frontiers
in Molecular Biosciences

ORIGINAL RESEARCH
published: 07 May 2021
doi: 10.3389/fmolb.2021.617016

Mapping OMIM Disease–Related Variations on Protein Domains Reveals an Association Among Variation Type, Pfam Models, and Disease Classes

Castrense Savojardo[1], Giulia Babbi[1], Pier Luigi Martelli[1] and Rita Casadio[1,2]*

[1] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy, [2] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council, Bari, Italy

OPEN ACCESS

Edited by:
*Arun Prasad Pandurangan,
MRC Laboratory of Molecular Biology
(LMB), United Kingdom*

Reviewed by:
*Yannick Mahlich,
Rutgers, The State University
of New Jersey, United States
Mauno Vihinen,
Lund University, Sweden*

***Correspondence:**
*Pier Luigi Martelli
pierluigi.martelli@unibo.it*

Specialty section:
*This article was submitted to
Biological Modeling and Simulation,
a section of the journal
Frontiers in Molecular Biosciences*

Received: *13 October 2020*
Accepted: *09 April 2021*
Published: *07 May 2021*

Citation:
*Savojardo C, Babbi G, Martelli PL
and Casadio R (2021) Mapping OMIM
Disease–Related Variations on Protein
Domains Reveals an Association
Among Variation Type, Pfam Models,
and Disease Classes.
Front. Mol. Biosci. 8:617016.
doi: 10.3389/fmolb.2021.617016*

Human genome resequencing projects provide an unprecedented amount of data about single-nucleotide variations occurring in protein-coding regions and often leading to observable changes in the covalent structure of gene products. For many of these variations, links to Online Mendelian Inheritance in Man (OMIM) genetic diseases are available and are reported in many databases that are collecting human variation data such as Humsavar. However, the current knowledge on the molecular mechanisms that are leading to diseases is, in many cases, still limited. For understanding the complex mechanisms behind disease insurgence, the identification of putative models, when considering the protein structure and chemico-physical features of the variations, can be useful in many contexts, including early diagnosis and prognosis. In this study, we investigate the occurrence and distribution of human disease–related variations in the context of Pfam domains. The aim of this study is the identification and characterization of Pfam domains that are statistically more likely to be associated with disease-related variations. The study takes into consideration 2,513 human protein sequences with 22,763 disease-related variations. We describe patterns of disease-related variation types in biunivocal relation with Pfam domains, which are likely to be possible markers for linking Pfam domains to OMIM diseases. Furthermore, we take advantage of the specific association between disease-related variation types and Pfam domains for clustering diseases according to the Human Disease Ontology, and we establish a relation among variation types, Pfam domains, and disease classes. We find that Pfam models are specific markers of patterns of variation types and that they can serve to bridge genes, diseases, and disease classes. Data are available as Supplementary Material for 1,670 Pfam models, including 22,763 disease-related variations associated to 3,257 OMIM diseases.

Keywords: protein variations, protein structure, protein domain, variation type, disease-related variations, disease variant databases, Pfam-disease association

INTRODUCTION

In the last decade, several efforts have been devoted to the problem of functional annotation of protein variants with the aim of relating variations to specific diseases (Vihinen, 2017, 2018). A collection of variations of genetic diseases is now available, and this prompted the investigation of molecular mechanisms responsible for protein failure (Schaafsma and Vihinen, 2018). Particularly, variations of non-synonymous proteins can promote the change of the active/binding sites and/or protein instability and can hamper protein–protein and ligand–protein interactions (Kucukkal et al., 2015; Ittisoponpisan et al., 2019; Ofoegbu et al., 2019). Molecular mechanisms can be, therefore, different, and different phenotypes may share common molecular mechanisms, independent of the different genes (Deans et al., 2015; Reeb et al., 2016; Babbi et al., 2019, and references therein). Several studies also focused on determining the most frequent protein variants associated with diseases, with the aim of helping functional annotation, starting from variant sequencing (Niroula and Vihinen, 2017; Zeng and Bromberg, 2019).

Different computational methods are available for the functional annotation of variations, based on different approaches. Routinely, given a specific variation, computational methods return with a computed reliability whether the change of a side chain in a protein is disease-related or not (Niroula and Vihinen, 2016).

An interesting aspect of disease-related protein variants is the protein instability promoted by the variations (Casadio et al., 2011; Savojardo et al., 2019, and references therein). Protein instability may be related to a disease, with this not being the only reason. For functional annotation of disease-related variations, routinely, the chemico-physical properties of the variation and the effect of the variation on the close environment in the protein structure are taken into consideration. It appears that the correlation among the strength of association to disease and the strength of association to the protein structure perturbation is moderate (Savojardo et al., 2019).

The problem of which phenotype is associated with a given variation or a set of variations has been scarcely addressed, and it remains unanswered, given the complexity of the scenario relating phenotypes to variations. Existing databases can relate genes to diseases and/or variations to diseases (MalaCards[1], Rappaport et al., 2017; GeneCards[2], Stelzer et al., 2016; DisGeNet[3], Piñero et al., 2020; eDGAR[4], Babbi et al., 2017; Humsavar[5], UniProt Consortium, 2019; OMIM[6], Amberger et al., 2015).

