Title
Most protein domains exist as multiple variants with distinct structure and function

Authors
Kristoffer Vitting-Seerup[1]

Affiliations
[1]The Bioinformatics Section, Department of Health Technology, The Technical University of Denmark (DTU), Denmark.

Contact
k.vitting.seerup@gmail.com

Brief unreferenced abstract

Protein domains are the active subunits of proteins that provide proteins with specific functions by adopting specific three-dimensional structures. These protein domains are currently described and analyzed as invariable molecular building blocks with fixed functions. Contradictory to that belief, we here show that most human protein domains exist as multiple distinct variants that, compared to each other, modulate, abolish, or enable new functionality of the domain.

Main text

Introduction

Most proteins have, throughout evolution, been created using protein domains as molecular building blocks[1]. The importance of these protein domains is hard to overstate as they encode the core functionality needed for most, if not all, cellular functions. Noteworthy examples include signal transduction[2] and the ability of proteins to bind to DNA, RNA, and other proteins[3]. Current state-of-the-art annotation of protein domains (e.g., Pfam[4]) relies on encapsulating known domain sequences into a reference model and matching those models to a sequence of interest. These tools are used in many types of analysis and range from mechanistic studies of proteins to interpreting the impact of mutations and are cited by thousands of papers each year.

While protein domains are defined, described, and analyzed as protein subunits with fixed functionality[4,5] that might not be the case. Numerous mutational and cancer studies find that

removing just a tiny part of a protein domain can either eliminate or modify domain function[6–8]. Such findings lead to an intriguing hypothesis: just like most genes produce protein isoforms with distinct functions, naturally occurring protein domain variants could exist. Such domain variants would originate either from alternative splicing or evolution, resulting in different genes having different isotypes of the same protein domain. To collectively describe such protein domain variants, we use the term "domain isotypes".

Domain isotypes exist in humans

To investigate if alternative splicing creates domain isotypes, we turned to cancer data. Specifically, we re-analyzed the isoform switches obtained from comparing tumors to adjacent healthy tissue[9] for partial domain loss/gain. Across the 12 cancer types analyzed, we identified 492 genes where an isoform switch resulted in substantial changes to, but not complete gain/removal of, the protein domain (Figure S1A-1, Table S1). Notably, 204 of these genes were found in more than one cancer type (Figure S1A-2), highlighting their functional importance. A prominent example is the isoform switch in the tumor suppressor ABI1 that inhibit the oncogene ABL1[10]. Accordingly, low expression of ABI1 is associated with metastasis and poor survival in many cancer types[11]. Intriguingly the ABI1 isoform switch (Figure 1A, S1A-3) results in increased usage of an isoform with a disrupted "Abi_HHR" domain isotype. Since the Abi_HRR domain is responsible for inhibiting the ABL1 oncogene, this switch results in cell transformation and worse patient outcome[12]. Thus, splicing can produce domain isotypes that are functionally important.

Encouraged by these disease-associated examples, we next asked if domain isotypes also exist in non-disease states. Since the difference between the domain isotypes in the ABI gene is "just" 10 amino acids, we systematically analyzed the length of protein domains identified across the manually curated Swiss-Prot canonical database (11076 domain-contain proteins, one isoform per gene analysis)[13]. We found substantial length differences when comparing protein domains from different genes (Figure 1B). Intriguingly many protein domains seem to follow individual patterns in how they deviate from the reference length (Figure S1C-1). Specifically, we find that the length distribution of 58.3% of protein domains contains two or more clusters (Figure 1C, S1C-2), suggesting domain isotypes are frequently created and used through evolution.

Enabling systematic analysis of domain isotypes

Encouraged by this, we developed pfamAnalyzeR, which enables detection and in-depth analysis of domain isotypes (available at https://github.com/kvittingseerup/pfamAnalyzeR, in the prosses of being submitted to Bioconductor). pfamAnalyzeR uses a highly stringent approach to detect domain changes (Figure S1D-1) and can distinguish between 5 domain isotypes (Figure 1D). These isotypes are the reference isotype and four isotypes that, compared to the reference isotype, are best described as a truncation, an insertion, a deletion, or combinations thereof ("complex")(See schematic illustration in Figure 1D).

