 doi:10.20944/preprints201811.0280.v1

Article

Mutational hotspots and protein interactome analyses of collagen-specific chaperone - HSP47

Alisha Parveen[1], Rajesh Kumar[2], Ravi Tandon[3], Sukant Khurana[4], Chandan Goswami[5] & Abhishek Kumar[6,*]

[1]Medical Research Center, Medical Faculty of Mannheim, University of Heidelberg, Mannheim, Germany;dralishaparveen@yahoo.com

[2]Center for Molecular Biology of Heidelberg University (ZMBH), DKFZ-ZMBH Alliance, Heidelberg, Germany;rajesh.mukesh.naga@gmail.com

[3]Laboratory of AiDS Research and Immunology, School of Biotechnology, Jawaharlal Nehru University, New Delhi, India; ravitandon@jnu.ac.in

[4]Pharmacology Department, Central Drug Research Institute - Lucknow, Uttar Pradesh, India; sukantkhurana@gmail.com

[5]National Institute of Science Education and Research, Bhubaneswar, Orissa, India; chandan@niser.ac.in

[6]Department of Genetics & Molecular Biology in Botany, Institute of Botany, Christian-Albrechts-University at Kiel, Germany; abhishek.abhishekkumar@gmail.com

* Correspondence: abhishek.abhishekkumar@gmail.com; Tel.: +49-17647164094

Abstract: Heat shock protein 47kDa (HSP47) serves as a client-specific chaperone, essential for collagen biosynthesis and its folding and structural assembly. To date, there is no comprehensive study on mutational hotspots and protein network for human HSP47. Using five different human mutational databases, we deduced a comprehensive list of human HSP47 mutations and we found 24 67, 50, 43 and 2 deleterious mutations from the 1000 genomes data, gnomAD, COSMICv86, cBioPortal, and CanVar. We identified thirteen top-ranked missense mutations of HSP47 with the stringent cut-off of CADD score (>25) and Grantham score (≥151) as Ser76Trp, Arg103Cys, Arg116Cys, Ser159Phe, Arg167Cys, Arg280Cys, Trp293Cys, Gly323Trp, Arg339Cys, Arg373Cys, Arg377Cys, Ser399Phe, and Arg405Cys with the arginine-cysteine change as the predominant mutation. We also found that HSP47 is up-regulated and down-regulated in 11 and 4 of cancers types. Upon constructing protein interactome map of human HSP47, we found that a set of molecular chaperones is interaction partners of HSP47, which included two copies each of CREB binding proteins, HSP27, HSP40, HSP70, HSP90, ubiquitin proteins and one copy each of cartilage associated protein (CRTAP), HSPH1, HSBP1, FK506-binding protein B (FKBP), kruppel-like factor (KLF13), peptidyl-prolyl isomerase PIPB and Prolyl 4-hydroxylase beta subunit (P4HB). This suggested a cocktail of different chaperones interact with HSP47. These findings will assist in the evaluation of roles of HSP47 in human disease including different types of cancers.

Keywords: HSP47; missense mutation; mutational hotspot; variant analysis; cancer database; chaperone

 doi:10.20944/preprints201811.0280.v1

1. Introduction

Heat shock protein 47 kDA (HSP47) serves as an endoplasmic reticulum (ER)-residing collagen-specific chaperone and it has the cavalier role in collagen biosynthesis and its structural assembly [1,2]. HSP47 protein is the product of the human SERPINH1 gene, which belongs to the group V6 in the indel-based group-wise classification of vertebrate serpins [3]. Structurally, HSP47 possesses a typical serpin domain (Pfam ID - PF00079 and InterPro ID - IPR000215), composed of threeβ-sheets (s) and nine α-helices (h) as sA-sC and hA-hI, respectively [4]. HSP47 lacks inhibitory function due to mutations in its reactive center loop (RCL) [1]. Over last five decades, HSP47 has been extensively characterized by biochemical and biophysical methods to demonstrate its roles in the collagen biosynthesis. Recently we have characterized the evolutionary history of HSP47 [1]. Human HSP47 is associated with several human diseases like a familial connective tissue disorder, known as *Osteogenesis imperfecta* (OI) [5], *rheumatoid arthritis* [6] and different cancer types [7]. A missense mutation in the HSP47 protein leads into a lethal form of OI which is triggered by improper regulation of collagen type I and it leads into bone fragilities and deformities, short stature, and shortened lifespan and the higher risk of bone fractures [8]. To understand further roles of HSP47, there is a need to evaluate mutational profiles of HSP47, expression patterns of HSP47 in human diseases and building understanding on HSP47-protein interaction partners. These are not a single study known until today, which focuses on these issues. Hence, an investigation focusing on these aspects of HSP47 is warranted. Herein, we depicted mutational hotspots for disease perfectives focusing using population genomics-based mutation resources such as 1000Genomes [9] and gnomAD [10] and cancer genomics-based mutation resources such as COSMIC version 86 [11], cBioPortal [12] and CanVar [13]. We also examined the expression pattern of HPS47 in the different cancer types. We have also constructed interaction maps of HSP47 and identified top 20 interactions partners and most of these are different heat shock proteins.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

2. Results

2.1. Overview of genetic variants of human HSP47 from the 1000 genomes dataset

There are 888 genetic variants for the full-length transcript (Ensembl Id. ENST00000533603) of human HSP47, deduced from 1092 human genomes analysis (**Table S1**). Top six variants types are intron variants (335), downstream gene variants (191), upstream gene variants (172), missense variants (82) and synonymous variants (55). Majority of these genetic variants are SNPs (**Figure 1**), which constitutes 88% of the total and remaining three major stockholders are somatic SNV (6%), deletion (2.8%) and insertion (2.4%).

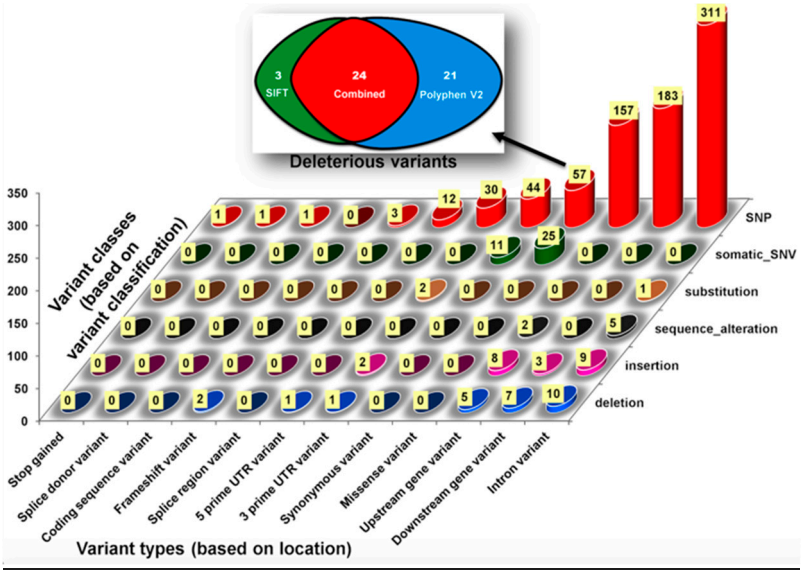


Figure 1. SNPs are major stakeholders of human genetic variants of HSP47 deduced from 1000G data. Deleterious nature of HSP47 missense variants were predicted using SIFT [14] and PolyPhen V2 [15].

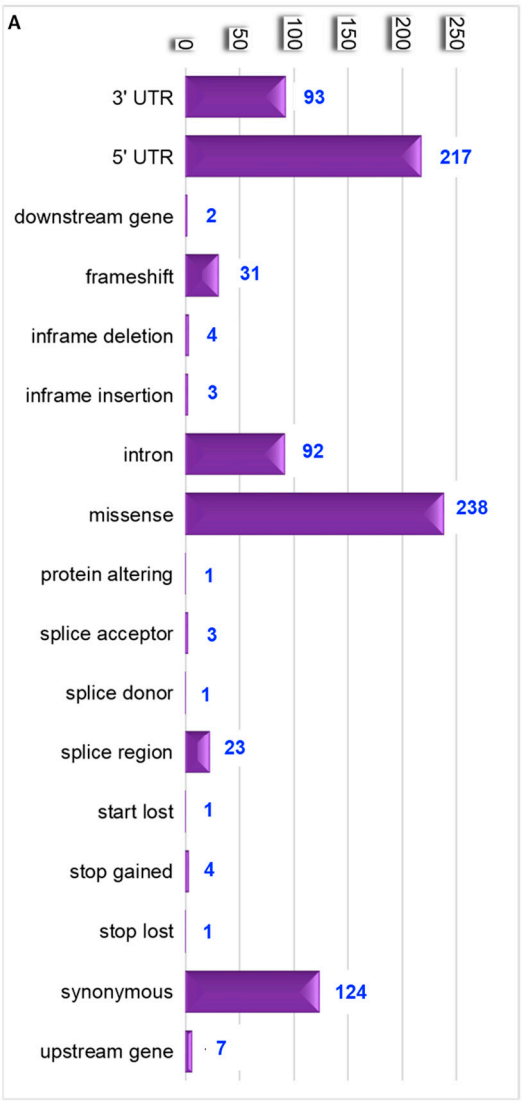
We have identified a total of 82 missense variants, which are causing 75 mutations in the human HSP47 protein sequence (**Table 1**). We have computed the deleterious nature of these missense variants using SIFT [14] and PolyPhen V2 [15] and there are 24 deleterious variants predicted both by SIFT (score <0.06) and PolyPhen V2 (score >0.45), while 21 are predicted deleterious by PolyPhen V2 only and 3 are only predicted by SIFT only (**Table 1**). We considered single tool prediction as partly deleterious nature, which is 24 in total.

2.2. Summary of missense variants of human HSP47 from gnomAD dataset

We deduced 828 HSP47 variants (Figure2A and Table S2) from the Genome Aggregation Database (gnomAD [10]) with top four variant types are missense (238, 28%), 5' UTR (217, 26%), synonymous(124, 15%) and 3' UTR (93, 26%). Of 238 missense variants, 67 were highly deleterious with CADD score >20 (Figure 2B). CADD score >10 >20 and >30 means top 10% top 1% and 0.1% probable functional variants in entire human genome, respectively [16]. Out of 67 highly deleterious variants, 61 missense variants have CADD>25 (yellow shade in Figure 2B). We applied Grantham score to these variants and at the CADD score >25 and Grantham score ≥151 (radical), only 9 missense variants have remained as top-ranked variants - Ser76Trp, Arg103Cys, Arg116Cys, Ser159Phe, Arg280Cys, Arg339Cys, Arg377Cys, Ser399Phe and Arg405Cys (yellow shade in Figure 2B). Most of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

these top deleterious mutations are arginine to cysteine changes. These mutations are described in details in the next sections.



B

Mutations	Structural location	Grantham score	CADD score
Leu43Phe	Nt	22	23.4
Ser47Gly	hA	56	24.6
Leu50Gln	hA	113	27.4
Leu54Phe	hA	22	23.8
Ser70Pro	Lo-s6B-hB	74	25.4
Pro71Ala	Lo-s6B-hB	27	24
Ser76Leu	hB	145	31
Ser76Trp	hB	177	28.6
Ser91Leu	hC	145	29.6
Arg103Cys	Lo-hC-hD	180	34
Leu115Gln	hD	113	26.4
Arg116Cys	hD	180	31
Leu118Pro	hD	98	24.3
Arg124Leu	hD	102	31
Arg124Pro	hD	103	29.7
Leu134Pro	s2A	98	25.9
Gly136Arg	s2A	125	26.7
Ser139Leu	Lo-s2A-hE	145	27.9
Phe142Leu	Lo-s2A-hE	22	27.2
Asp144Gly	hE	94	25.9
His153Asn	hE	68	27.6
His153Gln	hE	24	25.7
Cys156Trp	Lo-hE-s1A	215	24.4
Ser159Phe	s1A	155	28.4
Asn162Lys	Lo-s1A- hF	94	28
Asn174Lys	hF	94	26.8
Gly183Ser	Lo-hF-s3A	56	32
Gly183Arg	Lo-hF-s3A	125	31
Val192Ala	s3A	64	25.2
Asn202Lys	s3A	94	26.3
Phe205Val	s3A	50	26.4
Val219Met	hF1	21	32
Asn221Tyr	Lo-hF1-s4C	143	28.2
Arg222His	s4C	29	35
Val226Met	Lo-s4C- s3C	21	31
Val232Gly	s3C	109	29.1
Met237Ile	s3C	1	32
Arg239Trp	s3C	101	35
Gly241Val	s3C	109	29
Leu253Pro	Lo-s1B- s2B	98	28.4
Pro259Thr	s2B	38	28.6
Pro259Leu	s2B	98	32
Leu260Val	s2B	32	25.6
Glu276Lys	Lo-s3B-hF1	56	33
Glu276Ala	Lo-s3B-hF1	107	28.1
Arg280Cys	hG	180	34
Arg280Ser	hG	110	33
Arg280His	hG	29	34
Thr314Ile	s6A	89	28.2
Leu326Val	Lo-hI-hI1	32	27.9
Glu328Lys	hI1	56	34
Arg339Cys	Lo-hI1-s5A	180	32
Ala349Asp	Lo-hI1-s5A	126	32
Phe352Tyr	s5A	22	27.3
His353Arg	s5A	29	25.9
Arg377Cys	s4A/RCL	180	34
Ala384Thr	Lo-s1C-s4B	58	33
Ser399Phe	s5B	155	29.9
Ser399Tyr	s5B	144	29
Leu400Gln	s5B	113	28.6
Arg405Cys	s5B	180	34
Arg408Trp	s5B	101	34
Asp412Asn	Ct	23	29.2
Arg415Gly	Ct	125	28.6
Arg415Gln	Ct	43	32
Asp416Glu	Ct	0	26.1
Glu417Lys	Ct	56	34

Figure 2. Overview of HSP47 variants deduced from the Genome Aggregation Database (gnomAD [10]**).**

A. Location-wise distribution of HSP47 variants.

B. Deleterious nature of missense HSP47 variants using Grantham and CADD scores. Most deleterious missense variants were marked with yellow shade with Grantham score (≥150) CADD score (>25). Orange shade for missense variants shared by 1000 genomes data.