Protein domains have been adopted to explore associations between genes and human-inherited diseases (Zhang et al., 2011, 2016; Yates and Sternberg, 2013; Wiel et al., 2017, 2019). Models of protein domains are available in the Pfam database[7] (El-Gebali et al., 2019), and they enable the clustering of proteins into protein families, each represented by multiple sequence alignments, mainly based on protein structural alignments and cast into hidden Markov models (HMMs). Initially, similarities of disease phenotypes were exploited within a given domain–domain interaction network, and a Bayesian approach was proposed to prioritize candidate domains for human complex diseases (Zhang et al., 2011). Then, domain–disease associations were inferred from domain–protein, protein–disease, and disease–disease relationships (Zhang et al., 2016). In these studies, the bottom layer of variations in proteins, detected in large-scale sequencing experiments, was not taken into consideration, restraining the analysis only to the already known protein– or gene–disease associations. More recently (Wiel et al., 2017), with the notion of homologous domains in proteins, variants were aggregated to improve their interpretation, and a web server (MetaDome[8], Wiel et al., 2019) was made available for the pathogenicity analysis of genetic variants.

In a previous study (Savojardo et al., 2019), we introduced the notion of variation type, in order to take the physico-chemical properties of the variations into account as well (Casadio et al., 2011). After mapping genetic disease–related variations on a restricted set of human protein three-dimensional (3D) structures, we found that the distribution of disease variation types significantly varies across different structural/functional Pfam models.

In this study, relying on the relationship between genes and phenotypes, we ask the question as to which extent possible patterns of variation types framed into Pfam domains are significant for a reliable association to specific groups of maladies.

MATERIALS AND METHODS

Dataset Construction

The dataset adopted in this study was derived from the Humsavar database[5] release 2020_04 of August 2, 2020, listing all missense variants annotated in human UniProtKB/Swiss-Prot (UniProt Consortium, 2019) entries.

From the initial set of proteins included in the database, we only selected those reporting at least one variant implicated in the disease, excluding proteins reporting only polymorphisms not associated with disease insurgence. Moreover, any variation labeled as "unclassified" (i.e., with uncertain implications in disease) was filtered out. Finally, we only retained disease-related variations associated with a genetic disorder reported in the Online Mendelian Inheritance in Man (OMIM) catalog[9].

The set of neutral variations was extended using data retrieved from the GnomAD database (exome version 2.1.1) (Karczewski et al., 2020). Only variations occurring in our set of proteins, not already included in Humsavar and with clinical significance

[1] https://www.malacards.org/
[2] https://www.genecards.org/
[3] https://www.disgenet.org/
[4] http://edgar.biocomp.unibo.it
[5] https://www.uniprot.org/docs/humsavar
[6] https://www.omim.org/
[7] https://pfam.xfam.org/
[8] https://stuart.radboudumc.nl/metadome/
[9] https://omim.org/

labeled as “Benign/Likely benign” by ClinVar (release 2021-03-23) (Landrum et al., 2020), were retained.

Pfam (El-Gebali et al., 2019) annotations were retrieved from the Pfam-A region annotation file for *Homo sapiens* version 33.1 obtained *via* the Pfam FTP server[10]. From all the annotations available, we only retained those occurring at proteins included in our set of data and covering at least one disease-related variation.

Mapping OMIMs to Disease Ontology

The DO (Human Disease Ontology) OBO (Open Biological and Biomedical Ontology) file release of September 15, 2020, was downloaded[11] and used directly to retrieve annotations for each OMIM disease by means of cross-references. Each retrieved leaf DO term associated to a single OMIM was expanded up to the ontology root term, including all ancestors. Term expansion was computed using an *ad-hoc* script to parse the OBO file.

Computing the Disease Score

For each Pfam domain, we estimated a propensity score for the association to the disease as follows:

$$Score\ (pfam) = \frac{N_d^{pfam} / \left(N_d^{pfam} + N_p^{pfam}\right)}{N_d / \left(N_d + N_p\right)} \tag{1}$$

where N_d^{pfam} and N_p^{pfam} are the number of disease-related and polymorphism variations in the domain *pfam*, while N_d and N_p are the same numbers in the whole dataset. In the dataset, scores range from 1.40 down to 0.03.

Kullback–Leibler Divergence Between Distributions

Differences between probability distributions were evaluated using the Kullback–Leibler divergence:

$$D_{KL} = -\sum_{x \in X} p\,(x) \cdot log_2 \frac{q(x)}{p(x)} \tag{2}$$

where p and q are two discrete probability distributions defined on the same probability space X.

RESULTS

A Dataset of Variations With Annotated Pfam

Overall, our dataset comprises 50,746 variations occurring in 2,959 proteins implicated in 3,884 genetic disorders. Disease-related variations in these proteins are 29,949, accounting for 55% of the total variations. The remaining 20,797 variations are neutral (45%). **Table 1** shows summary statistics about the dataset analyzed in this study.

TABLE 1 | Summary of the OMIM-related variation dataset of this study.