Domain isotypes are omnipresent

We used pfamAnalyzeR to analyze all protein domains in the manually curated Swiss-Prot database of multiple species[13]. Intriguingly we find that most protein domains exist as both reference and non-reference iso-types (Figure 1E, S1E-1, Table S2), with higher frequencies found in proteins manually annotated as "domain containing" by UniProt (Figure S1E-2). To ensure the biological relevance of domain isotypes, we examined the expression of mRNA transcripts containing either reference or non-reference domain isotypes across the 55 human tissues in the GTEx data[14]. We find that reference and non-reference domain isotypes exist within expressed transcripts at almost the same rate across expression thresholds (Figure 1F), suggesting they are frequently used in humans. Accordingly, substantial fractions of non-reference isotype domains are even found when only considering the most expressed (major) transcripts from each gene (Figure S1F-1). Interestingly within the major transcripts, where there are changes in which transcript is the most expressed in different tissues, we find 935 (26.7%) protein domains where the domain isotype also changes across tissues (Figure 1G, S1F-1). Similarly, we find 94 (21.6%) domains with isotype switches across a small selection of 9 human cell types profiled at single cell resolution[15] (Figure 1G). Thus, protein domain isotypes are not an exception but the norm and seem to be used in a cell, tissue, and disease-specific manner.

Domain isotypes have distinct 3D structures

Having shown that domain isotypes are pervasive, we next asked: what is the biological function of domain isotypes? Following the "structure is function" axiom[16], we started with the three-dimensional (3D) structure. We analyzed all protein domains within the experimentally solved structures of human proteins in Protein Data Bank (PDB)[17] (Table S3). For each protein domain, we compared the 3D structure of reference domain isotypes to both reference and non-reference isotypes (as illustrated by arrows in Figure 1D). An interesting example is the

Histidine phosphatase domain "His_Phos_1". This domain facilitates the removal of phosphorylations via an essential histidine residue ("His_8") aided by a few other residues, including "His_{108}"[18]. While the "His_Phos_1" protein domains were virtually identical to each other within isotype category (mean $RMSD_{aligned\text{-}backbone}$ <1), the reference and truncated isotypes were strikingly different (mean $RMSD_{aligned\text{-}backbone}$ >11)(Figure 2A, S2A-1). This difference is exacerbated by the absence of the conserved catalytical residues "His_{108}"[18] from the truncated isotype (shown in blue in the reference isotype domains of Figure 2A, S2A-1). Importantly this is not just an artifact of the domain boundary, as the residue is missing even when considering an additional 50 amino acids downstream of the domain boundary (Figure S2A-2). The absence of a conserved catalytical residue naturally implies that the truncated isotype modifies the functionality of the domain[19]. Encouraged by this example, we summarized our results for the structural differences of domain isotypes across the 952 protein domains analyzed (Table S4). We find the 3D structures of different domain isotypes are remarkably different (Figure 2B, S2B-1)(Median $RMSD_{aligned\text{-}backbone}$ >3x higher than within reference isotype comparison, P = 2.14e-102, Wilcoxon tests). Surprisingly specific domain isotypes are even associated with larger structural differences that approach what is expected when comparing random domains (shaded area in Figure 2B, Table S5). In summary, the 3D structure of domain isotypes is surprisingly different, thus challenging the current view of domains as identical units.

Domain isotypes affect biological function

Since interaction between proteins and other macromolecules (e.g., protein, DNA, RNA) requires a specific 3D structure, the structural differences of domain isoforms could modify such interactions (both positively and negatively). In support of this, we find that non-reference protein domain isotypes are highly enriched for protein domains that facilitate interactions with DNA, RNA, and proteins[3] (over odds ratio per interaction type > 1.93, p-value < 3.23e-10) (Figure 2C)(Table S6). Indeed, by analyzing the specific amino acid motif that interacts with protein domains in solved 3D structures[3], we find at least 10 cases where domain isotypes were significantly associated with specific clusters of binding motifs (Methods). One example is the trypsin binding domain, where the reference and truncation isotype bind two distinct motif clusters with extreme specificity (FDR = 8.68e-18, fisher test) (Figure 2D). Thus, protein domain isotypes seem important for determining the interaction between proteins and other macromolecules. The importance of this is further emphasized when only considering protein isoforms within the same gene. Indeed, the human genome contains at least 7735 protein-

coding genes (38.71%) that encodes protein isoforms, which differ in the presence, absence or isotype of interaction domains (Figure S2D-1, S2D-2). Thus, protein domain isotypes also modulate the function of individual genes through alternative transcript usage.