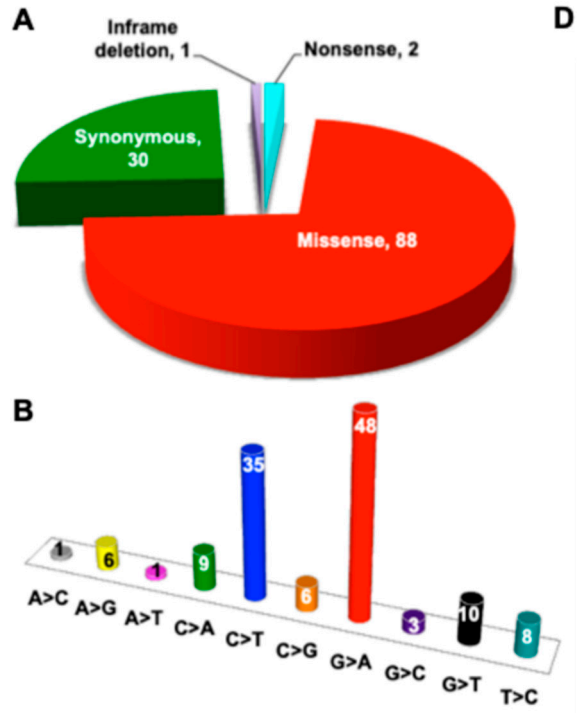
Ct – C-terminal end; h -α-helix; Lo – loop; Nt – N-terminal end; RCL - reactive center loop; s -β-sheet. Grantham scores - 0–50 - conservative, 51–100 - moderately conservative, 101–150 - moderately radical and ≥151 - radical.

2.3. Summary of HSP47 mutations in human cancer

We have evaluated the HSP47 mutations in different cancers using three different cancer mutational resources as COSMICv86, cBioPortal, and CanVar. COSMICv86 has 119 mutations in HSP47 surpassing various cancer types (**Figure 3 and Table S3**). These mutations are found in 121 cancer patients with 88 missense mutations, 30 synonymous mutations, 2 nonsense mutation and 1 inframe deletion (**Figure 3A**). Major types of nucleotide changes are G>A, C>T, G>T, and C>A (**Figure 3B**) with mutations 48 (40.34%), 35(29.41%), 10 (8.40%) and 9 (7.56%). On examining missense mutations,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

we found that these mutations are found in 18 different cancer types with top five being cancers of large intestine (23 mutations), liver (10), stomach (9), lung (8) and oesophagus (7) (**Figure 3C**). Upon computing deleterious nature of these COSMIC missense variants of HSP47 with CADD score (>20) and Grantham score (>50), we found 50 deleterious missense mutations on the various locations of HSP47 protein in various cancers (**Figure 3D**).



D

Mutations	Structural Location	CADD score	Grantham score	Cancer types (#samples)
Arg2Cys	Nt	22,3	180	St (1)
Ala19Thr	Nt	22	58	Pc (1)
Ala40Val	Nt	23,3	64	Li (1)
Gly49Asp	hA	21,3	94	Li (1)
Val69Ala	s6B	25,2	64	Pc (1)
Ala75Ser	hB	32	58	Liv (1)
Ala75Thr	hB	25,6	99	Th (2)
Ala90Thr	hC	32	58	En (1); Li (1)
Ser91Leu	hC	25,4	145	Ut (1)
Ala99Thr	Lo-hC-hD	27,9	58	Li (1)
Glu100Lys	Lo-hC-hD	24	56	En (1)
Asp104Gly	Lo-hC-hD	25,4	94	Hly (1); Pc (1)
Leu134Pro	s2A	34	98	Oe (1)
Gly136Arg	s2A	33	125	Liv (1)
Ser159Cys	s1A	31	112	Li (1)
Arg167Cys	hF	33	180	Hly (1)
Ala169Thr	hF	25,1	58	Li (1)
Glu175Lys	hF	32	56	Li (1)
Ala178Ser	hF	22,3	99	Sk (1)
Ala178Val	hF	24	64	Ut (1)
Asp182Asn	hF	33	81	Li (1); Lu (1)
Gly183Ser	Lo-hF-s3A	33	56	Oe (1); St (1)
Arg194Ser	s3A	23,1	110	Li (1); Br (1); Hly (1)
Thr195Met	s3A	22	81	Cns (1)
Ala198Thr	s3A	28,2	58	Li (1); Th (1)
Leu200Pro	s3A	27,1	98	Bt (1)
Arg228Trp	Lo-s4C-s3C	34	101	Sk (1)
Ser229Phe	Lo-s4C-s3C	23,6	155	Sk (1)
Thr231Ile	s3C	24,6	89	Bt (1); Sk (1)
Met235Thr	s3C	21,7	81	Liv (1); Ut (2)
Glu249Lys	s1B	23,4	56	Li (1)
Glu251Lys	s1B	24,1	56	Bt (1)
Lys252Asn	Lo-s1B-s2B	22,5	94	Uat (1)
Ala261Thr	s2B	24,3	58	Li (1)
Glu279Lys	hG	26	56	Li (1)
Arg280Cys	hG	32	180	Pc (2)
Trp293Leu	hG	35	61	Lu (1)
Trp293Cys	hG	34	215	Liv (1)
Ala303Thr	s2C	24,1	58	Br (1)
Ser305Tyr	s2C	28,3	144	Oe (3)
Leu321Pro	hI	34	98	En (1)
Gly323Trp	hI	34	184	Lu (1)
Leu324Pro	hI	35	98	Li (1); Ut (1)
Lys343Thr	Lo-hI1-s5A	34	78	Sk (1)
His353Asn	s5A	27,2	68	Li (1)
Ala356Thr	s5A	25	58	St (2)
Arg373Cys	s5A/RCL	28	180	Lu (1); Pc (1)
Arg373Pro	s5A/RCL	23,2	103	En (1)
Arg377Cys	s5A/RCL	34	180	Li (2)
Gly411Val	Ct	28,7	109	Liv (1)

Low High

Figure 3. Overview of somatic mutations of HSP47 identified from the COSMIC v86 database.

A. Overview of mutational types reveal large fraction is missense variants.

B. Summary of nucleotide changes that generates mutational profile for HSP47.

C. Summary of missense variants identified in samples of different cancers.

D. Top-ranked deleterious missense variants of HSP47 computed using CADD score (>20) and Grantham score (>50) with summary of cancer type and number of samples possessing deleterious variant. Ct – C-terminal end;

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

h -α-helix; Lo – loop; Nt – N-terminal end; RCL - reactive center loop; s -β-sheet. Grantham scores - 0–50 - conservative, 51–100 - moderately conservative, 101–150 - moderately radical and ≥151 - radical.

Using the cBioPortal, we identified 168 cancer mutations of HSP47 (**Table S4**), which included 163 missense mutations (**Table 2**), five nonsense mutations, three fusion mutations and one each of frameshift deletion and in frame deletion (**Table S4**). These mutations were found in 45 different types of cancers (**Tables 2 and S4**) with top cancer types with HSP47 mutations are 26 mutations for uterine endometrioid carcinoma, 15 each cutaneous melanoma and bladder urothelial carcinoma, 14 mutations in stomach adenocarcinoma and 12 mutations in colorectal adenocarcinoma (**Tables 2 and S4**). There are 91 different missense mutations of which are 43 are predicted as deleterious mutations (**Tables 2 and S4**). Upon applying the strict cut-off of the CADD score >25 and Grantham score ≥151 (radical), we found only 6 highly ranked missense variants as top-ranked variants - Arg167Cys, Arg280Cys, Trp293Cys, Gly323Trp, Arg373Cys and Arg377Cys (**Figure 3D**). Additionally, there are three nonsense mutations as Glu251*, Gln368* and Glu375* spanning to five different cancer types (**Table S4**). CanVar is a specific database of colorectal cancer samples and there are only 2 deleterious missense variants of HSP47 (**Table 3**).

2.4. Overview of deleterious mutation of human HSP47 deduced from various resources

We have identified a comprehensive list of deleterious or pathogenic mutations using five different resources. **Figure 5** depicts the sharing pattern of deleterious missense mutations of HSP47 in these 5 resources.

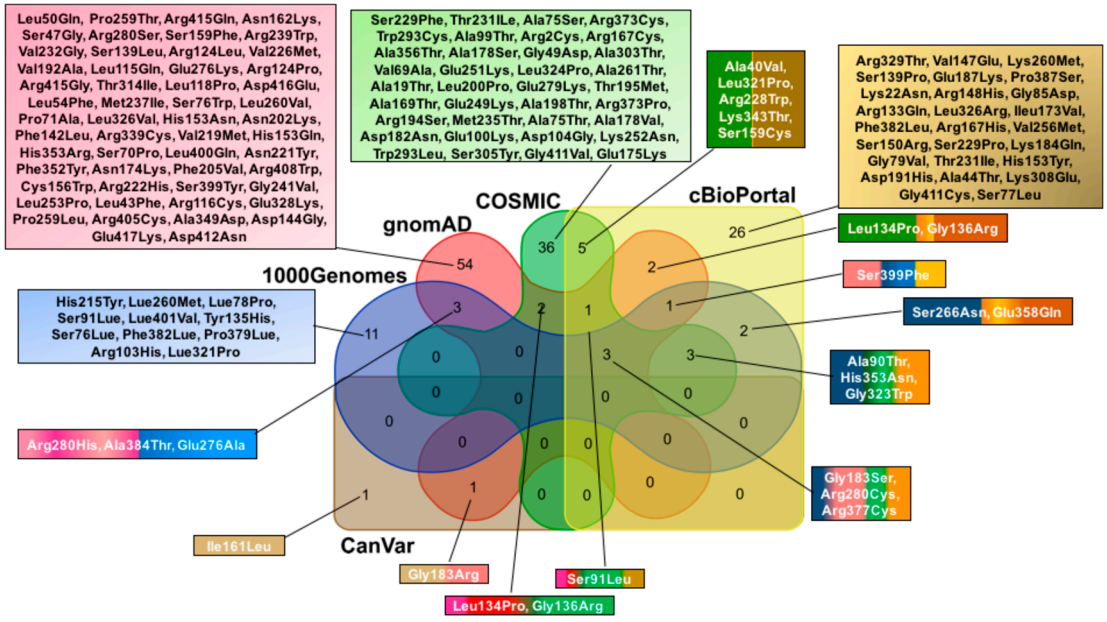


Figure 4. Distribution of deleterious missense SNPs of HSP47 derived from different resources like population genomics databases (1000Genomes [9] **and gnomAD** [10]**) and cancer databases (COSMIC version 86** [11]**, cBioPortal** [12] **and CanVar** [13]**).**

Only 3 deleterious missense variants were shared by these four databases (1000G-COSMIC-cBioPortal and gnomAD). While one, three, one variant was shared by a combination of three resources like com1000G-cBioPortal-gnomAD, 1000G-COSMIC-cBioPortal, COSMIC-cBioPortal-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

gnomAD, respectively. Similarly, combination of two databases such as COSMIC-cBioPortal and 1000G-gnomAD share 5 and 3 deleterious missense variants, respectively; whereas 2 each are shared by 1000G-cBioPortal, COSMIC-gnomAD, cBioPortal-gnomAD and one by CanVar-gnomAD. Each database has unique missense variants of HSP47, which are not shared by other databases like 1000G, gnomAD COSMIC, cBioPortal, and CanVar have 11, 54, 36, 26 and 1 deleterious missense mutations, respectively.

Upon computing top-ranked deleterious variants from all these resources using the stringent cut-off of the CADD score >25 and Grantham score ≥151, we found thirteen top-ranked deleterious missense mutations as Ser76Trp, Arg103Cys, Arg116Cys, Ser159Phe, Arg167Cys, Arg280Cys, Trp293Cys, Gly323Trp, Arg339Cys, Arg373Cys, Arg377Cys, Ser399Phe, and Arg405Cys. Eight out of these thirteen are arginine mutated into cysteine. We performed structural comparisons of these 13 deleterious mutations using protein models deduced using canine HSP47 (PDB Id - 3ZHA) as a template and structural changes are depicted in **Figure 5**.