Number of proteins associated with disease	2959
Number of diseases (OMIM)	3884
Number of variations	54746
Number of disease variations	29949 (55%)$^{\wedge}$
Number of neutral polymorphisms (on the same disease proteins)	24797 (45%)$^{\wedge}$
Number of disease proteins with Pfam covering disease variations	2513 (85%)$^{\#}$
Number of Pfams	1670
Number of diseases (OMIM) in proteins with Pfams	3257 (84%)$^{\circ}$
Number of variations covered by Pfams	31934 (68%)$^{\wedge}$
Number of disease variations covered by Pfams	22763 (71%)$^{+}$
Number of neutral polymorphisms covered by Pfams	9171 (29%)$^{+}$

$^{\wedge}$ *percentage computed with respect to the total number of variations (54746);* $^{\#}$ *percentage computed with respect to the total number of proteins (2959);* $^{\circ}$ *percentage computed with respect to the total number of diseases (3884);* $^{+}$ *percentage computed with respect to the total number of Pfam-covered variations (31934).*

Restricting the set of proteins to those having Pfam entries covering at least one disease-related variation, we ended up with 2,513 proteins (corresponding to 85% of the initial protein set) implicated in 3,257 distinct genetic diseases. Overall, 1,670 distinct Pfam entries were annotated on these proteins. A subset of 548 out of 1,670 Pfams occurs in two or more proteins in the set. The vast majority (96%) of Pfam entries are of type “Domain” or “Family,” while a very small fraction accounts for “Repeat,” “Coiled-coil,” “Motif,” and “Disordered” types.

After this reduction, we retained 31,934 variations covered by Pfams, distributed into 22,763 (71%) and 9,171 (29%) disease-related and neutral polymorphic variations, respectively.

Data shown in **Table 1** clearly indicate that the incidence of disease-related variations within Pfam domains is significantly higher than the background (71% against 55%).

Overall Pfam Association With Disease

We were interested in elucidating the overall association between Pfam and OMIM diseases. For each entry in the set of 1,670 Pfam domains in our dataset, we computed the score for the association to disease with the formula reported in Eq. 1. A value greater than 1 for this ratio highlights a higher abundance of disease variations in the Pfams than in the background. The complete result of this analysis is reported in **Supplementary Table 1** for all the 1,670 Pfam entries. About 48% of Pfam entries have a value greater than 1, as a consequence of the overall propensity of disease-related variations to be located within Pfam domains. In general, the distribution of scores is not random and reflects a differential disease association for the different Pfam entries.

In **Table 2**, we list the result for the 20 highest scoring Pfams covering 10 or more proteins. Scores with corrected *p*-values (**Supplementary Table 2**) equal to or lower than 0.1 are highlighted (top scoring Pfams are all significant at 0.1 level). Significance does not hold for some Pfams covering only few variations. In these cases, more data are needed in order to properly evaluate the association to the disease.

Interestingly, Pfam entries reported in **Table 2** can be grouped into few functional classes, including DNA-binding domains

[10] ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam33.1/proteomes/9606.tsv.gz

[11] https://disease-ontology.org/

TABLE 2 | The 20 highest scoring Pfam entries mostly associated with diseases.

Pfam ID	Pfam name	Pfam type	No of proteins	No of disease variations	No of neutral polymorphisms	Score§
PF00105	zf-C4	Domain	12	60	2	1.36*
PF00250	Forkhead	Domain	10	88	4	1.34*
PF00010	HLH	Domain	14	48	3	1.32*
PF00104	Hormone_recep	Domain	18	195	20	1.27*
PF00307	CH	Domain	11	48	6	1.25*
PF00046	Homeodomain	Domain	42	163	21	1.24*
PF07645	EGF_CA	Domain	17	301	46	1.22*
PF00096	zf-C2H2	Domain	23	80	13	1.21*
PF00029	Connexin	Family	10	319	53	1.20*
PF00017	SH2	Domain	11	72	12	1.20*
PF00520	Ion_trans	Family	48	1020	173	1.20*
PF00004	AAA	Domain	10	70	12	1.20*
PF00400	WD40	Repeat	19	52	9	1.20
PF02770	Acyl-CoA_dh_M	Domain	10	40	7	1.19
PF00169	PH	Domain	11	53	10	1.18
PF00005	ABC_tran	Domain	15	236	49	1.16*
PF07686	V-set	Domain	12	84	18	1.16
PF00271	Helicase_C	Family	17	65	15	1.14
PF00176	SNF2_N	Family	10	63	15	1.13
PF00089	Trypsin	Domain	21	258	87	1.13*

§Score is computed as defined in Eq. 1. Significance of each score was assessed using the Fisher exact test on the corresponding contingency table and correcting for multiple testing using the Benjamini-Hochberg procedure. Individual p-values are listed in ***Supplementary Table 2****. *Corrected P-values are equal or lower than 0.1.*

(accounting for eight domains/families), transmembrane domains (three), and enzymes (three).

Pfams Have Distinctive Patterns of Disease Variation Types

Going a step further in the analysis, we investigated the composition of disease-related variations occurring in different Pfam domains. In a previous study (Savojardo et al., 2019), the same analysis was performed on a small dataset of highly curated variations covered by 3D structures from Protein Data Bank (PDB). In this study, we extended and complemented the previous results using a larger dataset of Pfam domains and variations. To this aim, we first grouped residues according to their physico-chemical properties, obtaining four major groups, namely, apolar (GAVPLIM), aromatic (FWY), polar (STCNQH), and charged (DEKR) residues. We define a variation type in relation to the conservation or substitution of apolar (a), polar (p), aromatic (r), and charged (c) (**Figure 1**). Then, we computed Pfam-specific distributions of disease-related variations involving substitutions from one group to another (overall, 16 different substitution types are possible). Complete results are reported in **Supplementary Table 3** for all the 1,670 Pfam domains.