The structure is function axiom also suggest that protein domain isotypes could have distinct biological functions. We, therefore, analyzed if individual protein domains were overrepresented in the genes annotated to specific pathways or go-terms. This resulted in a significant association between 552 protein domains and 6018 gene-sets (FDR < 0.05, Table S7), suggesting these domains are crucial for the pathways. Next, we tested whether the significant gene-sets were preferentially enriched for reference or non-reference domain isotypes (see methods). We found that 77 protein domains had at least one gene-set significantly skewed towards an isotype (Figure 2E, Table S5), suggesting that a particular domain isotype was needed in that pathway. One of these is the Ras protein domain, which is significantly associated with 198 gene-sets, of which 20 were significantly skewed towards a domain isotype (Figure 2F). Interestingly the genes with the reference isotype of the Ras domain seem to be preferentially involved in vesicle formation and release, whereas the genes containing the non-reference isotypes were associated with cell movement (Figure 2F). Thus domain isotypes facilitate different biological functions.

Next, we utilized that some protein domains have already been associated with human diseases through overrepresentation of either mutations or naturally occurring disease-related genetic variations (SNPs)[20,21]. We find that non-reference protein domain isotypes are highly enriched for protein domains associated with all 17 disease groups tested (Figure S2G-1), with prominent examples being cardiovascular diseases, mental disorders, immune pathologies, and cancer (all having odds ratio > 4, P-value < 2.33e-07)(Figure 2G). Importantly this suggests that many non-reference domain isotypes are required for homeostasis and, when disrupted, can lead to a wide range of human diseases.

Discussion

In summary, we developed pfamAnalyzeR and used it to show that protein domain isotypes are a pervasive mechanism that modulates, disrupts, and enables new functionality of protein domains across cell, tissue, and disease states. Unfortunately, we think the results presented here underestimate the true diversity and importance of domain isotypes. Long-read technology consistently demonstrates that the current databases underrepresent the protein diversity[22,23].

This also means domain isotypes are probably underrepresented in the data analyzed here. Additionally, naturally occurring genetic variation (SNPs) affect splicing in most human protein-coding genes[14], probably resulting in domain isotype changes. Jointly this indicates that our results underestimate the true diversity and importance of domain isotypes. It is therefore also clear that considerable additional computational and experimental work is needed to characterize the functional role of domain isotypes, both as a group and for individual protein domains

This article challenges the current perception of protein domains as building blocks with a fixed function. Instead, our findings suggest that protein domain isotypes are used to modulate domain function, thereby increasing the flexibility of protein adaptation. These findings also suggest caution when transferring concussions about a protein domain from one setting to another (both other proteins and cell/tissues) as it is likely that different domain isotypes are used, whereby functional differences are expected. The widespread use of protein domain isoforms also highlights, along with many other considerations[9,14,24,25], the need to change how we think about molecular biology. Although it will take a lot of work, we need to progress from the current "gene-centric" research paradigm to a more nuanced "isoform-centric" paradigm.

Figure text

Figure 1: Domain isotypes are omnipresent. **A)** The isoform switch in ABI1 identified by comparing Lung Squamous Cell Carcinoma (LUSC) to adjacent healthy tissue. Boxes represent exons; where the higher sections annotate, the coding regions and protein domains are indicated by colored. The isoform usage is indicated on the left side. Note the differences in the Abi_HRR domain. **B)** The difference between the actual length of protein domains and their corresponding reference length across all human domains identified in SwissProt. **C)** The number of clusters found in the length distribution of each human protein domain. **D)** Schematic representation of the five different protein domain isotypes identified. Black boxes represent the reference model, and red/turquoise boxes indicate the sequence analyzed (red being reference isotype and turquoise being non-reference isotypes). Comparisons analyzed in this paper are indicated by arrows, and the collection of non-reference isotypes are highlighted to the left. **E)** The fraction of protein domains that exist as reference and non-reference isotypes in SwissProt of various species. **F)** The fraction (y-axis) of protein domain isotypes (color)

contained in transcripts expressed above a cutoff (x-axis) in the average expression profile of at least one human tissue. **G)** The most expressed transcript was identified from each gene in each human tissue, and genes where the major transcript was different in tissues or cell types (x-axis) were extracted. From these transcripts, we calculated the percent of encoded protein domains (Y-axis) where there was a domain isotype-switch due to the transcript switch. Numbers on bars indicate the actual number of domains identified.