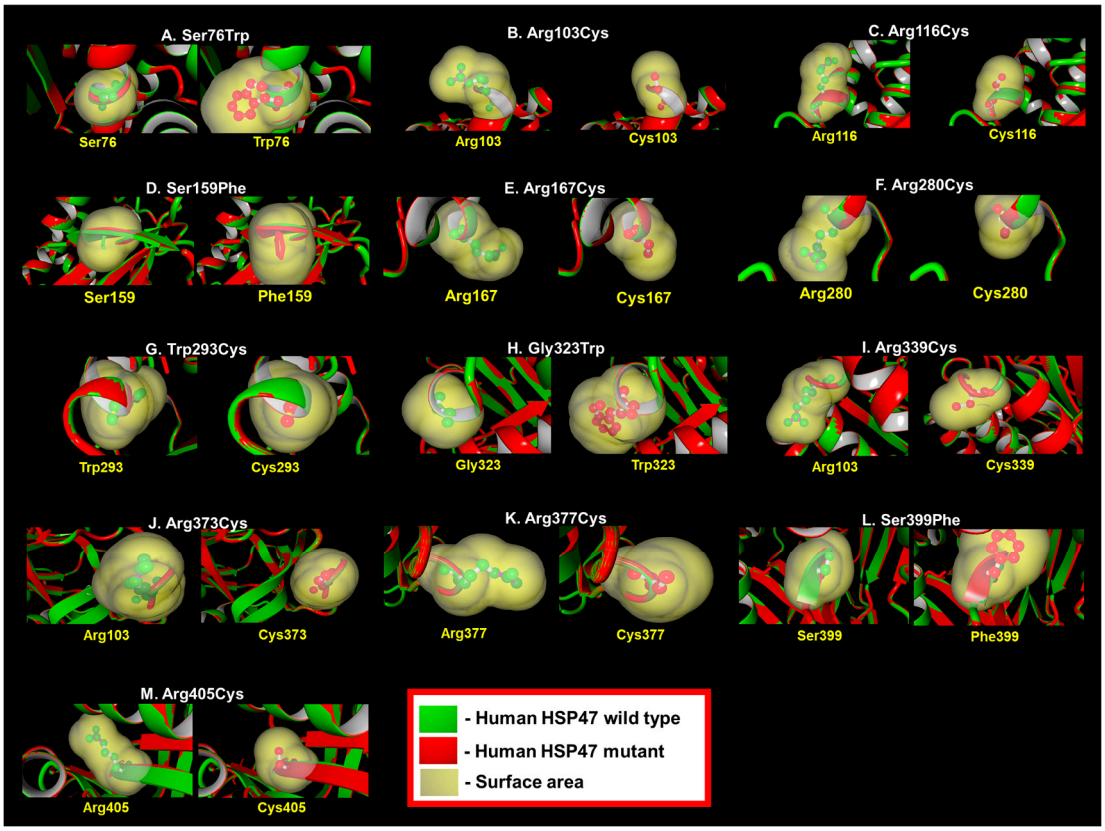


Figure 5. Overview of structural and conformational changes induced by 13 top-ranked deleterious mutations of HSP47. This structural study is based on the homology models of human HSP47 (wild type) and mutant HSP47 deduced using canine HSP47 structure (PDB Id - 3ZHA) as a template with SWISS-MODEL [17] and by superimposing and visualizing in YASARA [18].

Herein, we summarize and describe the implications of deleterious mutations of human HSP47. In the N-terminal segment, and only Leu6Pro is potentially deleterious from the 1000 genomes data, while, the gnomAD dataset has six deleterious mutations as Arg2His, Ala9Thr, Ala19Thr, Lys22Asn, Ala40Val, and Thr42Met (**Figure 6**). The COSMIC database has three deleterious mutations - Arg2Cys, Ala19Thr and Ala40Val with one sample mutated with each mutation in stomach, pancreatic and

 doi:10.20944/preprints201811.0280.v1

large intestine cancers (**Figure 3**). cBioPartal shares two deleterious mutations as Lys22Asn and Ala40Val with one sample each mutated in breast invasive ductal carcinoma and colorectal adenocarcinoma (**Table 2**).

In the core domain of HSP47, there are 24, 62, 46, 40 and 2 deleterious mutations deduced from the 1000 genomes (**Table 1**), gnomAD (**Figures 2 and 6**), COSMIC (**Figure 3**), cBioPortal (**Table 2**) and CanVar (**Table 3**), respectively.

The helix hA has no deleterious variant in 1000 genomes dataset (**Table 1**), but three variants are deleterious in gnomAD as Ser47Gly, Leu50Gln, and Leu54Phe (**Figure 3**), while COSMIC has one deleterious mutation as Gly49Asp with mutation in sample of large intestine (**Figure 3**) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

cBioPortal has one deleterious mutation as Ala44Thr in three samples of esophageal squamous cell carcinoma (**Table 2**).

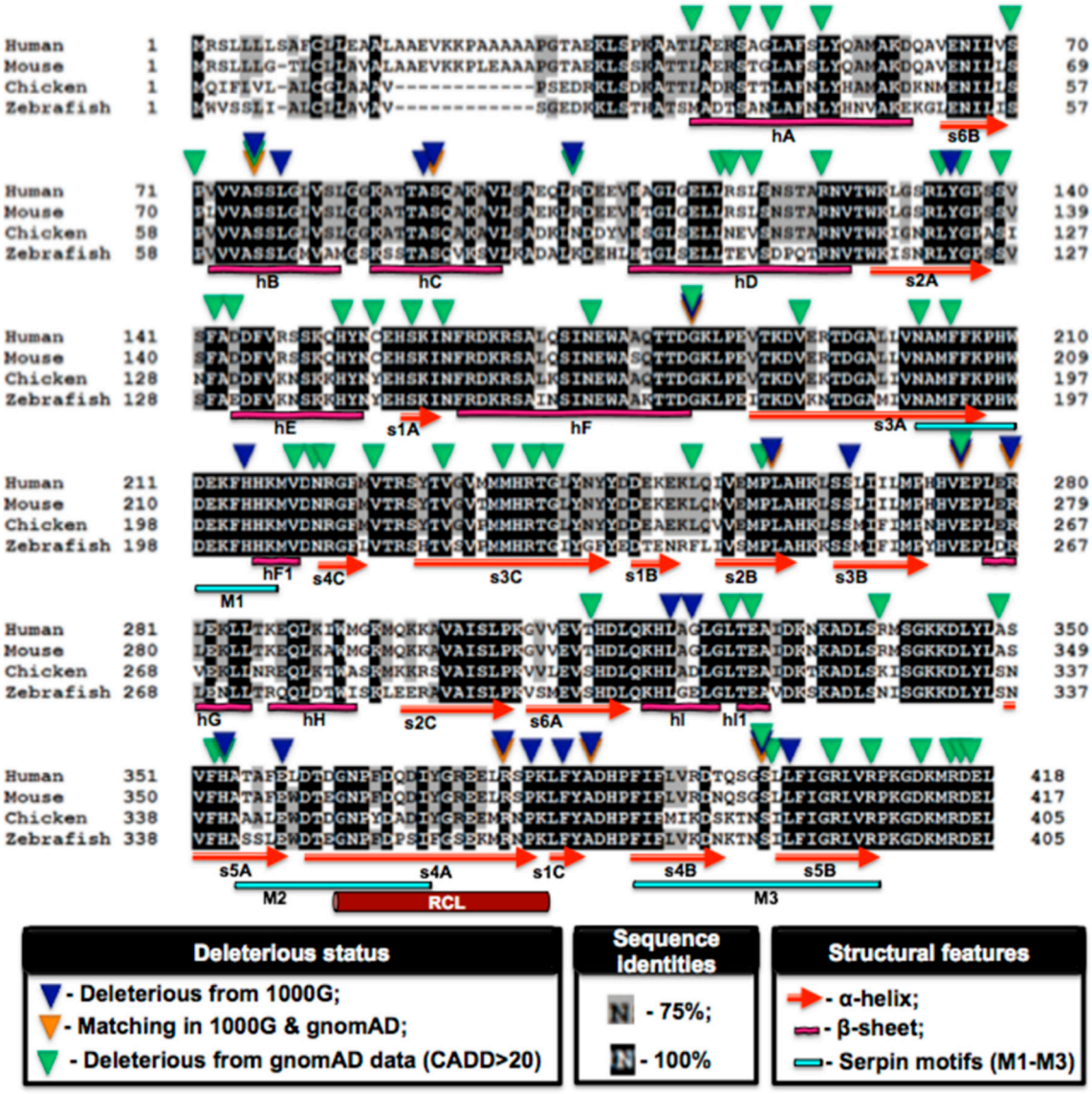


Figure 6. Protein sequence alignment of HSP47 proteins depicting overall deleterious mutational hotspot of HSP47 derived from 1000G and gnomAD [10]**.**

The sheet s6B has a single deleterious mutation in a sample of pancreatic cancer from the COSMIC dataset (**Figure 3**), but none in 1000 genomes (**Table 1**), gnomAD (**Figure 2**) and cBioPortal (**Table 2**). The loop between sheet s6B and helix hB possesses two pathogenic mutation Ser70Pro and Pro71Ala from GnomAD data (**Figure 2**), but none from other datasets. The helix hB has two deleterious mutations as Ser76Leu and Leu78Pro from the 1000 genomes (**Table 1**).

Mutation Leu78Pro will cause instability for the helix hB as generally, prolines are helix breaker residues. Hence, it has severe implication as the Leu78Pro mutation causes degradation of the ER-resident HSP7 via the proteasome and it leads into OI [8]. The gnomAD data has also three missense variants as Ser76Leu and Ser76Trp at the same position and Ser77Leu (**Figure 2**). COSMIC data has two deleterious mutations at the amino acid position 75 as Ala75Ser with one sample of liver cancer

 doi:10.20944/preprints201811.0280.v1

and Ala75Thr with two samples of thyroid cancer, respectively (**Figure 3**). The cBioPortal data has also Ser76Leu with one sample each with cutaneous melanoma and uterine endometrioid carcinoma and two deleterious variants as Ser77Leu with one sample mutated with diffuse large B-cell lymphoma (NOS) and Gly79Val in one sample each with cutaneous melanoma (**Table 2**). Ser76Trp is top-ranked deleterious variant (CADD score >25 and Grantham scores ≥151) and change of serine to tryptophan leads into increment conformational space in HSP47 protein′, depicted using surface area (**Figure 5A**). These mutations in the helix hB are adjutant to known OI-causing Leu78Pro mutation, hence these must be given priorities in future studies.

The loop between helices hB-hC harbors a pathogenic variant as Gly85Asp with one sample of uterine endometrioid carcinoma (1) gathered from the cBioPortal data (**Table 2**)

The helix hC has total three variants with two deleterious mutational sites as Ala90Thr and Ser91Leu from the 1000 genomes (**Table 1** and **Figure 6**). Mutation Ala90Thr will have higher implication as it causes large changes in the middle of the helix hC in comparison to small amino acid changes for Ser91Leu.Ser91Leu mutation is also possessed by gnomAD data (**Figures 2** and **6**).

Deleterious mutation, Arg103His is localized in the loop connecting helices hC and hD, which is also known as CD-loop (**Figure 6** and **Table 1**) and at the same position, there is another deleterious mutation as Arg103Cys found in the gnomAD data (**Figures 2** and **6**) and it is also found mutated in two samples of stomach adenocarcinoma of cBioPortal (**Table 2**). Arg103Cys is top-ranked deleterious mutation (CADD score >25 and Grantham scores ≥151) with shrinkage of total surface area with arginine changing into serine (**Figure 5B**).

This CD loop also harbors three pathogenic variants with first two as Ala99Thr andGlu100Lys with one each sample of large intestine and endometrium cancer (1) and the last variant Asp104Gly mutated in two types of cancers haematopoietic and lymphoid and pancreatic cancer from the COSMIC dataset (**Figure 3**).

The helix hD has only one partly predicted (by SIFT only) pathogenic mutation Ser117Lue. This helix has five mutations at the four locations as Leu115Gln, Arg116Cys, Leu118Pro and Arg124Leu/Arg124Pro in the gnomAD data (**Figures 2** and **6**). Out of these missense variants, Arg116Cys is top-ranked deleterious mutation (CADD score >25 and Grantham scores ≥151) with shrinkage of the total surface area in the helix D (**Figure 5C**).

The sheet s2A has a single deleterious mutation as Tyr135His, this position is occupied by aromatic amino acids (**Figure 6** and **Table 1**). We found two deleterious mutations as Leu134Pro and Gly136Arg in the gnomAD data (**Figures 2** and **6**) and these are matching with COSMIC data with one sample each for esophagus and liver cancer (**Figure 3**). There is one more critical mutation as Arg133Gln, mutated in the uterine endometrioid carcinoma deduced from the cBioPortal data (**Table 2**).

The loop connecting the sheet s2A and helix hE possesses two deleterious mutations as Ser139Leu and Phe142Leu in the gnomAD data (**Figures 2** and **6**) and at the position 139, there is another critical mutation as Ser139Pro, found in one sample of hepatocellular carcinoma of cBioPortal data (**Table 2**)

The helix hE has three critical mutations at two positions Asp144Gly and His153Asn/His153Gln in the gnomAD data (**Figures 2** and **6**), whereas cBioPortal data has 4 different pathogenic variants as Val147Glu, Arg148His, Ser150Arg and His153Tyr with one sample each from tubular stomach adenocarcinoma, colon adenocarcinoma, hepatocellular adenoma and mucinous adenocarcinoma of the colon and rectum (**Table 2**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

The loop connecting the helix hE and the sheet s1A has a single deleterious variant as Cys156Trp from gnomAD data (**Figures 2 and 6**) and. The sheet s1A has a single deleterious mutation as Ser159Phe from gnomAD data (**Figures 2 and 6**) and this missense mutation is in the list of the top 13 deleterious variants (CADD score >25 and Grantham scores ≥151) with noticeable increase in the total surface area by serine to phenylalanine leads into conformational changes in the sheet s1A (**Figure 5D**).

This sheet harbors another mutation at the same position as Ser159Cys with one sample of large intestine cancer (**Figure 3**) and one sample of colorectal adenocarcinoma (**Table 2**). There is another deleterious variant as Ile161Leu deduced from CanVar dataset (**Table 3**). The loop between of the sheet s1A and the helix hF has one deleterious mutation as Asn162Lys from gnomAD data (**Figures 2 and 6**).

The helix hF has a single deleterious mutation as Asn174Lys from gnomAD data (**Figures 2 and 6**).

This helix hF has six deleterious variants found in different cancers as Arg167Cys (haematopoietic and lymphoid cancer), Ala169Thr (large intestine cancer), Glu175Lys (large intestine cancer), Ala178Ser (large intestine cancer), Ala178Val (urinary tract cancer) and Asp182Asn (large intestine and lung cancer) from the COSMIC dataset (**Figure 3**). Two pathogenic mutations are also found in the cBioPortal data (**Table 2**) as Arg167His and Ileu173Val with one mutation each mutated in colorectal and esophagogastric adenocarcinoma. Arg167Cys is a top-ranked deleterious mutation (CADD score >25 and Grantham scores ≥151) and this mutation leads into depletion of the total surface area (**Figure 5E**).