In **Figure 1**, we show a heatmap reporting the frequencies of each substitution type for the 20 highest scoring Pfam entries described in the previous section and mostly associated with diseases. For each Pfam entry, we report the Pfam ID, the name, and two numbers in parentheses, indicating the number of proteins and disease-related variations covered by the specific Pfam. For comparison, the last row reports the overall distribution of substitution types computed on the whole set of variation types covered by Pfams.

The results shown in the heatmap of **Figure 1** indicate that the different Pfams are enriched in different variation types and that each Pfam shows a differential pattern with respect to the background. Interestingly, in some cases, the pattern of enriched variation types can be related with the overall function of the Pfam domain and/or the cellular context in which the domain/s are presumably operating.

In **Figure 2**, we report three examples, namely, a selection of DNA-binding domains, growth factors, and transmembrane domains. For DNA-binding domains, we observe a higher concentration of disease-related variations involving a substitution from a charged residue to any different residue type. Contrarily, for growth factor domains, we observe abundant variations involving substitutions from polar to any type of the residue, while transmembrane domains are mostly enriched in substitutions involving apolar wild types. These observations clarify a general trend, pointing to the specificity of the disease variation type per Pfams of functional classes.

From data analysis, we conclude that the distribution of the disease-related variation type patterns observed for the different Pfams is non-random and different from the background distribution (computed considering all the disease-related variation types occurring in Pfams). This observation confirms our previous results obtained with a smaller number of Pfam domains, directly related to human protein structures, and corroborates the notion that distinctive patterns of disease-related variation types are Pfam specific (Savojardo et al., 2019).

Linking the Pfam to Disease Ontology

As a final step of our investigation, we searched for a link between Pfam domains and disease ontology. Disease classification is not a trivial task. Different controlled vocabularies and ontologies

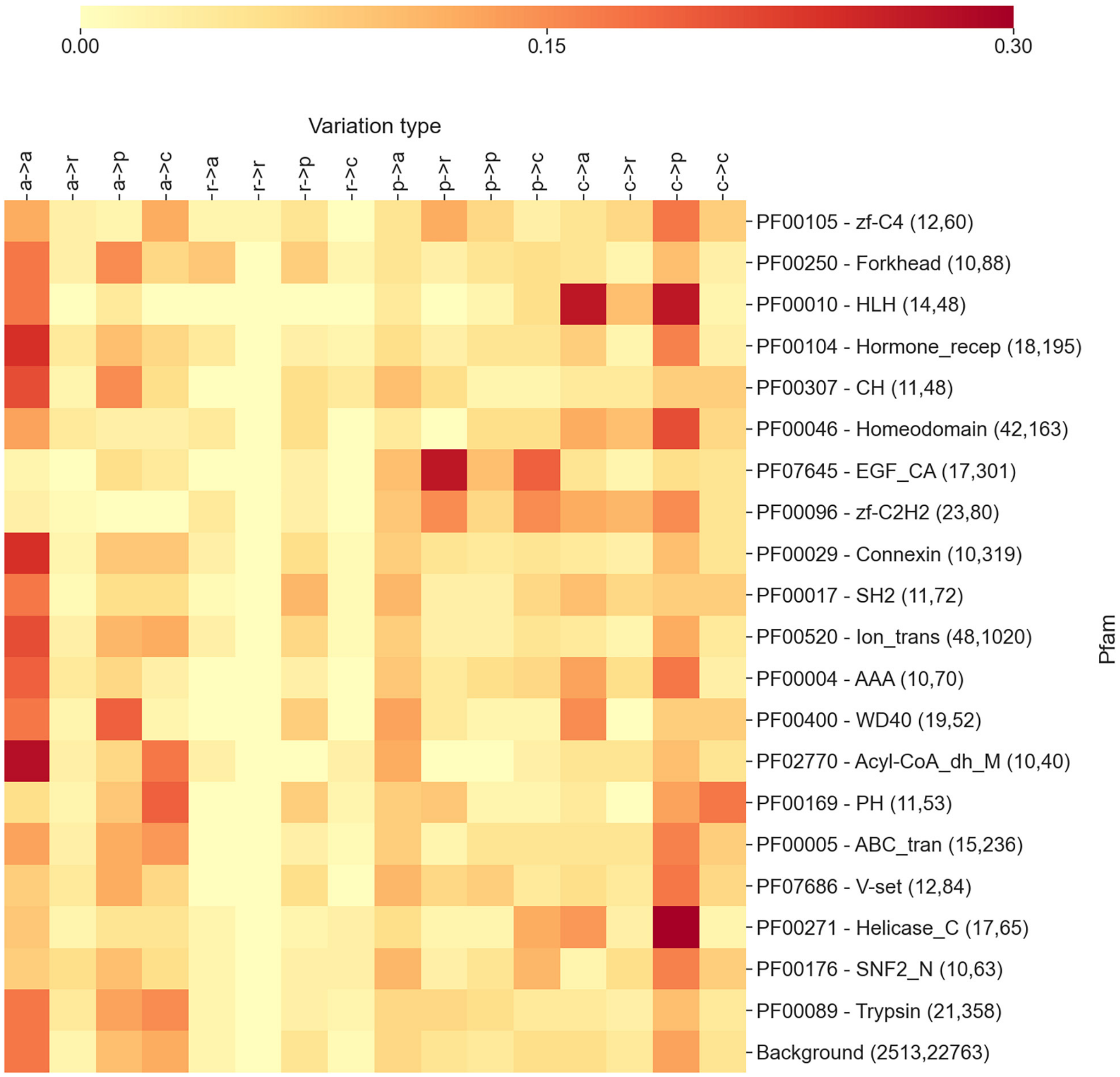

FIGURE 1 | The heatmap reporting the frequency of each variation type as observed within the 20 Pfam entries mostly associated with diseases. For each Pfam, the numbers within parentheses indicate the number of proteins and disease-related variations covered. In variation types, labels are as follows: a, apolar; r, aromatic; p, polar; and c, charged. Mean and median Kullback–Leibler divergences (Eq. 2) between individual Pfam distributions and the background are 2.1 and 2.1 bits, respectively.