Figure 2: Domain isotypes are functionally important. A) Structural alignment of three references (grey) and two truncated (red) isotypes of the His_Phos_1 protein domain. The catalytic residue His108 is indicated in blue. **B)** For each protein domain, the 3D structures of reference isotypes were compared to the structure of all identified isotypes (x-axis). The structural difference was quantified RMSD and averaged per domain (Y-axis) (higher means more different). The grey area indicates the RMSD expected for random domains by showing the area where the middle 90% of random comparisons fall. **C)** The enrichment (y-axis) of domains involved in interactions (color) for each domain isotype (x-axis). Triangles indicate FDR < 0.05. **D)** The number of interactions (color) found between trypsin domain isotype (x-axis) and clusters of amino acid binding motifs (y-axis) in experimentally solved 3D structures. For each cluster, the most frequent motif is indicated in brackets. **E)** For each protein domain, the number of significantly associated gene sets (x-axis)(FDR < 0.05) as a function of the median skew towards reference or non-reference domain isotype (y-axis). Color indicates if at least one gene set was significantly skewed towards an isotype (FDR < 0.05). **F)** For each-gene set significantly associated with the RAS domain, the skew towards reference or non-reference domain isotype (x-axis) and the associated certainty of the shift (p-value, y-axis). Color indicates significant skew (FDR < 0.05), and top 3 skewed gene-sets in each direction are highlighted. **G)** The enrichment of non-reference domain isotypes (y-axis) among domains associated with various disease classes (x-axis). Color denote FDR < 0.05.

Acknowledgments

Albin Sandlin for the encouraging discussion of the initial idea. For insightful feedback on manuscript drafts, Bo Porse, Elena Papaleo, and Lars Rønn Olsen.

References

1. Aziz, M. F. & Caetano-Anollés, G. Evolution of networks of protein domain organization. *Sci Rep-uk* 11, 12075 (2021).

2. Roskoski, R. Src protein–tyrosine kinase structure and regulation. *Biochem Bioph Res Co* 324, 1155–1164 (2004).

3. Mosca, R., Céol, A., Stein, A., Olivella, R. & Aloy, P. 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res* 42, D374–D379 (2014).

4. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* 49, gkaa913- (2020).

5. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240 (2014).

6. Klimovich, B. *et al.* p53 partial loss-of-function mutations sensitize to chemotherapy. *Oncogene* 41, 1011–1023 (2022).

7. Oren, M. & Rotter, V. Mutant p53 Gain-of-Function in Cancer. *Csh Perspect Biol* 2, a001107 (2010).

8. Kato, S. *et al.* Understanding the function–structure and function–mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc National Acad Sci* 100, 8424–8429 (2003).

9. Vitting-Seerup, K. & Sandelin, A. The Landscape of Isoform Switches in Human Cancers. *Mol Cancer Res* 15, 1206–1220 (2017).

10. Xiong, X. *et al.* Allosteric inhibition of the nonMyristoylated c-Abl tyrosine kinase by phosphopeptides derived from Abi1/Hssh3bp1. *Biochimica Et Biophysica Acta Bba - Mol Cell Res* 1783, 737–747 (2008).

11. Nath, D. *et al.* Abi1 loss drives prostate tumorigenesis through activation of EMT and non-canonical WNT signaling. *Cell Commun Signal Ccs* 17, 120 (2019).

12. Kumar, S. *et al.* Reciprocal regulation of Abl kinase by Crk Y251 and Abi1 controls invasive phenotypes in glioblastoma. *Oncotarget* 6, 37792–37807 (2015).

