 hits, based on different ranking criteria.

Rank	Ranked by distance to pocket center (Å)[a,b]	Ranked by distance to pocket center (Å)[a,b] – Nuclear proteins only	Ranked by RMSD with pharmacophore[b,c]	Ranked by RMSD with pharmacophore[b,c] – Nuclear proteins only	Ranked by number of relevant Pubmed hits[d]
1	NT5M	KDM4A *(Tudor)*	PLG	BIRC7	CDK2 (43)
2	KDM4A *(Tudor)*	L3MBTL *(2nd MBT)*	BIRC7	CLIC2	PPARG (32)
3	L3MBTL *(2nd MBT)*	TP53BP1 *(Tudor)*	SNX9	SCML2 *(2nd MBT)*	HBB & HBA1 (30)
4	PYGL	UNG	INSR	SPIN1 *(2nd Tudor-like)*	CHEK1 (22)
5	TP53BP1 *(Tudor)*	NR1H2	HCK	CHEK1	CASP3 (20)
6	UNG	L3MBTL *(3rd MBT)*	AMD1	NCBP2	STAT1 (18)
7	NR1H2	WDR5	ABO	UBE2B	SET (18)
8	L3MBTL *(3rd MBT)*	BPTF *(PHD)*	CLIC2	PHF19	CHEK2 (18)
9	PTPN1	L3MBTL3 *(3rd MBT)*	MMP13 & TIMP2	PIR	KDM4A (15)
10	MMP2	SCML2 *(2nd MBT)*	PLG	EGFR	TP53BP1 (15)
11	FDPS	SND1 *(Tudor)*	FCER1A	L3MBTL *(3rd MBT)*	TNFSF10 (14)
12	ACACB	L3MBTL2 *(4th MBT)*	HEXB	L3MBTL3 *(3rd MBT)*	DDB1 (13)
13	WDR5	EHMT1 *(Ankyrin)*	SCML2 *(2nd MBT)*	NR1I2	AURKA (13)
14	CUTA	MSH2 & MSH6	MET	WDR5	WDR5 (12)
15	BPTF *(PHD)*	DCPS	MASP2	DCK	MSH2 & MSH6 (12)
16	ERAF	PYGO1 *(PHD)*	LPA	KDM4A[e]	CUL4A (12)
17	NUDT2	CDYL *(Chromo)*	F2	PPP2R1A & PPP2CA	POLR2D & POLR2G (12)
18	CEL	UBE2B	GUSB	STAT1	NR1I2 (11)
19	HMGCL	TP53BP1	SPIN1	SND1 *(Tudor)*	CASP8 (10)
20	SH3GL3	CASP8	ITSN2	CDK2	MET (9)
21	NUDT5	DDB1	CHEK1	PPARG	EGFR (9)
22	ARF1	CHD1 *(Chromo)*	PSAP	HMOX1	MMP2 (8)
23	L3MBTL3 *(3rd MBT)*	TNFAIP3	RAP1GAP	CDK6	EHMT1 (8)
24	SCML2 *(2nd MBT)*	ERCC1	NCBP2	PTPN6	CHD1 (8)
25	FCGR2A	CRK	LPA	RHOA	HRAS (8)
26	LCK	PIK3C2A	PTPN11	MSH2 & MSH6	TGFBR2 (8)
27	SND1 *(Tudor)*	CDK6	LCK	CDYL *(Chromo)*	CASP7 (7)
28	L3MBTL2 *(4th MBT)*	PPP2R1A & PPP2CA	LYZ	CDK6	UBE2B (7)
29	HLA-A	STAT1	AOC3	RRM2B	SRF (7)
30	FCER1A	CSTB	DHFR	APPL1	ABL1 (7)
31	NUDT2	ASPA	UBE2B	ASPA	L3MBTL (6)
32	EHMT1 *(Ankyrin)*	SRF	PDE10A	L3MBTL (1st MBT)	CDK6 (6)