The loop connecting helix hF-sheet s3A has single deleterious missense variant as Gly183Ser (**Table 1** and **Figure 6**). GnomAD data has two mutations at the position 183 as Gly183Ser and Gly183Arg (**Figures 2 and 6**). The second deleterious variant Gly183Arg is also deduced from CanVar dataset (**Table 3**).

The COSMIC data has a single mutation as Gly183Ser with one sample each mutation in oesophagus and stomach cancer (**Figure 3**). The cBioPortal data has 3 mutations as Gly183Ser mutated in stomach adenocarcinoma and Lys184Gln and Glu187Lys – both mutated breast invasive ductal carcinoma (**Table 2**).

The sheet s3A of HSP47 has three pathogenic mutations as Val192Ala, Asn202Lys and Phe205Val found in gnomAD data (**Figures 2 and 6**). This sheet has one critical mutation as Asp191His with two samples of lung adenocarcinoma derived from the cBioPortal (**Table 2**), while COSMIC data has four pathogenic mutation as Arg194Ser with one each sample mutated for large intestine cancer, breast cancer, haematopoietic and lymphoid cancer, Thr195Met with one sample of central nervous system cancer, Ala198Thr with one each sample mutated for large intestine and thyroid cancer (1) and Leu200Pro with one sample mutated for biliary tract cancer (**Figure 3**).

The loop connecting sheet s3A-helix hF1 has two variants with one deleterious variant as His215Tyr in the highly conserved region (**Figure 6**). The small helix hF1 has a single deleterious mutation with high CADD score (32) as Val219Met from the gnomAD dataset (**Figures 2 and 6**). The loop between the helix hF1 and the sheet s4C has one critical missense variant (Asn221Tyr) with CADD score of 28.2 (**Figures 2 and 6**). Sheet s4C possess one pathogenic mutation as Arg222His (CADD score = 35).

The loop connecting sheets s4C-s3C harbours four critical variants - Val226Met from the gnomAD dataset (**Figures 2 and 6**). COSMIC data has two mutations as Arg228Trp and Ser229Phe with one sample each of skin cancer (**Figure 3**) and two from the cBioPortal data as Arg228Trp and Ser229Pro

 doi:10.20944/preprints201811.0280.v1

with one sample each mutated of hepatocellular carcinoma and cutaneous squamous cell carcinoma (**Table 2**).

Sheet s3C has four pathogenic mutations as Val232Gly, Met237Ile, Arg239Trp, Gly241Val deduced from the gnomAD dataset (**Figures 2 and 6**). The COSMIC has 2 deleterious mutations (CADD score >20) as Thr231Ile with one sample each for biliary tract and skin cancer and Met235Thr possessed by one sample of liver cancer and two samples for urinary tract cancer (**Figure 3**). Thr231Ile is also mutated in the cutaneous melanoma (**Table 2**). The sheet s1B has two variants as Glu249Lys and Glu251Lys are potentially critical possessed by one sample of large intestine and biliary tract cancer as deduced from the COSMIC (**Figure 3**).

The loop connecting sheets s1B-s2B has two critical variants as Lys252Asn (CADD score of 22.5) mutated in a sample of upper aerodigestive tract cancer as found in the COSMIC dataset (**Figure 3**) and Leu253Pro (CADD score of 28.4) is found in the gnomAD dataset (**Figures 2 and 6**).

The sheet s2B has one deleterious variant as Leu260Met, which highly conserved position (>90%, **Figure 6**) in 1000 genomes data and there are three variants at two positions as Pro259Thr/Pro259Leu and Leu260Val. There are 3 mutations with known to be mutated in cancer samples in this sheet with Val256Met and Leu260Met mutated in breast invasive ductal carcinoma and stomach adenocarcinoma, derived from cBioPortal (**Table 2**) and Ala261Thr is mutated in large intestine cancer, deduced from the COSMIC data (**Figure 3**). The sheet s3B has one deleterious mutation as Ser266Asn, in a residue, which is highly conserved with >70% conservation at the start of this sheet (**Figure 6**) and this variant is also found in the cBioPortal data with the mutation in colorectal adenocarcinoma (**Table 2**). The loop connecting the sheet s3B and the helix hG has total 2 critical variants at the highly conserved position 276 - Glu276Ala and Glu276Lys, first one is found in 1000 genomes data (**Figure 6**), while both are present in the gnomAD dataset (**Figures 2 and 6**). Helix hG has two deleterious variants at a highly conserved position 280 as Arg280Cys and Arg280His in the 1000genomes (**Table 2**), while at the same position has three critical variants as Arg280Cys, Arg280Ser, Arg280His deduced from the gnomAD dataset (**Figures 2 and 6**). COSMIC data has total four deleterious mutations in the helix hG as Glu279Lys mutated in one sample of large intestine cancer, Arg280Cys found in two samples of pancreatic cancer, Trp293Leu and Trp293Cys with one sample each for liver cancer (**Figure 3**). The cBioPortal data also has Arg280Cys as deleterious variant as possessed by three samples of head and neck squamous cell carcinoma and also by one sample of ampullary carcinoma, respectively (**Table 2**).

Both Arg280Cys and Trp293Cys are listed into the group of 13 top-ranked deleterious HSP47 mutations. Structurally, these two mutations lead into shrinkage of surface areas at respective positions (**Figures 5F-G**)

The sheet s2C has two pathogenic mutations as Ala303Thr with one sample mutated with breast cancer and Ser305Tyr possessed by three samples of oesophagus cancer, derived from the COSMIC dataset (**Figure 3**). At the position 305, Ser305Pro is found as partly deleterious (only by Polyphen V2) mutation in the 1000gemones data (**Table 2**). The loop between the helix hH and the sheet s2C has a single variant as Met297Ile with partly deleterious nature (only by SIFT, **Table 2**). The loop between sheets s2C-s6A has one critical variant as Lys308Glu with one sample mutated in the

 doi:10.20944/preprints201811.0280.v1

hepatocellular carcinoma, derived from cBioPortal data (**Table 2**). The sheet s6A has a single critical mutation as Thr314Ile, derived from cBioPortal (**Table 2**).

The helix hI has two deleterious mutations at the two highly conserved positions 321 and 323 as Leu321Pro and one as Gly323Trp from 1000genomes data (**Figure 4 and Table 1**), which have matches in the COSMIC dataset (**Figure 3**) with one sample each for endometrium and lung cancer and these two mutations are also found in three samples of uterine endometrioid carcinoma and small cell lung cancer (**Table 2**). There is another pathogenic mutation as Leu324Pro in the COSMIC dataset with mutations in large intestine and urinary tract cancer (**Figure 3**). This mutation is already known to cause OI in dogs [8]. Out of these missense variants of the helix hI, Gly323Trp is a top-ranked deleterious mutation (CADD score - 34 and Grantham score – 182) with the increment of the total surface area at the position 323 (**Figure 5H**). This mutation is critical as it is located penultimate to known IO-causing Leu324Pro.

The loop between helices hl and hl1 has one deleterious mutation as Leu326Val from the gnomAD dataset (**Figures 2 and 6**) and this variant is also found in the cBioPortal data with this mutation in two samples of uterine carcinosarcoma/uterine malignant mixed mullerian tumor (**Table 2**). The helix hl1 has one deleterious mutation as Glu328Lys from the gnomAD dataset (**Figures 2 and 6**). The loop between helix hl1 and the sheet s5A has two deleterious variants as Arg339Cys and Ala349Asp from the gnomAD dataset (**Figures 2 and 6**). The cBioPortal has a deleterious mutation as Arg329Thr with one sample of uterine endometrioid carcinoma (**Table 2**). The sheet s5A has two deleterious variants as Phe352Tyr and His353Arg from the gnomAD dataset (**Figures 2 and 6**). The Loop between helix hI1-sheet s5A has two partly deleterious mutations as Lys332Asn and Arg339Leu deduced from 1000genomes (Only by PolyPhen V2, **Table 2**). This loop has also a highly ranked mutation as Arg339Cys with CADD score - 32 and Grantham score – 180, derived from the gnomAD dataset (**Figures 2 and 6**) and this mutation leads into depletion of the total surface area at the position 339 (**Figure 5I**).

This loop also harbors another deleterious mutation as Lys343Thr with one sample with skin cancer in the COSMIC dataset (**Figure 3**) and two samples with cutaneous melanoma from the cBioPortal dataset (**Table 2**).

Sheet s5A has total 5 variants of which two pose deleterious mutations as His353Asn and Glu358Gln at the highly conserved position from 1000genomes data (**Figure 6 and Table 1**). These two mutations are also found to be mutated in one and four samples of colon adenocarcinoma and bladder urothelial carcinoma (**Table 2**) in the cBioPortal data, whereas COSMIC data has the first mutation found in one sample of large intestine and another critical mutation as Ala356Thr in two samples of stomach cancer (**Figure 3**).

The sheet s4A (within RCL) harbors 6 mutations sites –first four (Pro365Leu, Gly372Arg, Arg373Pro, and Glu375Val) being partly deleterious with only PolyPhen V2 prediction support and two being deleterious as Arg377Cys and Pro379Leu. The COSMIC dataset has 3 mutations as Arg373Cys with one sample each of lung and pancreatic cancer, Arg373Pro with one sample of endometrium cancer and Arg377Cys with 2 samples of large intestine cancer (**Figure 3**). Deleterious mutation Arg377Cys is also in the gnomAD dataset (**Figures 2 and 6**) and in the cBioPortal dataset with one sample each in colorectal adenocarcinoma and B-lymphoblastic leukemia/lymphoma (**Table 2**). Two arginine to cysteine mutations (Arg373Cys Arg377Cys) are highly deleterious in nature (with very high CADD and Grantham scores) with causes depletion in surface areas at respective positions in the sheet s4A

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

(within RCL, **Figure 5J-K**). The sheet s1C has single deleterious mutation as Phe382Leu, to a highly conserved position just after RCL from the 1000 genomes (**Table 1**) and this variant is also shared in the cBioPortal dataset with two and three samples found in bladder urothelial and uterine endometrioid carcinoma, respectively (**Table 2**). Similarly, the loop joining sheets s1C-s4B harbors a single deleterious variant as Ala384Thr from the 1000 genomes (**Table 1**) and gnomAD dataset (**Figures 2 and 6**). This loop also harbors a pathogenic mutation (Pro387Ser) with one sample in cutaneous melanoma from the cBioPortal data (**Table 2**). The Sheet s5B harbors two deleterious mutations (Ser399Phe and Leu401Val) and five deleterious mutations at four positions (Ser399Phe/Ser399Tyr, Leu400Gln, Arg405Cys, Arg408Trp) from the 1000 genomes (**Table 1**) and the gnomAD (**Figures 2 and 6**), respectively. The pathogenic variant Ser399Phe was found in one sample each of cutaneous squamous cell carcinoma and cutaneous melanoma from the cBioPortal data (**Table 2**). Two missense variants of the loop joining sheets s1C-s4B is grouped in the top 13 most deleterious variants list with very high CADD (>25) and Grantham (≥151) scores as Ser399Phe and Arg405Cys with increase and decrease in the surface area due to mutations at the positions 399 and 405, respectively (**Figures 5L-M**). At the C-terminal end, there are five deleterious variants as Asp412Asn, Arg415Gly, Arg415Gln, Asp416Glu, Glu417Lys from gnomAD (**Figures 2 and 6**), while at the position 411, there are two pathogenic variants as Gly411Val and Gly411Cys found mutated in one sample each of the liver cancer and cutaneous melanoma deduced from the COSMIC (**Figure 3**) and cBioPortal (**Table 2**), respectively.

2.3. Expression of HSP47 in different cancers tissues and normal tissues

Based on the mutational profile of HSP47, it has roles in different cancers. It is also important to know, what is the expression pattern of HSP47 in different cancer types. To evaluate expression patterns of HSP47, we extracted data from the database of differential expression of protein in cancer, dbDEPC 3.0 [19]. In eleven types of cancer, HSP47 is up-regulated with top four cancers based on the number of experiments - meningioma, colorectal cancer, hepatocellular carcinoma and breast cancer (**Figure 7 and Suppl. Table S5**). HSP47 is down-regulated in chordoma, lung adenocarcinoma and urinary bladder neoplasms (**Figure 7 and Suppl. Table S5**).

This also leads us to think, how are the expression patterns of HSP47 in different normal human tissues. To evaluate expression pattern, we have scanned three large resources of human gene expression datasets as human protein atlas (HPA, https://www.proteinatlas.org/), genotype-tissue expression (GTEx https://gtexportal.org) and FANTOM5 project (http://fantom.gsc.riken.jp/5/).

Upon evaluating protein level expression of HPS47 using the HPA resource, we found that human HSP47 protein is highly expressed in in the normal tissues of lung, kidney, breast, endometrium, ovary and placenta (**Figure 7B**), whereas expression of HPS47 is ranged in medium level in tissues of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

tonsil, smooth muscle, oral mucosa, esophagus, testis, vagina, cervix (uterine), soft tissue, and skin (**Figure 7B**).