13. Consortium, T. U. *et al.* UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 49, D480–D489 (2020).

14. Consortium, T. Gte. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 369, 1318–1330 (2020).

15. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat Biotechnol* 38, 708–714 (2020).

16. Abbot, E. S. The Causal Relations between Structure and Function in Biology. *Am J Psychology* 27, 245 (1916).

17. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* 28, 235–242 (2000).

18. Rigden, D. J. The histidine phosphatase superfamily: structure and function. *Biochem J* 409, 333–348 (2007).

19. Ma, R. *et al.* Mutational study of human phosphohistidine phosphatase: Effect on enzymatic activity. *Biochem Bioph Res Co* 337, 887–891 (2005).

20. Savojardo, C., Babbi, G., Martelli, P. L. & Casadio, R. Mapping OMIM Disease–Related Variations on Protein Domains Reveals an Association Among Variation Type, Pfam Models, and Disease Classes. *Frontiers Mol Biosci* 8, 617016 (2021).

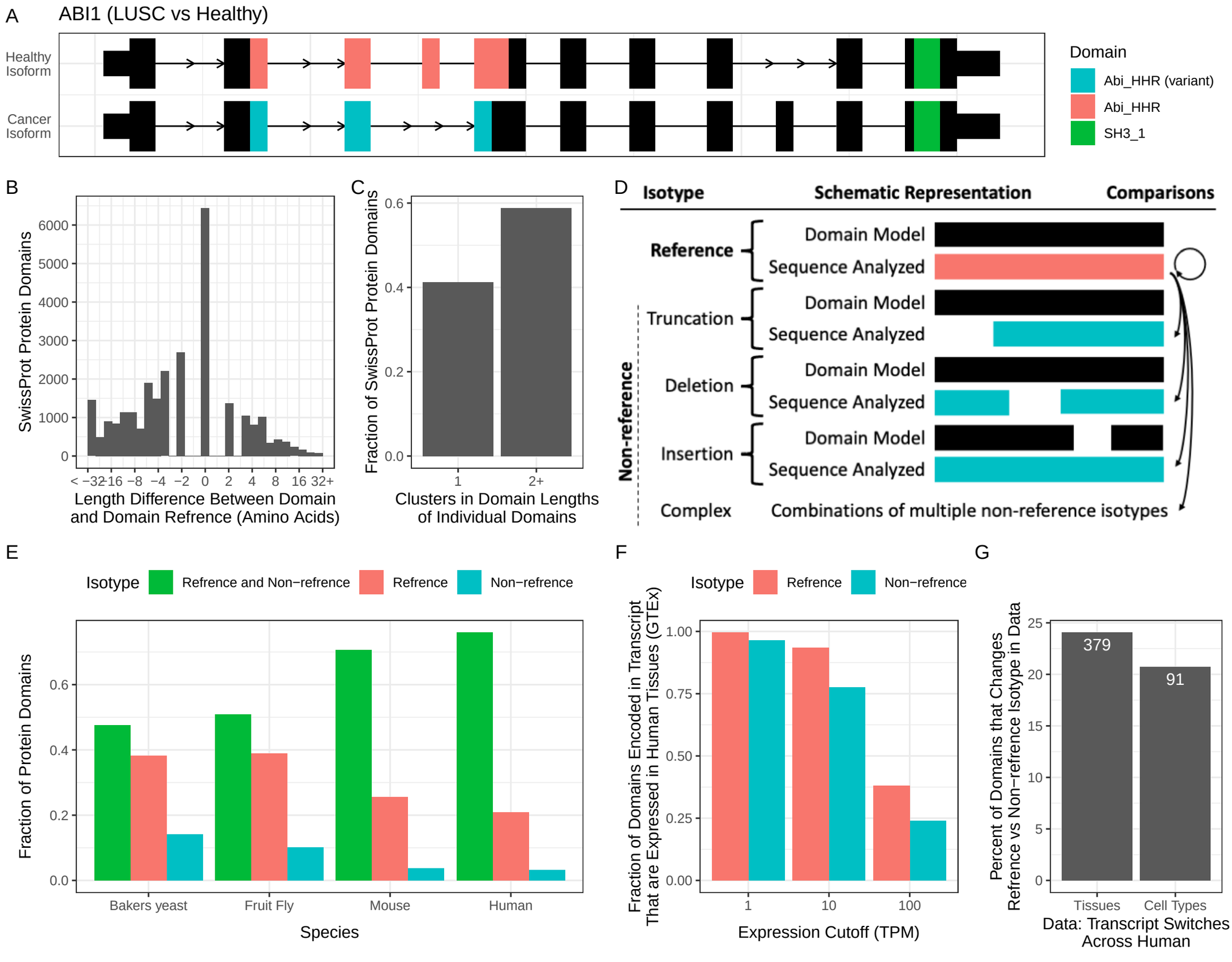
21. Peterson, T. A., Gauran, I. I. M., Park, J., Park, D. & Kann, M. G. Oncodomains: A protein domain-centric framework for analyzing rare variants in tumor samples. *Plos Comput Biol* 13, e1005428 (2017).