33	MSH2 & MSH6	HMOX1	PDE5A	L3MBTL	HLA-DRB1 (6)
34	PRNP	CDK2	EIF4G1	NR1H2	HLA-G (6)
35	CCBL1	LIMK2	HRAS	AURKA	MMP13 & TIMP2 (6)
36	DCPS	L3MBTL *(1st MBT)*	PTGIS	DCPS	CDYL (5)
37	CAMK1G	ABL1	EIF4E2	CRK	POLR2H (5)
38	GAS6	BIRC7	ARHGAP15	POLR2H	RHOA (5)
39	EIF4E2	MSH6	CAMK1G	SET	UNG (4)
40	CASP7	PHF19 *(Tudor)*	NUDT2	ERCC1	BPTF (4)
41	GUSB	XRCC4	CHIT1	DDB1	L3MBTL2 (4)
42	SEC23A	SPIN1 *(1st Tudor-like)*	PIK3CG	LIMK2	PIK3CG (4)
43	AMD1	POLR2H	ANXA5	CSTB	SAT1 (4)
44	NAMPT	NR1I2	NUDT2	XRCC4	DHFR (4)
45	ARHGDIB	SPIN1 *(2nd Tudor-like)*	NAMPT	NME1	APCS (4)
46	PYGO1 *(PHD)*	EGFR	PHF19	POLR2D & POLR2G	PTPN6 (4)
47	PLG	CDK2	PIR	UNG	NR1H2 (3)
48	PSAP	CLIC2	EGFR	PIK3C2A	SND1 (3)
49	CDYL (*Chromo)*	NME1	YES1	TP53BP1	GSN (3)
50	RBP5	CDK6	RAC1	ABL1	NME1 (3), ALB (3), SOS1 (3), THBD (3)

[a] Ranked from largest to lowest distance.

[b] Proteins are listed multiple times only if more than one unique site ranks in the top 50. Some proteins from the validation set are not listed or are ranked differently than in the validation study because one or multiple structures were not included in the human PDB data set for one of two reasons: 1) they originate from species other than human, or 2) they were structures available to the authors which were deposited to the Protein Data Bank after the human PDB list was compiled. A full list of proteins (gene symbols and full names) is provided in the Supporting Information (Table S1).

[c] The RMSD is calculated between the site and the query using the pharmacophoric points. Sites used to generate the pharmacophore queries are omitted from the hit list, since they result in RMSD = 0.0 Å.

[d] The number of Pubmed hits matching at least one keyword (histone, chromatin or epigenetic) is given between brackets. For sites at the interface between subunits of different proteins, the total number of Pubmed hits matching the keywords is calculated as the sum from the individual proteins forming the complex.

[e] This site is located in the catalytic domain of KDM4A, and not in the Tudor domain which was included in the validation set.

Table 4. List of 36 hits selected by visual inspection.

Gene Symbol	Full name
AMD1	adenosylmethionine decarboxylase 1
AOC3	amine oxidase, copper containing 3 (vascular adhesion protein 1)
ARF1	ADP-ribosylation factor 1
ASPA	aspartoacylase (Canavan disease)
CAMK1G	calcium/calmodulin-dependent protein kinase IG
CREBBP[a]	CREB binding protein
CHEK1	CHK1 checkpoint homolog
CHIT1	chitinase 1 (chitotriosidase)
DCPS	decapping enzyme, scavenger
EIF4E2	eukaryotic translation initiation factor 4E family member 2
EIF4G1	eukaryotic translation initiation factor 4 gamma, 1
F10	coagulation factor X
FCER1A	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide
FCGR2A	Fc fragment of IgG, low affinity IIa, receptor (CD32)
FKBP1A	FK506 binding protein 1A, 12 kDa
FKBP1B	FK506 binding protein 1B, 12.6 kDa
HEXB	hexosaminidase B (beta polypeptide)
INSR	insulin receptor
KDM4A[b]	lysine (K)-specific demethylase 4A
L3MBTL3[a,b]	l(3)mbt-like 3
LCK	lymphocyte-specific protein tyrosine kinase
LPA[c]	lipoprotein, Lp(a)
MMP2[c]	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
NAMPT	nicotinamide phosphoribosyltransferase
NCBP2	nuclear cap binding protein subunit 2, 20 kDa
NCF1	neutrophil cytosolic factor 1
NR1H2	nuclear receptor subfamily 1, group H, member 2
NT5M	5',3'-nucleotidase, mitochondrial
PCTP	phosphatidylcholine transfer protein
PLG[c]	plasminogen
SEC23A	Sec23 homolog A
SH3GL3	SH3-domain GRB2-like 3
TGFBR2	transforming growth factor, beta receptor II (70/80 kDa)

TP53BP1[b]	tumor protein p53 binding protein 1
UCK2	uridine-cytidine kinase 2
YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1

[a] These sites were below the selection threshold in Step 4, however they are located in known epigenetic targets and were deemed interesting upon visual inspection.