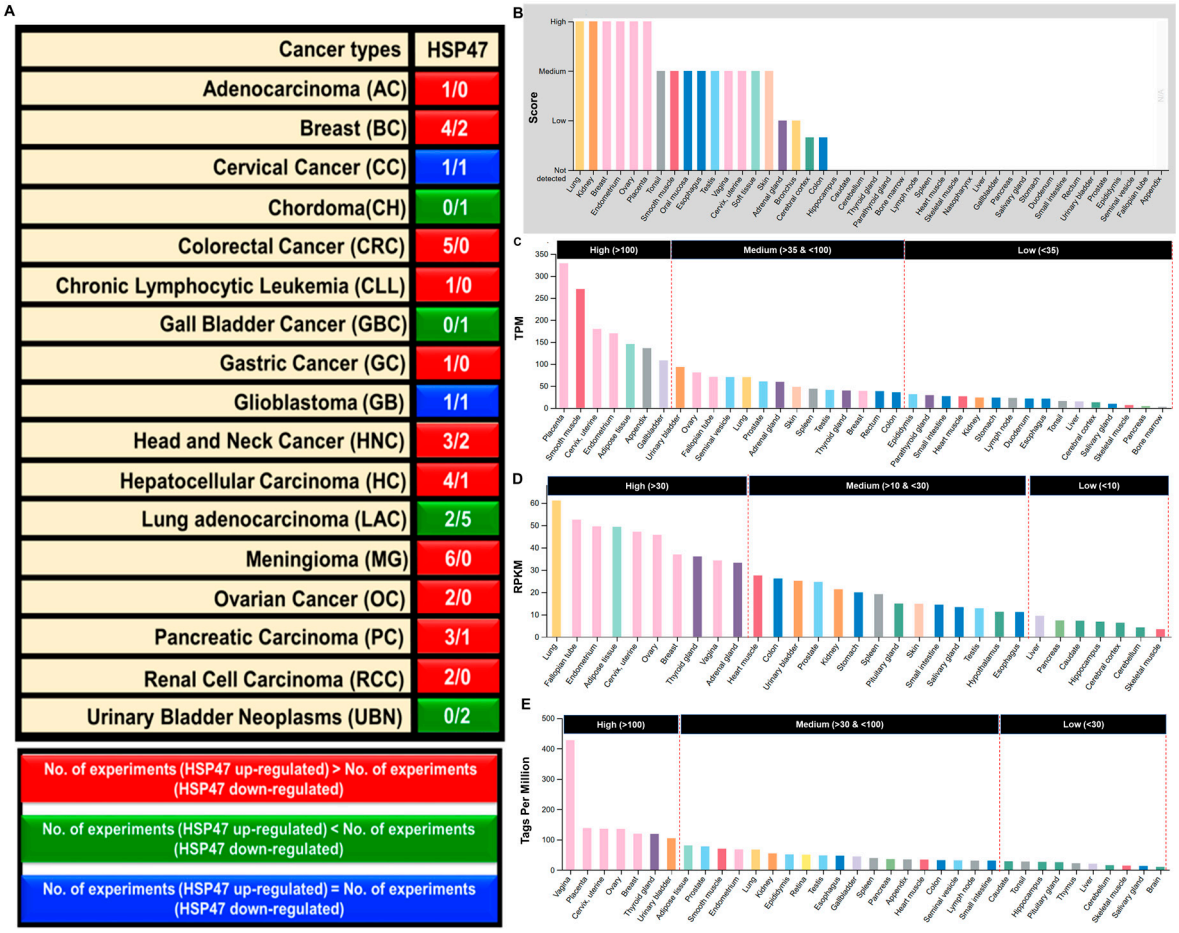


Figure 7. Overview of different expression patterns of human HSP47 in cancer and different normal tissue types.

A. Summary of HSP47 expression pattern in different cancer types. This expression pattern was deduced from dbDEPC 3.0 [19].

B. Summary of human HSP47 protein expression patterns in different normal tissues derived from human protein atlas (HPA, https://www.proteinatlas.org/).

C. Overview of of human HSP47 expression patterns using RNA-Seq data from HPA and expression values are depicted as mean transcripts per million (TPM), corresponding to mean values of the different individual samples from each tissue types.

D. Summary of RNA-seq based HSP47 expression patterns in different normal tissues and these values are shown as median reads per kilobase per million mapped reads (RPKM), derived from the genotype-tissue expression (GTEx https://gtexportal.org) datasets.

E. Overview of expression pattern of HSP47 in normal human tissues are reported as tags per million extracted through cap analysis of gene expression (CAGE) in the FANTOM5 project data (http://fantom.gsc.riken.jp/5/) Similar functional tissue groups are same color codes in B-E.

Low level of expression of HSP47 was found in tissues of adrenal gland, bronchus, cerebral cortex and colon (**Figure 7B**). We examined RNA-Seq data for HSP47 from the HPA resource, placenta tissues have highest expression level with 329.1 transcripts per million (TPM), whereas other normal

 doi:10.20944/preprints201811.0280.v1

tissues with higher level of expression pattern (>100 TPM) are smooth muscle (270.5 TPM), cervix, (uterine, 179.6 TPM), endometrium (169.8 TPM), adipose tissue (145.4 TPM), appendix (135.9 TPM) and gallbladder (108.8 TPM). Fourteen tissues have medium levels of HSP47 expression (<100 and >35 TPM) with top 2 being urinary bladder (93 TPM) and ovary (80.6 TPM) and the last two being rectal (38 TPM) and colon tissues (35.4 TPM). Sixteen tissues have low levels of HSP47 expression (<35 TPM) with top 2 being epididymis (31.1 TPM) and parathyroid gland (29 TPM) and the last two being pancreas (4 TPM) and bone tissues (2.2 TPM).

Using FANFOM5 dataset, we found that HSP47 is highly expressed (>100 tags per millions) in 7 normal tissues originating from vagina, placenta, cervix (uterine), ovary, breast, thyroid gland and urinary bladder (**Figure 7E**). We also found medium (>100 tags per millions) and lower levels of HSP47 expression in 19 and 10 tissues types, respectively (**Figure 7E**). Overall, we have found HSP47 expression patterns in several normal tissues using three different publicly available datasets.

2.4. A Cocktail of different chaperones interact with HSP47

To evaluate the protein interaction partners of HSP47, we have constructed the interactome map of human HSP47 protein. Remarkably top 20 protein-protein interaction partners (confidence score ≥

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

0.9) are different types of molecular chaperones (**Figure 8A**), which is clear from their names as they contain heat shock protein of specific kDa and member numbers (**Figure 8B and Table 4**).

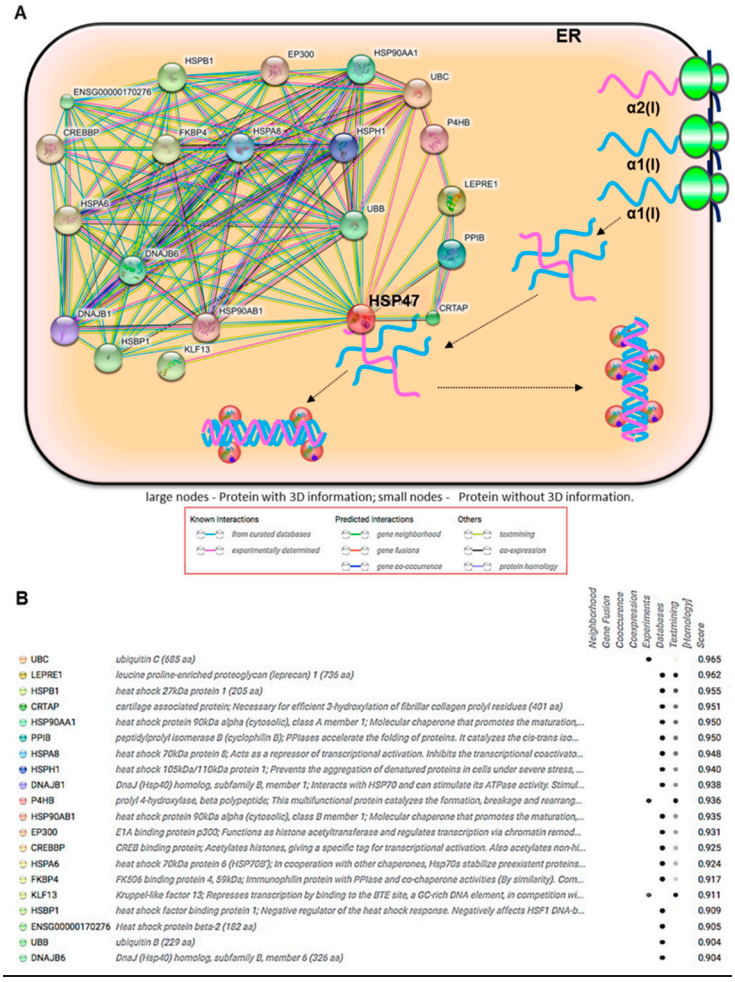


Figure 8. Protein interactome network of human HSP47 reveals several molecular chaperones are interaction partners for HSP47. This network is produced with help of STRING 10 [20] with confidence score > 0.9.

A. Interactome of human HSP47 protein.

B. Details of top protein-protein interaction partners of heat shock protein 47 (HSP47) with their confidence scores.

This suggested that a cocktail of different molecular chaperones is essential for the physiology of HSP47 in the endoplasmic reticulum (ER). These HSP47 interaction partners involve two paralogs involved in histone acetylations such as CREB binding protein (CREBBP) and E1A binding protein p300 (EP300) [21]. These two proteins are closely related size such as CREBBP and EP300 is 2442 and 2414 residues long, respectively and these two proteins possess multiple Pfam domains such as Zf-TAZ (Pfam ID - PF02135.15), KIX (PF02172.15), Bromo (PF00439.24), unknown domain (PF06001.12), HAT_KAT11 (PF08214.10), ZZ (PF00569.16) and other unknown domain (PF09030.9), respectively (**Figure 9A-B**). These two proteins work as histone acetyltransferases and they regulate transcription and/or cell cycle progression by modulating the chromatin structure [21]. These two prominent chromatin remodelers, which operate as scaffolds, which stabilize other protein-protein partners with the transcription complex and these are involved in crucial physiological roles such as development,

 doi:10.20944/preprints201811.0280.v1

growth, and homeostasis [21]. *CREBBP* and *EP300* genes are localized in the human genome (**Table 4**) on the chromosomes 16 (cytoplasmic band 16p13.3) and 22 (22q13.2), respectively. Mutations in these genes cause a rare neurodevelopmental syndrome of known as the Rubinstein-Taybi syndrome (RSTS, OMIM #180849, #613684), which is characterized by errors in facial appearance, skeletal and dysmorphic abnormalities, microcephaly, enlargement of thumbs and first toes, and impaired intellectual and postnatal growth [22].

Cartilage associated protein (CRTAP) is 401 amino acids long without any known protein domain (**Figure 9C**). It is encoded by *CRTAP* gene is localized on the human chromosome 3 (cytoplasmic band 3p22.3, **Table 4**). CRTAP forms the collagen prolyl 3-hydroxylation complex with P3H1 and cyclophilin B (CyPB) in the ER, which 3-hydroxylates the pro986 residue of α1(I) and α1(II) collagen chains [23]. It is associated with a small percentage (5–7%) of patients with severe to lethal OI types VII (OMIM # 610682) and there are five known mutations in *CRTAP* gene, which lead into either prevention of production of any cartilage associated proteins, or reduction in the production of cartilage associated proteins. Irregularities in the production of cartilage associated proteins cause problems in formation of collagen, which ultimately results into the severe form of OI [23].

There are two hsp40 proteins are in this list of HSP47 interaction partners as DnaJ (Hsp40) homolog subfamily B member 1 (DNAJB1) and member 6 (DNAJB6) of residue size of 326 and 340. DNAJB1 possesses 2 protein domains as DnaJ (PF00226.31) in the N-terminal end (4-65 residues) and DnaJ_C (PF01556.18) in the C-terminal end (164-323 residues) while DNAJB6 only harbors DnaJ (PF00226.31) in the N-terminal end (3-66) (**Figure 9D-E**). These two proteins are encoded by genes *DNAJB1* and *DNAJB6* and these genes are localized on human chromosomes 19 (19p13.12) and 7 (7q36.3) (**Table 4**). J-domain is highly conserved domains amongst hsp40 proteins, which is associated with protein folding and protein disaggregation with partnering with hsp70 [24,25]. These two proteins are associated with human diseases caused by errors in protein folding [26,27].