such as the Human Phenotype Ontology (HPO)[12] (Köhler et al., 2019) or the DO (Schriml et al., 2019) are available for this purpose. However, none of the ontologies provides a full coverage of the entire space of OMIM diseases, ranging from 82% coverage of HPO to 74% of DO. Moreover, ontologies like HPO are not specifically designed to describe a disease. Instead, they are devised to describe clinically relevant phenotypes. In the current study, we used the DO ontology because, in spite of a slightly lower coverage, it provides a better and less ambiguous classification of diseases.

To obtain a high-level disease classification, we collected all the 3,257 OMIM diseases linked to variations occurring in our 1,670 Pfam domains and mapped them to a set of 17 first-level DO terms. These include 12 terms describing diseases affecting anatomical entities (all child terms of "DOID:7 – disease of anatomical entity" like cardiovascular, endocrine, gastrointestinal, etc.), cellular proliferation diseases (DOID:14566), mental health diseases (DOID:150), metabolic diseases (DOID:0014667), physical disorders (DOID:0080015), and syndromes (DOID:225). We were able to map 2,454 out of 3,257 OMIMs to at least one of the above DO terms. On average, each OMIM was mapped to 1.01 DO,

[12] https://hpo.jax.org/app/

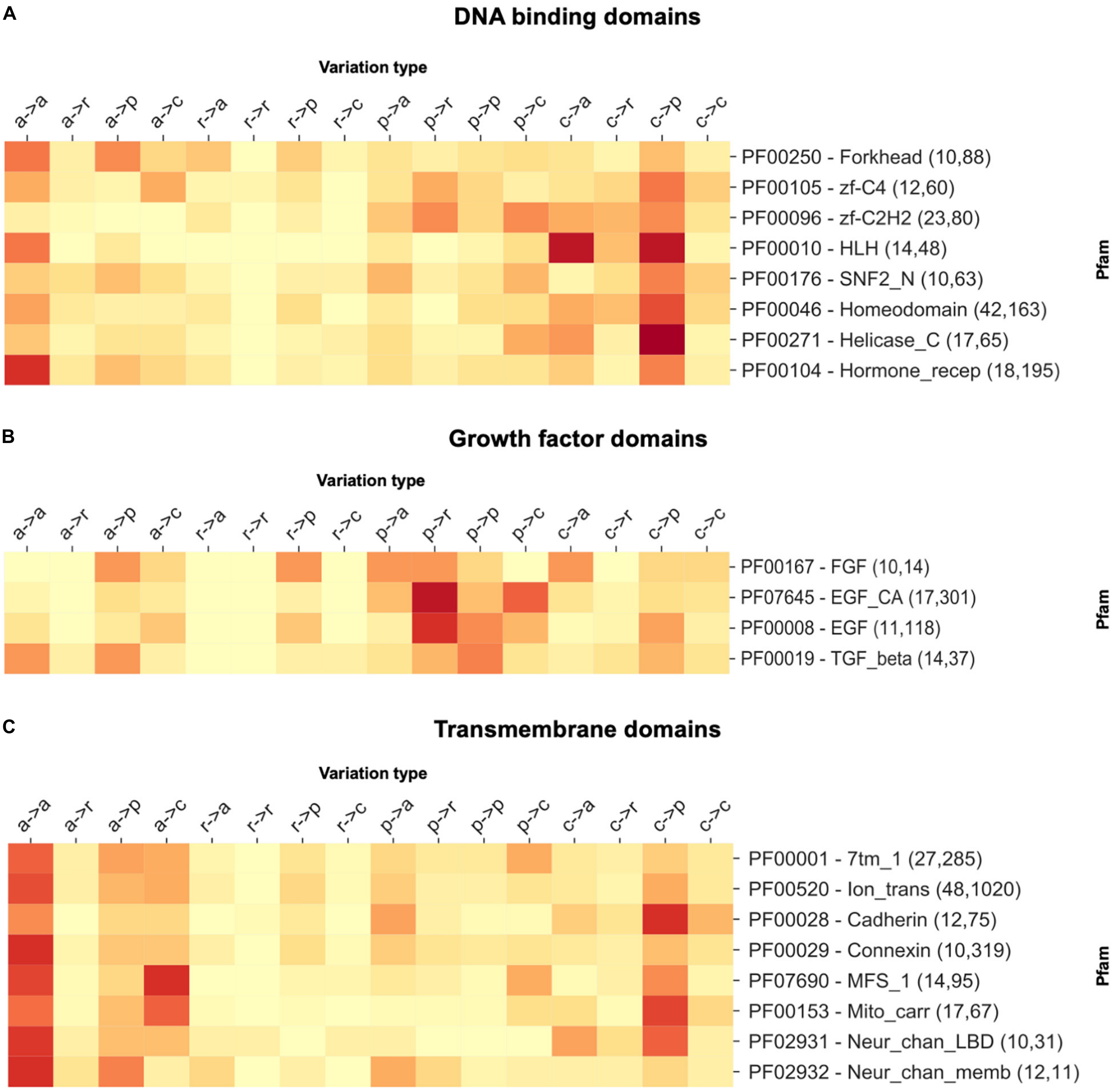

FIGURE 2 | The heatmap reporting the frequency of each variation type as observed within a selection of **(A)** DNA-binding, **(B)** growth factor, and **(C)** transmembrane domains. For each Pfam, the numbers within parentheses indicate the number of proteins and disease-related variations covered. In variation types, labels are as follows: a, apolar; r, aromatic; p, polar; and c, charged.

providing an almost strict classification of each OMIM into a single DO term.