[b] Allosteric sites on proteins containing Me-Lys binding domains.

[c] More than one unique site identified.

REFERENCES

(1) Kirchmair, J.; Markt, P.; Distinto, S.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Langer, T.; Wolber, G. The Protein Data Bank (PDB), Its Related Services and Software Tools as Key Components for In Silico Guided Drug Discovery. *J. Med. Chem.* **2008**, *51*, 7021-7040.

(2) Henrich, S.; Salo-Ahen, O. M. H.; Huang, B.; Rippmann, F. F.; Cruciani, G.; Wade, R. C. Computational approaches to identifying and characterizing protein binding sites for ligand design. *J. Mol. Recognit.* **2009**, *Published Online: 10 Sep 2009*.

(3) Levitt, D.; Banaszak, L. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* **1992**, *10*, 229-234.

(4) Laskowski, R. SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J. Mol. Graph.* **1995**, *13*, 323-330.

(5) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.* **1998**, *7*, 1884-1897.

(6) Hendlich, M.; Rippmann, F.; Barnickel, G. LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J. Mol. Graph. Model.* **1997**, *15*, 359-363.

(7) Huang, B.; Schroeder, M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct. Biol.* **2006**, *6*, 19.

(8) Brady, G.; Stouten, P. Fast prediction and visualization of protein binding pockets with PASS. *J. Comput. Aided Mol. Des.* **2000**, *14*, 383-401.

(9) Weisel, M.; Proschak, E.; Schneider, G. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem. Cent. J.* **2007**, *1*, 7.

(10) An, J.; Totrov, M.; Abagyan, R. Pocketome via Comprehensive Identification and Classification of Ligand Binding Envelopes. *Mol. Cell. Proteomics* **2005**, *4*, 752-761.

(11) Laurie, A.; Jackson, R. Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics* **2005**, *21*, 1908-1916.

(12) Huang, B. MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omics* **2009**, *13*, 325-330.

(13) Binkowski, T. A.; Naghibzadeh, S.; Liang, J. CASTp: Computed Atlas of Surface Topography of proteins. *Nucleic Acids. Res.* **2003**, *31*, 3352-3355.

(14) Ferrè, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. SURFACE: a database of protein surface regions for functional annotation. *Nucleic Acids. Res.* **2004**, *32*, D240-D244.

(15) Gold, N. D.; Jackson, R. M. A Searchable Database for Comparing Protein-Ligand Binding Sites for the Analysis of Structure-Function Relationships. *J. Chem. Inf. Model.* **2006**, *46*, 736-742.

(16) Lopez, G.; Valencia, A.; Tress, M. FireDB - a database of functionally important residues from proteins of known structure. *Nucleic Acids. Res.* **2007**, *35*, D219-D223.

(17) Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of Protein Active Site Structures for Functional Annotation of Proteins and Drug Design. *Proteins Struct. Funct. Bioinform.* **2006**, *65*, 124-135.

(18) Ferrè, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M. Functional annotation by identification of local surface similarities: a novel tool for structural genomics. *BMC Bioinformatics* **2005**, *6*, 194.

(19) Liu, Z.-P.; Wu, L.-Y.; Wang, Y.; Chen, L.; Zhang, X.-S. Predicting gene ontology functions from protein's regional surface structures. *BMC Bioinformatics* **2007**, *8*, 475.

(20) Glaser, F.; Pupko, T.; Paz, I.; Bell, R. E.; Bechor-Shental, D.; Martz, E.; Ben-Tal, N. ConSurf: identification of functional regions in proteins by surface-mapping. *Bioinformatics* **2003**, *19*, 163-164.

(21) Ausiello, G.; Gherardini, P. F.; Marcatili, P.; Tramontano, A.; Via, A.; Helmer-Citterich, M. FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. *BMC Bioinformatics* **2008**, *9(Suppl 2)*, S2.

(22) Schmitt, S.; Kuhn, D.; Klebe, G. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *J. Mol. Biol.* **2002**, *323*, 387-406.