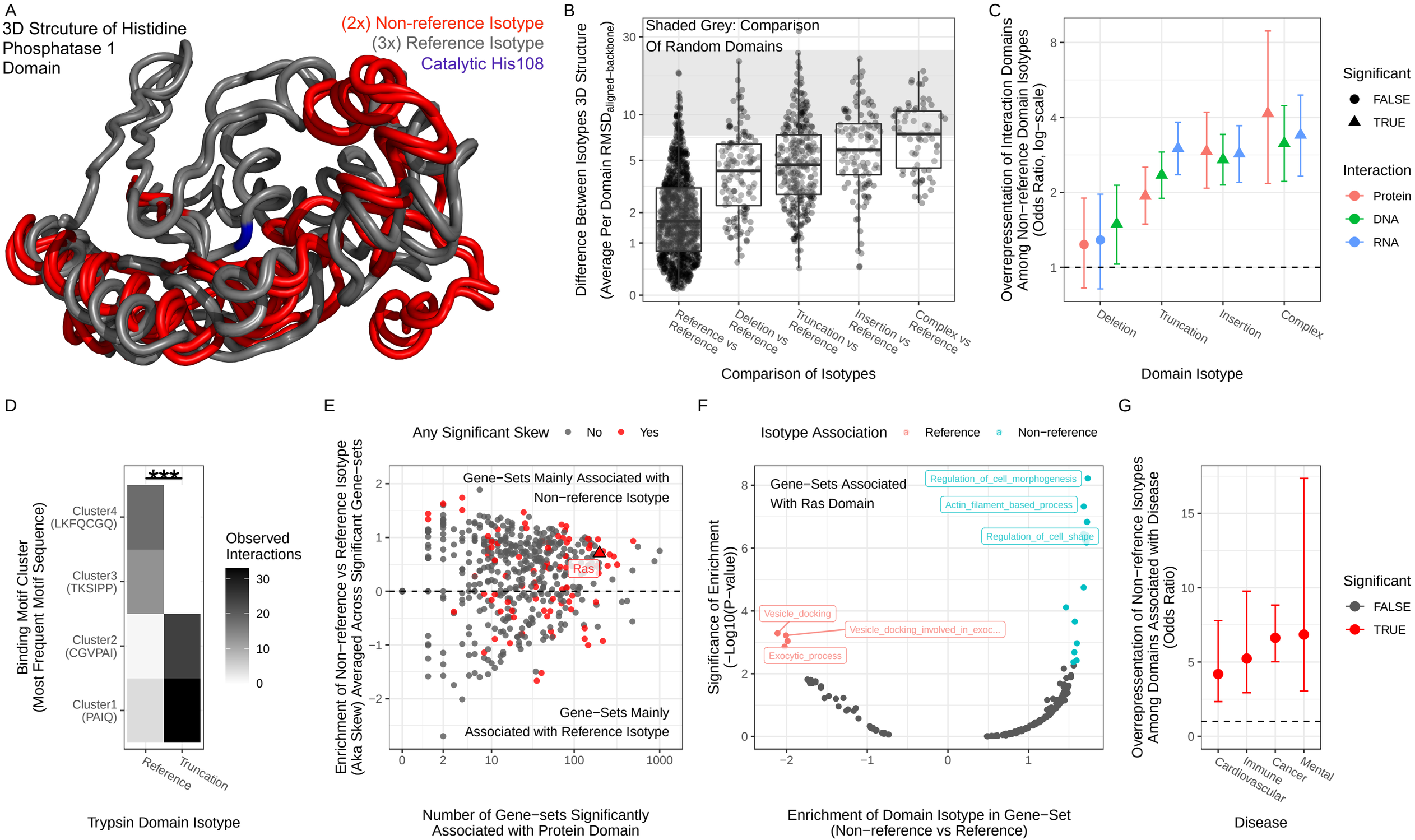
22. Glinos, D. A. *et al.* Transcriptome variation in human tissues revealed by long-read sequencing. *Biorxiv* 2021.01.22.427687 (2021) doi:10.1101/2021.01.22.427687.