With this mapping, we computed a Pfam-specific distribution of DO-associated disease classes. Complete results are reported in **Supplementary Table 4** for all the 1,670 Pfam entries considered in this study. The data provided in this study indicate that disease classes are not evenly distributed among different Pfam domains, again suggesting a differentiated association between the Pfam and phenotypes.

In **Figure 3**, we show an extract of our analysis, focusing on the 20 highest scoring Pfam domains associated with diseases. The heatmap reports, for each Pfam, the frequency of disease types (in the 17 different classes detailed above) as retrieved from OMIMs associated with substitutions occurring on the specific Pfam. In brackets, close to each Pfam name, we list the number of proteins, disease variations, and OMIMs associated to the Pfam.

Even in this case, the distributions of disease classes appear to be very different from the background (reported in the last row of the heatmap). Remarkably, the aggregation of Pfams into more general functional classes provides an additional level of interpretation. Considering **Figure 3**, we can observe that DNA-binding domains are mostly associated with syndromes, nervous system, and endocrine system disease classes, while enzymes are mostly involved in the metabolic disease

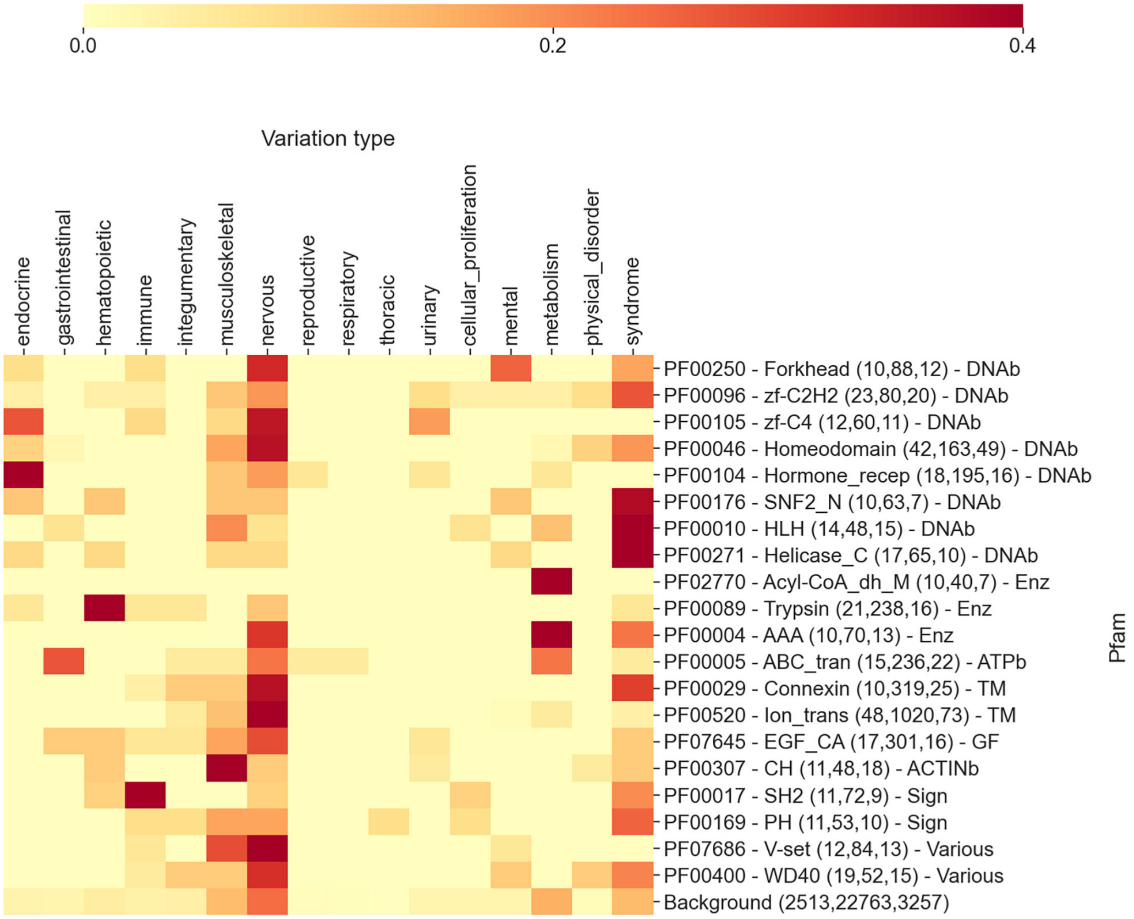

FIGURE 3 | The heatmap reporting, for each Pfam, the frequency of diseases (grouped into 17 different classes extracted from Disease Ontology) as retrieved from OMIMs, after the association *via* the disease type with Pfam. The numbers within parentheses are the number of proteins, the number of disease variations, and the number of Online Mendelian Inheritance in Man (OMIM) diseases associated with the Pfam, respectively. Each Pfam is labeled according to its functional class: DNAb, DNA-binding domain; Enz, enzymatic domain; TM, transmembrane; GF, growth factor; ACTINb, actin-binding domain; Sign, signaling; and Various, various functions associated. Mean and median Kullback–Leibler divergences (Eq. 2) between individual Pfam distributions and the background are 2.5 and 2.7 bits, respectively.