(23) Kuhn, D.; Weskamp, N.; Schmitt, S.; Hüllermeier, E.; Klebe, G. From the Similarity Analysis of Protein Cavities to the Functional Classification of Protein Families Using Cavbase. *J. Mol. Biol.* **2006**, *359*, 1023-1044.

(24) Jambon, M.; Imberty, A.; Deléage, G.; Geourjon, C. A New Bioinformatic Approach to Detect Common 3D Sites in Protein Structures. *Proteins Struct. Funct. Genet.* **2003**, *52*, 137-145.

(25) Jambon, M.; Andrieu, O.; Combet, C.; Deléage, G.; Delfaud, F.; Geourjon, C. The SuMo server: 3D search for protein functional sites. *Bioinformatics* **2005**, *21*, 3929-3930.

(26) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **2008**, *24*, i105.

(27) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A Common Reference Framework for Analyzing/Comparing Proteins and Ligands. Fingerprints for Ligands And Proteins (FLAP): Theory and Application. *J. Chem. Inf. Model.* **2007**, *47*, 279-294.

(28) Boobbyer, D. N.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure. *J. Med. Chem.* **1989**, *32*, 1083-1094.

(29) Wade, R. C.; Goodford, P. J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 148-156.

(30) Wade, R. C.; Clark, K. J.; Goodford, P. J. Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **1993**, *36*, 140-147.

(31) *ICM 3.6-1*: Molsoft LLC: San Diego, CA.

(32) Abagyan, R. ICM Manual v.3.6; Molsoft LLC: San Diego, 2008.

(33) Taverna, S. D.; Li, H.; Ruthenburg, A. J.; Allis, C. D.; Patel, D. J. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat. Struct. Mol. Biol.* **2007**, *14*, 1025-1040.

(34) Collins, R. E.; Northrop, J. P.; Horton, J. R.; Lee, D. Y.; Zhang, X.; Stallcup, M. R.; Cheng, X. The ankyrin repeats of G9a and GLP histone methyltransferases are mono- and dimethyllysine binding modules. *Nat. Struct. Mol. Biol.* **2008**, *15*, 245-250.

(35) Lan, F.; Collins, R. E.; Cegli, R. D.; Alpatov, R.; Horton, J. R.; Shi, X.; Gozani, O.; Cheng, X.; Shi, Y. Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature* **2007**, *448*, 718-723.

(36) Wang, W. K.; Tereshko, V.; Boccuni, P.; MacGrogan, D.; Nimer, S. D.; Patel, D. J. Malignant Brain Tumor Repeats: A Three-Leaved Propeller Architecture with Ligand/Peptide Binding Pockets. *Structure* **2003**, *11*, 775-789.

(37) Marti, D. N.; Schaller, J.; Llinás, M. Solution Structure and Dynamics of the Plasminogen Kringle 2-AMCHA Complex: 3_1-Helix in Homologous Domains. *Biochemistry* **1999**, *38*, 15741-15755.

(38) Mochalkin, I.; Cheng, B.; Klezovitch, O.; Scanu, A. M.; Tulinsky, A. Recombinant Kringle IV-10 Modules of Human Apolipoprotein(a): Structure, Ligand Binding Modes, and Biological Relevance. *Biochemistry* **1999**, *38*, 1990-1998.

(39) Couture, J.-F.; Collazo, E.; Trievel, R. C. Molecular recognition of histone H3 by the WD40 protein WDR5. *Nat. Struct. Mol. Biol.* **2006**, *13*, 698-703.

(40) Groemping, Y.; Lapouge, K.; Smerdon, S. J.; Rittinger, K. Molecular Basis of Phosphorylation-Induced Activation of the NADPH Oxidase. *Cell* **2003**, *113*, 343-355.

(41) Kang, C. B.; Ye, H.; Dhe-Paganon, S.; Yoon, H. S. FKBP family proteins: immunophilins with versatile biological functions. *Neurosignals* **2008**, *16*, 318-325.

(42) Feng, L.; Gell, D. A.; Zhou, S.; Gu, L.; Kong, Y.; Li, J.; Hu, M.; Yan, N.; Lee, C.; Rich, A. M.; Armstrong, R. S.; Lay, P. A.; Gow, A. J.; Weiss, M. J.; Mackay, J. P.; Shi, Y. Molecular Mechanism of AHSP-Mediated Stabilization of α-Hemoglobin. *Cell* **2004**, *119*, 629-640.