23. Gupta, I. *et al.* Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* 36, 1197–1202 (2018).

24. Ji, Y., Mishra, R. K. & Davuluri, R. V. In silico analysis of alternative splicing on drug-target gene interactions. *Sci Rep-uk* 10, 134 (2020).

25. Barnkob, M. B., Vitting-Seerup, K. & Olsen, L. R. Target isoforms are an overlooked challenge and opportunity in chimeric antigen receptor cell therapy. *Immunotherapy Advances - In press* (2022).


A
ABI1 (LUSC vs Healthy)
Healthy Isoform
Cancer Isoform
Domain
Abi_HHR (variant)
Abi_HHR
SH3_1
B
SwissProt Protein Domains
Length Difference Between Domain and Domain Refrence (Amino Acids)
C
Fraction of SwissProt Protein Domains
Clusters in Domain Lengths of Individual Domains
D
Isotype
Schematic Representation
Comparisons
Reference
Domain Model
Sequence Analyzed
Non-reference
Truncation
Deletion
Insertion
Complex
Combinations of multiple non-reference isotypes
E
Isotype
Refrence and Non−refrence
Refrence
Non−refrence
Fraction of Protein Domains
Bakers yeast
Fruit Fly
Mouse
Human
Species
F
Fraction of Domains Encoded in Transcript That are Expressed in Human Tissues (GTEx)
Expression Cutoff (TPM)
G
Percent of Domains that Changes Refrence vs Non−refrence Isotype in Data
379
91
Tissues
Cell Types
Data: Transcript Switches Across Human


A
3D Strcuture of Histidine Phosphatase 1 Domain
(2x) Non-reference Isotype
(3x) Reference Isotype
Catalytic His108
B
Shaded Grey: Comparison Of Random Domains
Difference Between Isotypes 3D Structure (Average Per Domain RMSD aligned−backbone)
Reference vs Reference
Deletion vs Reference
Truncation vs Reference
Insertion vs Reference
Complex vs Reference
Comparison of Isotypes
C
Overrepresentation of Interaction Domains Among Non−reference Domain Isotypes (Odds Ratio, log−scale)
Deletion
Truncation
Insertion
Complex
Domain Isotype
Significant
FALSE
TRUE
Interaction
Protein
DNA
RNA
D
***
Binding Motif Cluster (Most Frequent Motif Sequence)
Cluster4 (LKFQCGQ)
Cluster3 (TKSIPP)
Cluster2 (CGVPAI)
Cluster1 (PAIQ)
Reference
Truncation
Observed Interactions
Trypsin Domain Isotype
E
Any Significant Skew
No
Yes
Gene−Sets Mainly Associated with Non−reference Isotype
Ras
Gene−Sets Mainly Associated with Reference Isotype
Enrichment of Non−reference vs Reference Isotype (Aka Skew) Averaged Across Significant Gene−sets
Number of Gene−sets Significantly Associated with Protein Domain
F
Isotype Association
Reference
Non−reference
Gene−Sets Associated With Ras Domain
Regulation_of_cell_morphogenesis
Actin_filament_based_process
Regulation_of_cell_shape
Vesicle_docking
Vesicle_docking_involved_in_exoc...
Exocytic_process
Significance of Enrichment (−Log10(P−value))
Enrichment of Domain Isotype in Gene−Set (Non−reference vs Reference)
G
Overrepresentation of Non−refrence Isotypes Among Domains Associated with Disease (Odds Ratio)
Cardiovascular
Immune
Cancer
Mental
Disease
Significant
FALSE
TRUE