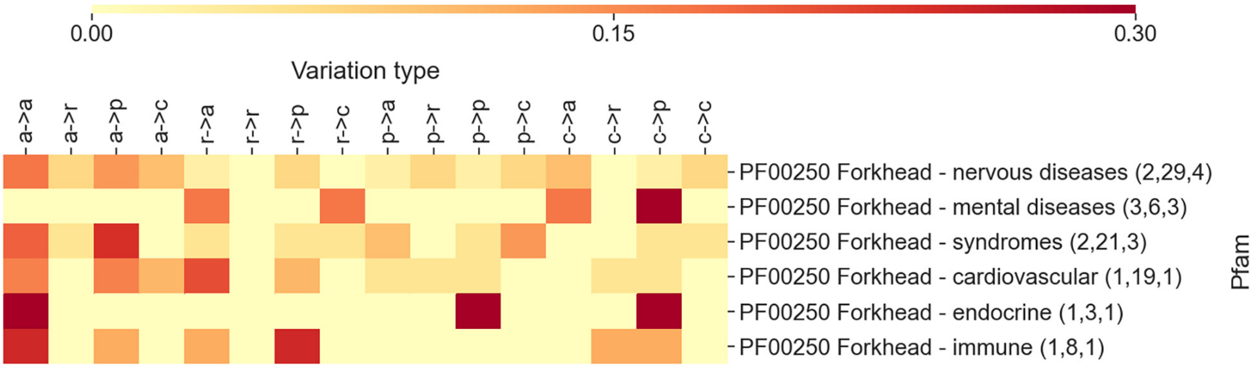

FIGURE 4 | The heatmap reporting, for the Pfam entry PF00250 Forkhead, the frequency of each variation type as observed after separating variations according to disease classes. The numbers within parentheses are the number of proteins, the number of disease variations, and the number of Online Mendelian Inheritance in Man (OMIM) diseases associated with the Pfam (and covered by DOID), respectively.

class. Transmembrane domains show the prevalence of nervous and integumentary disease classes, while growth factors and actin-binding domains are enriched in musculoskeletal diseases. Finally, signaling Pfam domains are prominently associated with immune system diseases. Overall, many of these findings are in line with what we expected. Protein domains have different functions and are involved into different biological processes. Variations occurring in these domains, when disruptive, lead to diseases that are connected to the biological processes in which the proteins are mainly involved. For instance, the fact that variations occurring in transmembrane domains are often linked to neurological diseases is a direct consequence of the involvement of transmembrane proteins (among other functions) in neurotransmission. Similarly, variations in enzymes routinely lead to metabolic diseases.

Some of the Pfams reported in **Figure 3** are associated to more than one disease types. For example, diseases that are associated to the Forkhead domain (PF00250) are distributed into five classes, namely, nervous, mental, endocrine, immune diseases, and syndromes. In **Figure 4**, an additional heatmap is shown trying to link the disease types to the patterns of variation types. Specifically, the patterns of variation types are reported after isolating variations linked to OMIMs in the different disease classes. Interestingly, the patterns show an evident difference among each other. This confirms the level of association that links domains to variation types and diseases.

CONCLUSION AND PERSPECTIVES

In this study, we consider, for the time being, only diseases of genetic origins, with the belief that cancer-related somatic variations are as yet not satisfactorily clustered according to tissue specificity of the plague.

This study, as well as the previous ones (Yates and Sternberg, 2013; Wiel et al., 2017, 2019), aims at establishing a direct mapping among variations, diseases, and phenotypes *via* the protein domains. Our novelty is the introduction of the variation type as a distinguished feature of association to the Pfam domain and to the phenotype. Our findings complement previous ones (Wiel et al., 2017) with the inclusion of the variation type, which adds to the classification of variations and their impact on the protein function, stability, and interaction in the specific context where the gene is active.

The link among the variation type, Pfam domain, and phenotype can greatly reduce the number of possible steps to understand which variations are disease-related or which are not and which phenotype they may promote. In perspective, the association among the variation type, protein domain/s, and phenotype may greatly simplify the problem of genetic variant annotation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RC, PM, and CS: conceptualization, methodology, and writing. CS: software. GB and CS: data curation and visualization. RC and PM: supervision. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the PRIN2017 grant (project 2017483NH8_002), delivered to CS from the Italian Ministry of University and Research.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.617016/full#supplementary-material

REFERENCES

Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* 43, D789–D798. doi: 10.1093/nar/gku1205

Babbi, G., Martelli, P. L., and Casadio, R. (2019). PhenPath: a tool for characterizing biological functions underlying different phenotypes. *BMC Genom.* 20, 548–558. doi: 10.1186/s12864-019-5868-x

Babbi, G., Martelli, P. L., Profiti, G., Bovo, S., Savojardo, C., and Casadio, R. (2017). eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genom.* 18, 554–564. doi: 10.1186/s12864-017-3911-3