(43) Botuyan, M. V.; Lee, J.; Ward, I. M.; Kim, J.-E.; Thompson, J. R.; Chen, J.; Mer, G. Structural Basis for the Methylation State-Specific Recognition of Histone H4-K20 by 53BP1 and Crb2 in DNA Repair. *Cell* **2006**, *127*, 1361-1373.

(44) Shimada, M.; Niida, H.; Zineldeen, D. H.; Tagami, H.; Tanaka, M.; Saito, H.; Nakanishi, M. Chk1 Is a Histone H3 Threonine 11 Kinase that Regulates DNA Damage-Induced Transcriptional Repression. *Cell* **2008**, *132*, 221-232.

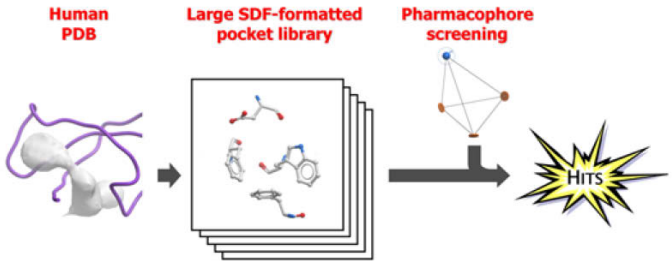
(45) Ng, S. S.; Kavanagh, K. L.; McDonough, M. A.; Butler, D.; Pilka, E. S.; Lienard, B. M. R.; Bray, J. E.; Savitsky, P.; Gileadi, O.; Delft, F. v.; Rose, N. R.; Offer, J.; Scheinost, J. C.; Borowski, T.; Sundstrom, M.; Schofield, C. J.; Oppermann, U. Crystal structures of histone demethylase JMJD2A reveal basis for substrate specificity. *Nature* **2007**, *448*, 87-92.

(46) Mujtaba, S.; He, Y.; Zeng, L.; Yan, S.; Plotnikova, O.; Sachchidanand; Sanchez, R.; Zeleznik-Le, N.; Ronai, Z.; Zhou, M. Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation. *Mol. Cell* **2004**, *13*, 251-263.

SYNOPSIS TOC

Pharmacophore screening of the Protein Data Bank for specific binding site chemistry

Valérie Campagna-Slater, Andrew G. Arrowsmith, Yong Zhao and Matthieu Schapira



Supporting Information.

Pharmacophore screening of the Protein Data Bank for specific binding site chemistry

Valérie Campagna-Slater[1], *Andrew G. Arrowsmith*[1], *Yong Zhao*[1] *and Matthieu Schapira*[1,2]

[1]Structural Genomics Consortium, University of Toronto, MaRS Centre, South Tower, 7th floor, 101 College Street, Toronto, Ontario, Canada, M5G 1L7

[2]Department of Pharmacology and Toxicology, University of Toronto, Medical Sciences Building, 1 King's College Circle, Toronto, Ontario, Canada, M5S 1A8

Figure S1. Two of the Kringle domains identified as possessing a similar chemistry and geometry to known Me-Lys readers: a) PLG Kringle 2 domain (PDB: 1b2i), and b) LPA Kringle IV-10 domain (PDB: 3kiv). Bound ligands are shown in light yellow. (additional residues not extracted to the virtual library are shown in grey)