FK506-binding protein 4 (FKBP4) is 59 kDA immunophilin protein. FKBP4 protein is 459 amino acids long composed of two FKBP_C (PF00254.28) domains in the regions of 44-134 and 162-249 residues and two tetratricopeptide repeat domains (TPR_1, PF00515.28, and TPR_2, PF07719.17) in the regions of 321-352 and 354–386 (**Figure 9F**). These tetratricopeptide repeats (TPRs) are required for interactions with HSP70 and HSP90 as co-chaperones [28]. With these properties, FKBP4 is involved in protein folding and cellular trafficking [28]. This protein is encoded by *FKBP4* gene mapped in the region of 12p13.33 on the human chromosome 12 (**Table 4**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

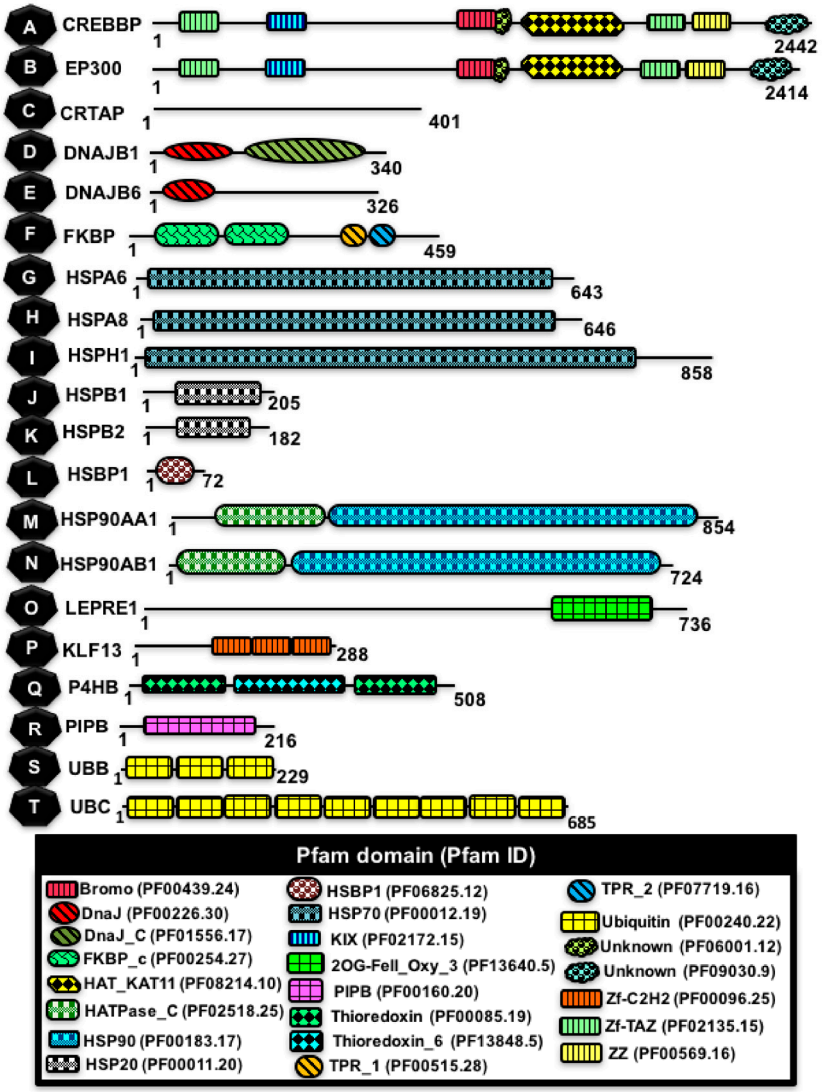


Figure 9. Overview of protein domain architecture of top 20 proteins interacting with HSP47. Pfam protein domains and corresponding Pfam IDs are listed in the box. A.CREBBP - CREB binding protein; B.EP300 - E1A binding protein p300;C.CRTAP - Cartilage associated protein;D.DNAJB1 - DnaJ (Hsp40) homolog subfamily B member 1; E.DNAJB6 - DnaJ (Hsp40) homolog subfamily B member 6: F.FKBP - FK506-binding protein 4;G.HSPA6 - Heat shock 70kDa protein 6; H.HSPA8 - Heat shock 70kDa protein 8; I.HSPH1 - Heat shock 105kDa/110kDa protein 1 J.HSPB1 - heat shock protein beta-1; K.HSPB2 - heat shock protein beta-2; L.HSBP1 - Heat shock factor binding protein 1; M.HSP90AA1 - Heat shock protein Hsp 90-alph(cytosolic), class A member 1;N.HSP90AB1 - Heat shock protein Hsp 90-alpha (cytosolic), class B member 1; O.LEPRE1 - Leucine proline-enriched proteoglycan (leprecan) 1; P.KLF13 - Kruppel-like factor 13; Q.P4HB - Prolyl 4-hydroxylase beta subunit; R.PIPB - Peptidyl-prolyl isomerase B; S.UBB - Ubiquitin B;T.UBC - Ubiquitin C.

There are two HSP70 homologs as interaction partners of HSP47 as heat shock 70kDa protein 6 **(**HSPA6) and HSPA8 (also known as heat shock cognate 71 kDa protein, Hsc70), both of these proteins harbor HSP70 (PF00012.19) protein domain (**Figure 9G-H**). These two proteins are encoded by the genes - *HSPA6 and HSPA8*, which are mapped to chromosomal regions 1q23.3 and 11q24.1 in the human genome, respectively (**Table 4**). These two ubiquitous molecular chaperones (HSPA6 and HSPA8) are members of core Hsp70 machinery and these proteins have critical roles in proper protein

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 November 2018 doi:10.20944/preprints201811.0280.v1

folding, protein degradation, protein translocation across membranes and protein-protein interactions [29]. Another interaction partner of HSP47 network is heat shock 105kDa/110kDa protein 1 (HSPH1), which also contains HSP70 (PF00012.20) in the region of 3 – 704 with total protein length 858 (**Figure 9I**). Gene *HSPH1* mapped on 13q12.3 genomic fragment (**Table 4**), which encodes for HSPH1 protein.

There are two heat-shock protein 27 (HSP27) homologs - heat shock factor binding protein 1 (HSPB1) and 2 (HSPB2) with size 205 and 182 amino acids with a protein domain HSP20 (PF00011.20) in region of 88–183 and 70-162, respectively (**Figure 9J-K**). HSPB1 gene is localized on human chromosome 7 (7q11.23) while HSPB2 is mapped to 11q23.1 region in the chromosome 11 (**Table 4)**. It encodes for an enzyme, which is a member of a heat shock protein family. Under environmental stress, HSPB1 translocate from cytoplasm to nucleus and helps other protein for correct folding. The main role of this gene is the differentiation of a wide range of cell type. Mutation in this gene leads to Charcot-Marie-Tooth Disease, Axonal, Type 2F, and distal hereditary motor neuropathy, Type Iib diseases. HSPB1 is involved in many cellular processes such as apoptosis, thermotolerance, protein disaggregation and cell differentiation and development. HSPB2 has a crucial role in binding and activating myotonic dystrophy protein kinase (DMPK), hence it is also called as called as myotonic dystrophy kinase binding protein (MKBP). This protein HSPB1/MKBP is a major player in maintenances of muscle structure and function [30]. Hsp27 has a highly conserved α-crystallin domain that is enriched with β-sheet structures. The sHsps bind to aggregated proteins in ATP-independent manner and which are subsequently tackled by either by HSP70 system (Hsp70 plus Hsp40 system) or Hsp70/104 bichaperone [31] system for protein disaggregation. Disaggregated proteins either refolded back into native proteins or degraded by autophagy and/or proteasomal system. In addition, Hsp27 recently was shown to be involved in cancer related retinopathy, suggesting its role in developing cancer therapeutics [32].

HSBP1 gene is localized in the genomic fragment of 16q23.3 on the chromosome 16 (**Table 4**), encodes for HSBP1 protein, which is 76 amino acids long with HSBP1 (PF06825.12) domain in the region of 10-60 (**Figure 9L**). HSBP1 is a member of small heat shock proteins (sHSPs) family and this protein prevents the aggregation of denatured and stress-induced misfolded proteins [33].

There are two HSP90 homologs in protein-protein interaction partners as HSP90AA1 (or Hsp90α) and HSP90AB1 (Hsp90β), belong to HSP90 family, which is a well-characterized, well-documented conserved and critical eukaryotic chaperone family [34]. These homologs *HSP90AA1* and *HSP90AB1* are mapped into the human chromosomes 14 (14q32.31) and 6 (6p21.1), respectively (**Table 4**). These two proteins have two types of protein domains such as HATPase_C (PF02518.25) and HSP90 (PF00183.17) in the N-terminal and the C-terminal end (**Figure 9M-N**). HSP90 proteins are required for the proper function of other chaperones. These HSP90 proteins are essential for the maturation, structural maintenance and protein folding, intracellular trafficking, and other signal transduction events [34,35]. HSP90AB1 was shown to be overexpressed during cancer, which prevents misfolding, and degradation of both mutated (for example Ras and p53) and over-expressed oncoproteins (for example p53 and Her2) [36].

Leucine proline-enriched proteoglycan 1 (*LEPRE1*, leprecan) gene is located on the human chromosome 1 (cytoplasmic location 1p34.2) (**Table 4**). LEPRE1 encode to prolyl 3-hydroxylase 1 (P3H1), which is a member of collagen prolyl hydroxylase family with 736 amino acid long and it possesses a single domain of 96 residues long as OG-Fe(II) oxygenase superfamily (2OG-FeII_Oxy_3, PF13640.5) in the

 doi:10.20944/preprints201811.0280.v1

region of 584-661 (**Figure 9O**). PPIB/CyPB plays the instrumental role in the formation of the collagen prolyl 3-hydroxylation complex with P3H1 and CRTAP in the ER [23]. The activity required for proper collagen synthesis and assembly [23]. Mutation in this gene is associated with OI type VIII.

Kruppel-like factor 13 (KLF13) protein is encoded by *KLF13* gene is localized on human chromosome 15 (**Table 4**) and KLF13 protein is 288 amino acids long with three copies of Zf-C2H2 (PF00096.25) domain from mid to the C-terminal end (**Figure 9P)**. It is a member of kruppel-like factors (KLFs) family of Cys2-His2 (C2H2) zinc-finger transcription factors and it has play function in a myriad of physiological roles during cell differentiation and development processes [37].

P4HB gene is localized on human chromosome 17 (cytoplasmic ban 17q25.3), which encodes for prolyl 4-hydroxylase beta subunit (P4HB) protein of size 508 amino acids with 3 protein domains made of two thioredoxin (PF00085.19) in the N-terminal (25-131 residues) and the C-terminal ends 368-472 residues) and one thioredoxin_6 (PF13848.5) in the middle located in 161 – 345 residues (**Figure 9Q**). This protein is a member of the disulfide isomerase family and it is also called protein disulfide isomerase (PDI). P4HB/PDI is the ubiquitously expressed protein which helps in the correction of disulfide bridges in nascent polypeptide chains [38]. Hence P4HB/PDI plays an instrumental role in the protein folding and the cellular concentration of this protein is critical for protein aggregation/disaggregation [38]. Mutations in this protein is involved in a new form of OI-like disorder, known as Cole-Carpenter syndrome [38].

Peptidyl-prolyl isomerase B **(**PPIB) gene is located on human chromosome 15 (cytoplasmic band 15q22.31), which encodes for peptidyl-prolyl isomerase B (PPIB) of size 216 residues with pro_isomerase (PF00160) domain in the region of 47-204 residues (**Figure 9R**) and it is also known as cyclophilin B (CyPB)**.** PPIB/CyPB plays the instrumental role in the formation of the collagen prolyl 3-hydroxylation complex with P3H1 and CRTAP in the ER [23]. Mutational variation in this gene leads to recessive forms of OI. PPIases enzyme helps in that catalysis process of the cis-trans isomerization of proline imidic peptide bonds in proteins and it ultimately assists protein folding and stability [23].PPIB, a member of peptidyl-prolyl cis-trans isomerase (PPIase) has β-barrel structure as cyclophilin and localized inside the endoplasmic reticulum (ER) lumen [39]. Due to its localization to this specialized cellular compartment, it is involved in many biological processes such as post-translational modification and proper folding of proteins such as type I collagen [40].

Finally, there are two ubiquitin proteins, which are interaction partner of HSP47 as ubiquitin B (UBB) and ubiquitin C and these proteins are variable in protein length with 229 and 685 amino acids and similarly these two possess 3 and 9 ubiquitin domains (72 amino acids each; PF00240.22), respectively (**Figure 9S-T**). UBB and UBC are encoded by UBB and UBC genes mapped on chromosomes 17 (17p11.2) and 12 (12q24.31), respectively (**Table 4**). These highly conserved are eukaryotic proteins that are involved in protein ubiquitination, which is a multifaceted dynamic post-translational change with help of the ubiquitin code present in the 72 amino acids of ubiquitin domain [41] with Pfam ID - PF00240.22. Resultant of protein ubiquitination is the clearance of aberrant proteins for their possible degradation by the proteasome and hence, this process is associated with various physiological roles and also with regulations of various signaling pathways [41]. Mutations in these two ubiquitins are related to different human diseases such as Huntington's disease, Alzheimer's disease and polyglutamine disease [42].

 doi:10.20944/preprints201811.0280.v1

3. Discussion

HSP47 is a critical regulator of the collagen maturation and associated embryonic development. However, despite great efforts on discovering the molecular mechanisms and clinical relevance of HSP47 gene and protein functions, also detailed molecular phylogenic analyses was carried out previously [1]. Although mutations of HSP47 have known to play roles in human diseases, yet detailed survey of mutational hotspot was lacking [8]. We have addressed mutational profiles using germline and somatic mutations in various resources.

Eukaryotes have two primary cell types as germ and somatic cells and mutation in these two cell types are called germline and somatic mutations, respectively. Germline mutations are Mendelian mutations, which serve major source for evolutionary change, and it contributes to familial diseases [43]. Similarly, somatic mutations are the primary cause of sporadic diseases including cancer [43]. Mutational hotspot identifications are essential for predicting functionally critical mutations often known as deleterious or pathogenic mutation. The current study has profiled the largest mutational hotspot dataset of HSP47. To our best knowledge, the current study is the most comprehensive study on focusing mutational profiles of HSP47 aided by five different resources such as population genomics-based mutation resources - 1000Genomes [9] and gnomAD [10] and cancer genomics-based mutation resources - COSMICv86 [11], cBioPortal [12] and CanVar [13]. It yielded 24 67, 50, 43 and 2 deleterious mutations from 1000Genomes [9] and gnomAD [10], COSMIC version 86 [11], cBioPortal [12] and CanVar [13]. Large fraction of which are not shared by other databases like 1000Genomes [9] and gnomAD [10], COSMIC version 86 [11], cBioPortal [12] have 11, 54, 36, 26 and 1 deleterious missense mutations, respectively (**Figures 1-4 and 6**). With the strict cut-off of the CADD score >25 and Grantham score ≥151, we identified thirteen top-ranked deleterious missense mutations as Ser76Trp, Arg103Cys, Arg116Cys, Ser159Phe, Arg167Cys, Arg280Cys, Trp293Cys, Gly323Trp, Arg339Cys, Arg373Cys, Arg377Cys, Ser399Phe, and Arg405Cys with the major fraction being arginine-cysteine mutation. These mutations show conformational changes when evaluated using structural comparisons based on protein models deduced using canine HSP47 (PDB Id - 3ZHA) as the template (**Figure 5)**.