Casadio, R., Vassura, M., Tiwari, S., Fariselli, P., and Martelli, P. L. (2011). Correlating disease related mutations to their effect on protein stability: a large-scale analysis of the human proteome. *Hum. Mutat.* 32, 1161–1170. doi: 10.1002/humu.21555

Deans, A. R., Lewis, S. E., Huala, E., Anzaldo, S. S., Ashburner, M., Balhoff, J. P., et al. (2015). Finding our way through phenotypes. *PLoS Biol.* 13:e1002033. doi: 10.1371/journal.pbio.1002033

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Ittisoponpisan, S., Islam, S., Khanna, T., Alhuzimi, E., David, A., Sternberg, M., et al. (2019). Can predicted protein 3D-structures provide reliable insights into whether missense variants are disease-associated? *J. Mol. Biol.* 431, 2197–2212. doi: 10.1016/j.jmb.2019.04.009

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443. doi: 10.1038/s41586-020-2308-7

Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J. O. B., Danis, D., Gourdine, J. P., et al. (2019). Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 47, D1018–D1027. doi: 10.1093/nar/gky1105

Kucukkal, T. G., Petukh, M., Li, L., and Alexov, E. (2015). Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. *Curr. Opin. Struct. Biol*. 32, 18–24. doi: 10.1016/j.sbi.2015.01.003

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., et al. (2020). ClinVar: improvements to accessing data. *Nucleic Acids Res.* 48, D835–D844. doi: 10.1093/nar/gkz972

Niroula, A., and Vihinen, M. (2016). Variation Interpretation Predictors: Principles. Types, Performance, and Choice. *Hum. Mutat.* 37, 579–597. doi: 10.1002/humu.22987

Niroula, A., and Vihinen, M. (2017). Predicting Severity of Disease-Causing Variants. *Hum. Mutat.* 38, 357–364. doi: 10.1002/humu.23173

Ofoegbu, T., David, A., Kelley, L., Mezulis, S., Islam, S., Mersmann, S., et al. (2019). PhyreRisk: a dynamic web application to bridge genomics, proteomics and 3D structural data to guide interpretation of human genetic variants. *J. Mol. Biol.* 431, 2460–2466. doi: 10.1016/j.jmb.2019.04.043

Piñero, J., Ramírez-Anguita, J. M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., et al. (2020). The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* 48, D845–D855. doi: 10.1093/nar/gkz1021

Rappaport, N., Twik, M., Plaschkes, I., Nudel, R., Iny Stein, T., Levitt, J., et al. (2017). MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res.* 45, D877–D887. doi: 10.1093/nar/gkw1012

Reeb, J., Hecht, M., Mahlich, Y., Bromberg, Y., and Rost, B. (2016). Predicted Molecular Effects of Sequence Variants Link to System Level of Disease. *PLoS Comput. Biol*. 12:e1005047. doi: 10.1371/journal.pcbi.1005047

Savojardo, C., Babbi, G., Martelli, P. L., and Casadio, R. (2019). Functional and Structural Features of Disease-Related Protein Variants. *Int. J. Mol. Sci.* 20, 1530–1544. doi: 10.3390/ijms20071530

Schaafsma, G. C. P., and Vihinen, M. (2018). Representativeness of variation benchmark datasets. *BMC Bioinformatics* 19, 461–479. doi: 10.1186/s12859-018-2478-6

Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., et al. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res*. 47, D955–D962. doi: 10.1093/nar/gky1032

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinformatics* 54, 1301–1333. doi: 10.1002/cpbi.5

UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Vihinen, M. (2017). How to Define Pathogenicity, Health, and Disease? *Hum. Mutat.* 38, 129–136. doi: 10.1002/humu.23144

Vihinen, M. (2018). Systematics for types and effects of DNA variations. *BMC Genomics* 28, 974–989. doi: 10.1186/s12864-018-5262-0

Wiel, L., Baakman, C., Gilissen, D., Veltman, J. A., Vriend, G., and Gilissen, C. (2019). MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. *Hum. Mutat*. 40, 1030–1038. doi: 10.1002/humu.23798

Wiel, L., Venselaar, H., Veltman, J. A., Vriend, G., and Gilissen, C. (2017). Aggregation of population-based genetic variation over protein domain homologues and its potential use in genetic diagnostics. *Hum. Mutat.* 38, 1454–1463. doi: 10.1002/humu.23313

Yates, C. M., and Sternberg, M. (2013). Proteins and domains vary in their tolerance of Non-Synonymous Single Nucleotide Polymorphisms (nsSNPs). *J. Mol. Biol.* 425, 1274–1286. doi: 10.1016/j.jmb.2013.01.026

Zeng, Z., and Bromberg, Y. (2019). Predicting Functional Effects of Synonymous Variants: A Systematic Review and Perspectives. *Front. Genet.* 10:914. doi: 10.3389/fgene.2019.00914

Zhang, W., Chen, Y., Sun, F., and Jiang, R. (2011). Domain RBF: a Bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC Syst. Biol.* 5, 55–75. doi: 10.1186/1752-0509-5-55

Zhang, W., Coba, M. P., and Sun, F. (2016). Inference of domain-disease associations from domain-protein, protein-disease and disease-disease relationships. *BMC Syst. Biol*. 10, 63–89. doi: 10.1186/s12918-015-0247-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2021 Savojardo, Babbi, Martelli and Casadio. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.