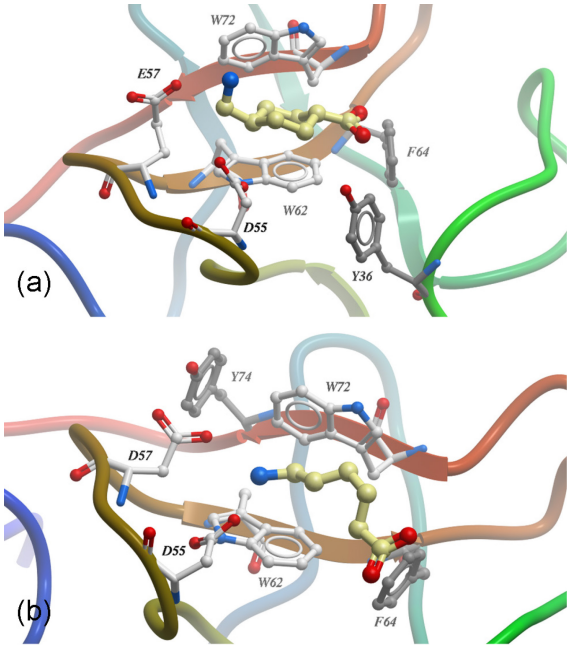


Figure S2. The chemistry and geometry of some proline recognition sites is similar to that of Me-Lys readers: a) FKBP1A (additional residues not extracted to the virtual library are shown in grey), and b) ERAF.

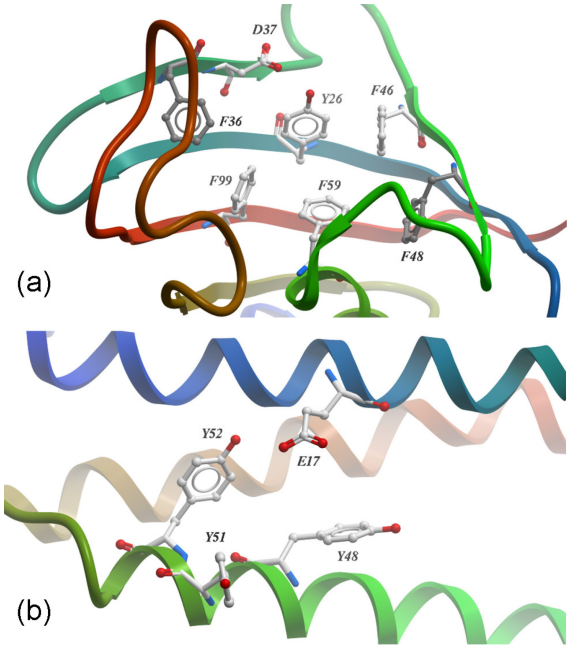


Table S1. List of 206 proteins (or protein complexes) in which putative aromatic sites were identified (after filtering, Methods – **Step 4**).

Gene Symbol	Full name
ABL1	c-abl oncogene 1, receptor tyrosine kinase
ABO	ABO blood group (alpha 1-3-N-acetylgalactosaminyltransferase, alpha
ACACB	acetyl-Coenzyme A carboxylase beta
ACE	angiotensin I converting enzyme 1
AKR1B1	aldo-keto reductase family 1, member B1
ALB	albumin
AMD1	adenosylmethionine decarboxylase 1
ANXA5	annexin 5
AOC3	amine oxidase, copper containing 3
APCS	serum amyloid P component
APPL1	adaptor protein, phosphotyrosine interaction, PH domain and leucine
ARF1	ADP-ribosylation factor 1
ARHGAP15	Rho GTPase activating protein 15
ARHGAP5	Rho GTPase activating protein 5
ARHGDIA	Rho GDP dissociation inhibitor (GDI) alpha
ARHGDIB	Rho GDP dissociation inhibitor (GDI) beta
ASL	argininosuccinate lyase
ASPA	aspartoacylase
ASS1	argininosuccinate synthetase 1
AURKA	serine/threonine protein kinase 6
BIRC7	livin inhibitor of apoptosis
BIRC8	baculoviral IAP repeat-containing 8
BPTF	bromodomain PHD finger transcription factor
C14orf129	chromosome 14 open reading frame 129
C2	complement component 2
C5	complement component 5
CAMK1G	calcium/calmodulin-dependent protein kinase IG
CASP3	caspase 3
CASP7	caspase 7
CASP8	caspase 8
CCBL1	kynurenine aminotransferase I
CD1D	CD1D antigen
CDK2	cyclin-dependent kinase 2
CDK6	cyclin-dependent kinase 6
CDK7	cyclin-dependent kinase 7
CDYL	chromodomain protein, Y chromosome-like
CEL	carboxyl ester lipase
CES1	carboxylesterase 1
CHD1	chromodomain helicase DNA binding protein 1