To date, this study also provides the largest report on the mutational hotspot analyses of any member of serpin superfamily, previously studied [44-47]. These mutational hotspots will be serving a major resource for understanding of the biology of HSP47, particularly how HSP47 regulates collagen maturations. Structure of canine HSP47 is known and it suggested that HSP47 do not undergo any conformational changes and two HSP47 monomers are required for stabilizing a single collagen triplex and crucial residues of collagen-binding are known. However, known OI causing mutations - Leu78Pro (in humans) and Leu326Pro (in dachshund) are far away from critical collagen binding sites and yet these can impose deleteriousness. This hints for further investigations of roles of HSP47 and its mutational sites.

Since HSP47 is the potential prognostic biomarker in cancer studies [48], these variants serve platforms for investigations for potential roles in the different cancer types. We also found that HPS47 is differentially expressed in different cancers and also in several normal tissues (Figure 6).

Knowledge of protein-protein interactions is crucial for understanding how singling networks functions flanking a protein. We performed interactome map of HSP47 analyses, which revealed that major stakeholders of top interaction partners of HSP47 are other molecular chaperones (**Figure 7**). This list included two CREB binding protein homologs (CREBBP-EP300), two HSP27 homologs (HSPB1-HSPB2), two HSP40 homologs (DNAJB1-DNAJB6), two HSP70 homologs (HSPA6-HSPA8),

 doi:10.20944/preprints201811.0280.v1

two HSP90 homologs (HSP90AA1 and HSP90AB1), two ubiquitin proteins (UBB-UBC), single copy of CRTAP, FKBP, HSPH1, HSBP1, KLF13 P4HB and PIPB. This finding is coinciding with other interactome analyses that chaperones interact with each other in the large interactome, also called as chaperome [49]. Previously, it is known that and CRTAP HSP47, P3H1, and PPIB/CyPB plays the instrumental role in the formation of the collagen prolyl 3-hydroxylation complex in the ER [23]. Several of HSP47 interaction partners also are markers of the panel of osteogenesis imperfecta (Version 1.12) under Rare Disease 100K (https://panelapp.genomicsengland.co.uk/panels/196/). Taken together, our protein-protein network of HSP47 is a critical protein network in the collagen-related disorders. Hence, remaining members of interactions partners must also be given into the consideration for future evaluations.

In conclusion, this study provides the largest repository of mutational hotspots of HSP47. It also sums up the expression pattern of HSP47 in human cancer types. This also reports on top protein-protein interaction patterns, which are cocktails of different heat-shock proteins. These findings will setup newer investigations focusing on the roles of HSP47 in human diseases from perspective of genetic variant and protein-protein network.

 doi:10.20944/preprints201811.0280.v1

669

670 ## 3. Materials and Methods

671 ### 4.1. Characterization of genetic variants of human HSP47 variants from the 1000 genomes project

672 We unraveled the genetic variants of human HSP47 from different human populations available in
673 the 1000 genomes project (phase 3, release 16) [9]. We assessed the impacts of missense variants on
674 the human HSP47 protein using two widely accepted tools, SIFT [14] and PolyPhen V2 [15], as
675 described previously [44-47].

676

677 ### 4.2. Extraction of human HSP47 variants from the gnomAD database

678 We also extracted HSP47 variants from the Genome Aggregation Database (gnomAD [10]) and
679 gnomAD data is derived from 123,136 exomes and 15,496 genomes (website:
680 http://gnomad.broadinstitute.org/).

681

682 ### 4.3. Assessment of HSP47 mutations involved in different cancers

683 To evaluate the HSP47 mutations involved in different cancer types, we extracted HSP47 mutations
684 in cancer samples reported in three different cancer resources namely, COSMIC version 86 [11],
685 cBioPortal [12] and CanVar [13].

686

687 ### 4.4. Ranking HSP47 variants of gnomAD and COSMIC using CADDv1.3

688 These variants were ranked the combined annotation dependent depletion (CADD) v1.3 [16], as used
689 previously [50]. This tool is a comprehensive package for variant ranking at the moment and it
690 measures the levels of deleterious for different variants using CADD PHRED cutoff score >10 >20 and
691 >30 for top 10% top 1% and 0.1% probable functional variants, respectively [16].

692

693 ### 4.5. Building homology models with top 13 mutations of HSP47

694 Out of all missense mutations, we collected most deleterious 13 mutations based on the very high
695 cut-off as CADD score >25 and Grantham scores ≥151, which stands for radical or highly deleterious,
696 while other lower grouping of the Grantham scores conservative (0–50), moderately conservative
697 (51–100) and moderately radical (101–150). We constructed homology models of human HSP47 and
698 mutant HPS47 harboring these 13 missense mutations using canine HSP47 structure (PDB Id - 3ZHA)
699 with SWISS-MODEL [17]. We superimposed and further visualized these two resulting homology
700 models using YASARA [18].

701

702 ### 4.6. Evaluation of HSP47 expression in different cancers tissues and normal tissues

703 We performed an evaluation of expression patterns of HSP47 in different cancer types using dbDEPC
704 3.0, the database of differential expression of protein in cancer [19]. We also examined HSP47
705 expression in normal human tissues using three resources namely, human protein atlas (HPA,
706 https://www.proteinatlas.org/), genotype-tissue expression (GTEx https://gtexportal.org) and
707 FANTOM5 project (http://fantom.gsc.riken.jp/5/).

4.7. Detection of protein interaction partners of HSP47

We detected putative protein interaction partners of human HSP47 using STRING 10 (website: http://version10a.string-db.org [20]) with confidence score higher than 0.9 with searching options of top 20 interaction partners

4.8. Identifications of protein Pfam domains

We identified Pfam protein domains from proteins interacting with HSP47 using HMMER3 with Pfam 32.0 (Sept 2018) dataset.

Supplementary Materials: Supplementary materials can be found online. Supplementary materials contain five supplementary tables as S1-S5.

Author Contributions: Conceptualization, A.K.; Formal analysis, A.K., A.P.; Investigation, A.K., A.P.; Data curation, A.K., A.P., R.K., R.T., S.K., C.G.; Writing—original draft preparation, A.K., A.P., R.K., ; Writing—review and editing A.K., A.P., C.G.; Visualization, A.K., A.P.; Supervision, A.K., C.G.; Project administration, A.K.

Funding: This research received no external funding

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

1000G	1000 Genomes
CADD	Combined annotation dependent depletion
CAGE	Cap analysis of gene expression
CanVar	Cancer variation resource
cBioPortal	cBio cancer genomics portal
CREBBP	CREB binding protein
CRTAP	Cartilage associated protein;
dbDEPC3.0	Database of differentially expressed proteins in human cancers, version 3.0
DNAJB1	DnaJ (Hsp40) homolog subfamily B member 1
DNAJB6	DnaJ (Hsp40) homolog subfamily B member 6
EP300	E1A binding protein p300
FANTOM5	Functional annotation of the mammalian genome project version 5
FKBP	FK506-binding protein 4
gnomAD	Genome aggregation database
GTEx	Genotype-tissue expression
HPA	Human protein atlas
HSBP1	Heat shock factor binding protein 1
HSP47	Heat shock protein 47 kDA
HSP90AA1	Heat shock protein Hsp 90-alph(cytosolic), class A member 1
HSP90AB1	Heat shock protein Hsp 90-alpha (cytosolic), class B member 1
HSPA6	Heat shock 70kDa protein 6
HSPA8	Heat shock 70kDa protein 8;
HSPH1	Heat shock 105kDa/110kDa protein 1
HSPB1	Heat shock protein beta-1
HSPB2	Heat shock protein beta-2
LEPTR1	Leucine proline-enriched proteoglycan (leprecan) 1
KLF13	Kruppel-like factor 13
OMIM	Online Mendelian Inheritance in Man
P4HB	Prolyl 4-hydroxylase beta subunit
PIPB	Peptidyl-prolyl isomerase B

 doi:10.20944/preprints201811.0280.v1

TPM	Transcripts per million
UBB	Ubiquitin B
UBC	Ubiquitin C

725

References

1. Kumar, A.; Bhandari, A.; Sarde, S.J.; Goswami, C. Ancestry & molecular evolutionary analyses of heat shock protein 47 kDa (HSP47/SERPINH1). *Sci Rep* **2017**, *7*, 10394, doi:10.1038/s41598-017-10740-0.
2. Ito, S.; Nagata, K. Biology of Hsp47 (Serpin H1), a collagen-specific molecular chaperone. *Seminars in cell & developmental biology* **2016**, 10.1016/j.semcdb.2016.11.005, doi:10.1016/j.semcdb.2016.11.005.
3. Kumar, A. Bayesian phylogeny analysis of vertebrate serpins illustrates evolutionary conservation of the intron and indels based six groups classification system from lampreys for ~500 MY. *PEERJ* **2015**, 10.7717/peerj.1026, 1-2, doi:10.7717/peerj.1026.
4. Silverman, G.A.; Bird, P.I.; Carrell, R.W.; Church, F.C.; Coughlin, P.B.; Gettins, P.G.; Irving, J.A.; Lomas, D.A.; Luke, C.J.; Moyer, R.W., et al. The serpins are an expanding superfamily of structurally similar but functionally diverse proteins. Evolution, mechanism of inhibition, novel functions, and a revised nomenclature. *J Biol Chem* **2001**, *276*, 33293-33296.
5. Marshall, C.; Lopez, J.; Crookes, L.; Pollitt, R.C.; Balasubramanian, M. A novel homozygous variant in SERPINH1 associated with a severe, lethal presentation of osteogenesis imperfecta with hydranencephaly. *Gene* **2016**, *595*, 49-52, doi:10.1016/j.gene.2016.09.035.
6. Hattori, T.; von der Mark, K.; Kawaki, H.; Yutani, Y.; Kubota, S.; Nakanishi, T.; Eberspaecher, H.; de Crombrugghe, B.; Takigawa, M. Downregulation of rheumatoid arthritis-related antigen RA-A47 (HSP47/colligin-2) in chondrocytic cell lines induces apoptosis and cell-surface expression of RA-A47 in association with CD9. *J Cell Physiol* **2005**, *202*, 191-204, doi:10.1002/jcp.20112.
7. Zhu, J.; Xiong, G.; Fu, H.; Evers, B.M.; Zhou, B.P.; Xu, R. Chaperone Hsp47 Drives Malignant Growth and Invasion by Modulating an ECM Gene Network. *Cancer Res* **2015**, *75*, 1580-1591, doi:10.1158/0008-5472.CAN-14-1027.
8. Christiansen, H.E.; Schwarze, U.; Pyott, S.M.; AlSwaid, A.; Al Balwi, M.; Alrasheed, S.; Pepin, M.G.; Weis, M.A.; Eyre, D.R.; Byers, P.H. Homozygosity for a missense mutation in SERPINH1, which encodes the collagen chaperone protein HSP47, results in severe recessive osteogenesis imperfecta. *Am J Hum Genet* **2010**, *86*, 389-398, doi:10.1016/j.ajhg.2010.01.034.
9. Abecasis, G.R.; Auton, A.; Brooks, L.D.; DePristo, M.A.; Durbin, R.M.; Handsaker, R.E.; Kang, H.M.; Marth, G.T.; McVean, G.A. An integrated map of genetic variation from 1,092 human genomes. *Nature* **2012**, *491*, 56-65, doi:10.1038/nature11632.
10. Lek, M.; Karczewski, K.J.; Minikel, E.V.; Samocha, K.E.; Banks, E.; Fennell, T.; O'Donnell-Luria, A.H.; Ware, J.S.; Hill, A.J.; Cummings, B.B., et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **2016**, *536*, 285-291, doi:10.1038/nature19057.
11. Forbes, S.A.; Bhamra, G.; Bamford, S.; Dawson, E.; Kok, C.; Clements, J.; Menzies, A.; Teague, J.W.; Futreal, P.A.; Stratton, M.R. The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics* **2008**, *Chapter 10*, Unit 10 11, doi:10.1002/0471142905.hg1011s57.
12. Gao, J.; Aksoy, B.A.; Dogrusoz, U.; Dresdner, G.; Gross, B.; Sumer, S.O.; Sun, Y.; Jacobsen, A.; Sinha, R.; Larsson, E., et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* **2013**, *6*, pl1, doi:10.1126/scisignal.2004088.
13. Chubb, D.; Broderick, P.; Dobbins, S.E.; Houlston, R.S. CanVar: A resource for sharing germline variation in cancer patients. *F1000Res* **2016**, *5*, 2813, doi:10.12688/f1000research.10058.1.
14. Ng, P.C.; Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* **2003**, *31*, 3812-3814.

15. Adzhubei, I.A.; Schmidt, S.; Peshkin, L.; Ramensky, V.E.; Gerasimova, A.; Bork, P.; Kondrashov, A.S.; Sunyaev, S.R. A method and server for predicting damaging missense mutations. *Nature methods* **2010**, *7*, 248-249, doi:10.1038/nmeth0410-248.

16. Kircher, M.; Witten, D.M.; Jain, P.; O'Roak, B.J.; Cooper, G.M. A general framework for estimating the relative pathogenicity of human genetic variants. **2014**, *46*, 310-315, doi:10.1038/ng.2892.

17. Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F.T.; de Beer, T.A.P.; Rempfer, C.; Bordoli, L., et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res* **2018**, *46*, W296-W303, doi:10.1093/nar/gky427.

18. Land, H.; Humble, M.S. YASARA: A Tool to Obtain Structural Guidance in Biocatalytic Investigations. *Methods Mol Biol* **2018**, *1685*, 43-67, doi:10.1007/978-1-4939-7366-8_4.