CHEK1	checkpoint kinase 1
CHEK2	protein kinase CHK2
CHIT1	chitotriosidase
CLIC2	chloride intracellular channel 2
CRK	v-crk sarcoma virus CT10 oncogene homolog
CRYBB1	crystallin, beta B1
CST5	cystatin D
CSTB	cystatin B
CTSK	cathepsin K
CUL4A	cullin 4A
CUTA	cutA divalent cation tolerance homolog
CYP2A6	cytochrome P450, family 2, subfamily A, polypeptide 6
CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8
DAO	D-amino-acid oxidase
DCK	deoxycytidine kinase
DCPS	mRNA decapping enzyme
DDB1	damage-specific DNA binding protein 1
DDX58	DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide RIG-I
DECR1	2,4-dienoyl CoA reductase 1
DHFR	dihydrofolate reductase
DPP4	dipeptidylpeptidase IV
DPP6	dipeptidyl-peptidase 6
EGFR	epidermal growth factor receptor
EHMT1	euchromatic histone-lysine N-methyltransferase 1
EIF3B	eukaryotic translation initiation factor 3, subunit 9 eta, 116kDa
EIF4E2	eukaryotic translation initiation factor 4E member 2
EIF4G1	eukaryotic translation initiation factor 4 gamma, 1
EPOR	erythropoietin receptor
ERAF	erythroid associated factor
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4
ERCC1	excision repair cross-complementing 1
F10	coagulation factor X
F2	coagulation factor II
F8	coagulation factor VIII
F9	coagulation factor IX
FCER1A	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide
FCGR2A	Fc fragment of IgG, low affinity IIa, receptor
FCGRT	Fc fragment of IgG, receptor, transporter, alpha
FDPS	farnesyl diphosphate synthase
FIS1	tetratricopeptide repeat domain 11
FKBP1A	FK506 binding protein 1A, 12kDa
FKBP1B	FK506 binding protein 1B, 12.6 kDa
FOLH1	folate hydrolase 1
GAS6	growth arrest-specific 6

GCDH	glutaryl-Coenzyme A dehydrogenase
GRHPR	glyoxylate reductase/hydroxypyruvate reductase
GSN	gelsolin
GUSB	glucuronidase, beta
HBB & HBA1	beta globin & alpha 1 globin
HCK	hemopoietic cell kinase
HEXB	hexosaminidase B
HLA-A	major histocompatibility complex, class I, A
HLA-B	major histocompatibility complex, class I, B
HLA-DMA & HLA-DMB	major histocompatibility complex, class II, DM alpha & major histocompatibility complex, class II, DM beta
HLA-DRB1	major histocompatibility complex, class II, DR beta 1
HLA-G	major histocompatibility complex, class I, G
HMGCL	3-hydroxy-3-methylglutaryl CoA lyase
HMOX1	heme oxygenase (decyclizing) 1
HMOX2	heme oxygenase (decyclizing) 2
HRAS	v-Ha-ras Harvey rat sarcoma viral oncogene homolog
HSD11B1	11-beta-hydroxysteroid dehydrogenase 1
IL6R	interleukin 6 receptor
INSR	insulin receptor
IRAK4	interleukin-1 receptor-associated kinase 4
ITGAV	integrin alpha-V
ITK	IL2-inducible T-cell kinase
ITSN2	intersectin 2
KDM4A	jumonji domain containing 2A
KHK	ketohexokinase
L3MBTL	l(3)mbt-like
L3MBTL2	l(3)mbt-like 2
L3MBTL3	l(3)mbt-like 3
LCK	lymphocyte-specific protein tyrosine kinase
LIMK2	LIM domain kinase 2
LOC100133811	N/A
LPA	lipoprotein Lp(a)
LSS	lanosterol synthase
LTF	lactotransferrin
LYZ	lysozyme
MAOB	monoamine oxidase B
MAP1LC3B	microtubule-associated proteins 1A/1B light chain 3
MASP2	mannan-binding lectin serine protease 2
MET	met proto-oncogene
MMP13 & TIMP2	matrix metalloproteinase 13 & TIMP metallopeptidase inhibitor 2
MMP2	matrix metalloproteinase 2
MSH2 & MSH6	mutS homolog 2 & mutS homolog 6
MSH6	mutS homolog 6

MYO5A	myosin VA
NAMPT	nicotinamide phosphoribosyltransferase
NCBP2	nuclear cap binding protein subunit 2, 20kDa
NCF1	neutrophil cytosolic factor 1
NME1	non-metastatic cells 1, protein (NM23A) expressed in
NR1H2	nuclear receptor subfamily 1, group H, member 2
NR1I2	nuclear receptor subfamily 1, group I, member 2
NT5C2	5'-nucleotidase, cytosolic II
NT5M	5',3'-nucleotidase, mitochondrial
NUDT2	nudix-type motif 2
NUDT5	nudix-type motif 5
P4HA1	prolyl 4-hydroxylase, alpha I subunit
PAPSS1	3'-phosphoadenosine 5'-phosphosulfate synthase 1
PC	pyruvate carboxylase
PCTP	phosphatidylcholine transfer protein
PDE10A	phosphodiesterase 10A
PDE5A	phosphodiesterase 5A
PECI	peroxisomal D3,D2-enoyl-CoA isomerase
PHF19	PHD finger protein 19
PIK3C2A	phosphoinositide-3-kinase, class 2 alpha polypeptide
PIK3CG	phosphoinositide-3-kinase, catalytic, gamma polypeptide
PIR	pirin
PKLR	pyruvate kinase, liver and RBC
PLG	plasminogen
PNLIPRP2	pancreatic lipase-related protein 2
POLG2	DNA polymerase subunit gamma-2, mitochondrial
POLR2D & POLR2G	DNA directed RNA polymerase II polypeptide D & DNA directed RNA polymerase II polypeptide G
POLR2H	RNA polymerase II, polypeptide H
PPARG	peroxisome proliferative activated receptor gamma
PPCDC	phosphopantothenoylcysteine decarboxylase
PPP2R1A & PPP2CA	alpha isoform of regulatory subunit A, protein phosphatase 2 & protein phosphatase 2, catalytic subunit, alpha isoform
PRNP	prion protein
PSAP	prosaposin
PTGES3	prostaglandin-E synthase 3
PTGIS	prostaglandin I2 synthase
PTPN1	protein tyrosine phosphatase, non-receptor type 1
PTPN11	protein tyrosine phosphatase, non-receptor type 11
PTPN22	protein tyrosine phosphatase, non-receptor type 22 (lymphoid)
PTPN6	protein tyrosine phosphatase, non-receptor type 6
PTPRO	receptor-type protein tyrosine phosphatase O
PYGL	liver glycogen phosphorylase
PYGO1	pygopus homolog 1

RAB11A	Ras-related protein Rab-11A
RAB27B	RAB27B, member RAS oncogene family
RAC1	ras-related C3 botulinum toxin substrate 1
RAP1GAP	RAP1 GTPase activating protein
RBP4	retinol-binding protein 4, plasma
RBP5	retinol binding protein 5, cellular
RHOA	ras homolog gene family, member A
RRM2B	ribonucleotide reductase M2 B (TP53 inducible)
SAT1	spermidine/spermine N1-acetyltransferase 1
SCML2	sex comb on midleg-like 2
SEC13	SEC13 protein
SEC23A	SEC23-related protein A
SERPINC1	serpin peptidase inhibitor, clade C, member 1
SET	SET translocation (myeloid leukemia-associated)
SH3GL3	SH3-domain GRB2-like 3
SLC4A1	solute carrier family 4, anion exchanger, member 1
SND1	staphylococcal nuclease domain containing 1
SNX9	sorting nexin 9
SOS1	son of sevenless homolog 1
SPIN1	spindlin
SRF	serum response factor (c-fos serum response element-binding
STAT1	signal transducer and activator of transcription 1
TF	transferrin
TGFBR2	transforming growth factor, beta receptor II
THBD	thrombomodulin
TNFAIP3	tumor necrosis factor, alpha-induced protein 3
TNFSF10	tumor necrosis factor (ligand) superfamily, member 10
TP53BP1	tumor protein p53 binding protein 1
TRAF2	TNF receptor-associated factor 2
TTK	TTK protein kinase
UBE2B	ubiquitin-conjugating enzyme E2B
UCK1	uridine-cytidine kinase 1
UCK2	uridine-cytidine kinase 2
UNG	uracil-DNA glycosylase
USP14	ubiquitin specific protease 14
WDR5	WD repeat domain 5
XRCC4	X-ray repair cross complementing protein 4
YES1	viral oncogene yes-1 homolog 1