19. Yang, Q.; Zhang, Y.; Cui, H.; Chen, L.; Zhao, Y.; Lin, Y.; Zhang, M.; Xie, L. dbDEPC 3.0: the database of differentially expressed proteins in human cancer with multi-level annotation and drug indication. *Database (Oxford)* **2018**, *2018*, doi:10.1093/database/bay015.

20. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P., et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **2015**, *43*, D447-452, doi:10.1093/nar/gku1003.

21. Ogryzko, V.V.; Schiltz, R.L.; Russanova, V.; Howard, B.H.; Nakatani, Y. The transcriptional coactivators p300 and CBP are histone acetyltransferases. *Cell* **1996**, *87*, 953-959.

22. Milani, D.; Manzoni, F.M.; Pezzani, L.; Ajmone, P.; Gervasini, C.; Menni, F.; Esposito, S. Rubinstein-Taybi syndrome: clinical features, genetic basis, diagnosis, and management. *Ital J Pediatr* **2015**, *41*, 4, doi:10.1186/s13052-015-0110-1.

23. Chang, W.; Barnes, A.M.; Cabral, W.A.; Bodurtha, J.N.; Marini, J.C. Prolyl 3-hydroxylase 1 and CRTAP are mutually stabilizing in the endoplasmic reticulum collagen prolyl 3-hydroxylation complex. *Hum Mol Genet* **2010**, *19*, 223-234, doi:10.1093/hmg/ddp481.

24. Kim, J.H.; Alderson, T.R.; Frederick, R.O.; Markley, J.L. Nucleotide-dependent interactions within a specialized Hsp70/Hsp40 complex involved in Fe-S cluster biogenesis. *Journal of the American Chemical Society* **2014**, *136*, 11586-11589, doi:10.1021/ja5055252.

25. Qiu, X.B.; Shao, Y.M.; Miao, S.; Wang, L. The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. *Cell Mol Life Sci* **2006**, *63*, 2560-2570, doi:10.1007/s00018-006-6192-6.

26. Mansson, C.; Arosio, P.; Hussein, R.; Kampinga, H.H.; Hashem, R.M.; Boelens, W.C.; Dobson, C.M.; Knowles, T.P.; Linse, S.; Emanuelsson, C. Interaction of the molecular chaperone DNAJB6 with growing amyloid-beta 42 (Abeta42) aggregates leads to sub-stoichiometric inhibition of amyloid formation. *The Journal of biological chemistry* **2014**, *289*, 31066-31076, doi:10.1074/jbc.M114.595124.

27. Aprile, F.A.; Kallstig, E.; Limorenko, G.; Vendruscolo, M.; Ron, D.; Hansen, C. The molecular chaperones DNAJB6 and Hsp70 cooperate to suppress alpha-synuclein aggregation. *Sci Rep* **2017**, *7*, 9039, doi:10.1038/s41598-017-08324-z.

28. Assimon, V.A.; Southworth, D.R.; Gestwicki, J.E. Specific Binding of Tetratricopeptide Repeat Proteins to Heat Shock Protein 70 (Hsp70) and Heat Shock Protein 90 (Hsp90) Is Regulated by Affinity and Phosphorylation. *Biochemistry* **2015**, *54*, 7120-7131, doi:10.1021/acs.biochem.5b00801.

29. Kampinga, H.H.; Craig, E.A. The HSP70 chaperone machinery: J proteins as drivers of functional specificity. *Nat Rev Mol Cell Biol* **2010**, *11*, 579-592, doi:10.1038/nrm2941.

30. Prabhu, S.; Raman, B.; Ramakrishna, T.; Rao Ch, M. HspB2/myotonic dystrophy protein kinase binding protein (MKBP) as a novel molecular chaperone: structural and functional aspects. *PLoS One* **2012**, *7*, e29810, doi:10.1371/journal.pone.0029810.

31. Mogk, A.; Schlieker, C.; Friedrich, K.L.; Schonfeld, H.J.; Vierling, E.; Bukau, B. Refolding of substrates bound to small Hsps relies on a disaggregation reaction mediated most efficiently by ClpB/DnaK. *The Journal of biological chemistry* **2003**, *278*, 31033-31042, doi:10.1074/jbc.M303587200.

32. Yang, S.; Dizhoor, A.; Wilson, D.J.; Adamus, G. GCAP1, Rab6, and HSP27: Novel Autoantibody Targets in Cancer-Associated Retinopathy and Autoimmune Retinopathy. *Translational vision science & technology* **2016**, *5*, 1, doi:10.1167/tvst.5.3.1.

33. Yamagishi, N.; Ishihara, K.; Saito, Y.; Hatayama, T. Hsp105 but not Hsp70 family proteins suppress the aggregation of heat-denatured protein in the presence of ADP. *FEBS letters* **2003**, *555*, 390-396.

34. Mollapour, M.; Neckers, L. Post-translational modifications of Hsp90 and their contributions to chaperone regulation. *Biochimica et biophysica acta* **2012**, *1823*, 648-655, doi:10.1016/j.bbamcr.2011.07.018.

35. Picard, D. Heat-shock protein 90, a chaperone for folding and regulation. *Cellular and molecular life sciences : CMLS* **2002**, *59*, 1640-1648.

36. Schulz, R.; Streller, F.; Scheel, A.H.; Ruschoff, J.; Reinert, M.C.; Dobbelstein, M.; Marchenko, N.D.; Moll, U.M. HER2/ErbB2 activates HSF1 and thereby controls HSP90 clients including MIF in HER2-overexpressing breast cancer. *Cell death & disease* **2014**, *5*, e980, doi:10.1038/cddis.2013.508.

37. Bieker, J.J. Kruppel-like factors: three fingers in many pies. *J Biol Chem* **2001**, *276*, 34355-34358, doi:10.1074/jbc.R100043200.

38. Rauch, F.; Fahiminiya, S.; Majewski, J.; Carrot-Zhang, J.; Boudko, S.; Glorieux, F.; Mort, J.S.; Bachinger, H.P.; Moffatt, P. Cole-Carpenter syndrome is caused by a heterozygous missense mutation in P4HB. *Am J Hum Genet* **2015**, *96*, 425-431, doi:10.1016/j.ajhg.2014.12.027.

39. Kim, J.; Choi, T.G.; Ding, Y.; Kim, Y.; Ha, K.S.; Lee, K.H.; Kang, I.; Ha, J.; Kaufman, R.J.; Lee, J., et al. Overexpressed cyclophilin B suppresses apoptosis associated with ROS and Ca2+ homeostasis after ER stress. *Journal of cell science* **2008**, *121*, 3636-3648, doi:10.1242/jcs.028654.

40. Cabral, W.A.; Perdivara, I.; Weis, M.; Terajima, M.; Blissett, A.R.; Chang, W.; Perosky, J.E.; Makareeva, E.N.; Mertz, E.L.; Leikin, S., et al. Abnormal type I collagen post-translational modification and crosslinking in a cyclophilin B KO mouse model of recessive osteogenesis imperfecta. *PLoS genetics* **2014**, *10*, e1004465, doi:10.1371/journal.pgen.1004465.

41. Swatek, K.N.; Komander, D. Ubiquitin modifications. *Cell Res* **2016**, *26*, 399-422, doi:10.1038/cr.2016.39.

42. Fischer, D.F.; De Vos, R.A.; Van Dijk, R.; De Vrij, F.M.; Proper, E.A.; Sonnemans, M.A.; Verhage, M.C.; Sluijs, J.A.; Hobo, B.; Zouambia, M., et al. Disease-specific accumulation of mutant ubiquitin as a marker for proteasomal dysfunction in the brain. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* **2003**, *17*, 2014-2024, doi:10.1096/fj.03-0205com.

43. Shendure, J.; Akey, J.M. The origins, determinants, and consequences of human mutations. *Science* **2015**, *349*, 1478-1483, doi:10.1126/science.aaa9119.

44. Kumar, A.; Bhandari, A.; Goswami, C. Surveying genetic variants and molecular phylogeny of cerebral cavernous malformation gene, CCM3/PDCD10. *Biochemical and Biophysical Research Communications* **2014**, *455*, 98-106, doi:10.1016/j.bbrc.2014.10.105.

45. Kumar, A.; Bhandari, A.; Sarde, S.J.; Goswami, C. Sequence, phylogenetic and variant analyses of antithrombin III. *Biochemical and Biophysical Research Communications* **2013**, *440*, 714-724, doi:10.1016/j.bbrc.2013.09.134.

853 46. Kumar, A.; Bhandari, A.; Sarde, S.J.; Goswami, C. Genetic variants and evolutionary analyses of heparin
854 cofactor II. *Immunobiology* **2014**, *219*, 713-728, doi:10.1016/j.imbio.2014.05.003.
855 47. Kumar, A.; Sarde, S.J.; Bhandari, A. Revising angiotensinogen from phylogenetic and genetic variants
856 perspectives. *Biochemical and biophysical research communications* **2014**, *446*, 504-518, doi:10.1016/j.bbrc.2014.02.139.
857 48. Lee, H.W.; Kwon, J.; Kang, M.C.; Noh, M.K.; Koh, J.S.; Kim, J.H.; Park, J.H. Overexpression of HSP47 in
858 esophageal squamous cell carcinoma: clinical implications and functional analysis. *Dis Esophagus* **2015**,
859 10.1111/dote.12359, doi:10.1111/dote.12359.
860 49. Palotai, R.; Szalay, M.S.; Csermely, P. Chaperones as integrators of cellular networks: changes of cellular
861 integrity in stress and diseases. *IUBMB Life* **2008**, *60*, 10-18, doi:10.1002/iub.8.
862 50. Kumar, A.; Bandapalli, O.R.; Paramasivam, N.; Giangiobbe, S.; Diquigiovanni, C.; Bonora, E.; Eils, R.;
863 Schlesner, M.; Hemminki, K.; Forsti, A. Familial Cancer Variant Prioritization Pipeline version 2 (FCVPPv2)
864 applied to a papillary thyroid cancer family. *Sci Rep* **2018**, *8*, 11635, doi:10.1038/s41598-018-29952-z.

Tables

Table 1. Summary of missense variants of human HSP47 derived from 1000Genomes dataset [9].

doi:10.20944/preprints201811.0280.v1

Mutation*	Structural location	SIFT	SIFT score	PolyPhenV2	PolyPhenV2 score	Variant ID	PosChr11	Alleles	gMAF	NtgMAF	Exon	Variant Class	Source	Evidences
Leu6Pro	N-terminal extension	DEL	0,03	UNK	0	rs11539325	75277411	T/C	-	-	eI	SNP	dbS	-
Glu20Lys	N-terminal extension	TOL	0,77	UNK	0	rs200397594	75277452	G/A	-	-	eI	SNP	dbS	-
Gly31Ala	N-terminal extension	TOL	1	UNK	0	rs140588417	75277486	G/C	-	-	eI	SNP	dbS	Mo,ESP
Ala33Pro	N-terminal extension	TOL	0,18	UNK	0	rs150061926	75277491	G/C	0	C	eI	SNP	dbS	Mo,ESP
Ala41Pro	N-terminal extension	TOL	0,23	BEN	0,133	rs7105528	75277515	G/C	-	-	eI	SNP	dbS	Mo,
Gln56Arg	Helix hA	TOL	0,66	BEN	0,01	ss1341923074	75277561	A/G	0	G	eI	SNP	1KGp3	-
Ala63Val	Loop BTN helix hA-sheet s6B	TOL	0,28	BEN	0,15	ss1341923075	75277582	C/T	0,001	T	eI	SNP	1KGp3	-
Ser76Leu	Helix hB	DEL	0	PRD	0,998	rs376824871	75277621	C/T	-	-	eI	SNP	dbS	ESP
Leu78Pro	Helix hB	DEL	0	PRD	1	rs137853892	75277627	T/C	-	-	eI	SNP	dbS	-
Ala90Thr	Helix hC	DEL	0	POD	0,887	COSM931998	75277662	G/A	-	-	eI	Ss	Cos	-
Ser91Leu	Helix hC	DEL	0,01	POD	0,586	COSM1298619	75277666	C/T	-	-	eI	Ss	Cos	-
Gln92Glu	*Helix hC*	*TOL*	*0,06*	*POD*	*0,884*	*rs112365393*	*75277668*	*C/G*	-	-	*eI*	*SNP*	*dbS*	-
Ser98Asn	Loop BTN helices hC-hD	TOL	0,53	BEN	0,033	ss1341923079	75277687	G/A	0,001	A	eI	SNP	1KGp3	-
Glu100Lys	Loop BTN helices hC-hD	TOL	0,91	BEN	0,005	COSM931999	75277692	G/A	-	-	eI	Ss	Cos	-
Leu102Gln	Loop BTN helices hC-hD	TOL	0,25	PRD	0,991	rs377583721	75277699	T/A	-	-	eI	SNP	dbS	ESP
Arg103Cys	Loop BTN helices hC-hD	TOL	0,06	POD	0,494	COSM323275	75277702	G/A	-	-	eI	Ss	Cos	-
Arg103His	Loop BTN helices hC-hD	DEL	0,02	POD	0,802	COSM323275	75277702	G/A	-	-	eI	Ss	dbS	ESP

doi:10.20944/preprints201811.0280.v1

Glu105Gln	Loop BTN helices hC-hD	TOL	0,08	PRD	0,918	rs376810317	75277707	G/C	-	-	eI	SNP	dbS	-
Glu106Asp	Loop BTN helices hC-hD	TOL	0,88	BEN	0,002	ss1341923080	75277712	G/C	0	C	eI	SNP	1KGp3	-
Ala109Gly	Helix hD	TOL	0,2	BEN	0,01	rs148959638	75277720	C/G	-	-	eI	SNP	dbS	Fre,ESP
Leu111Met	Helix hD	TOL	0,05	PRD	0,956	ss1341923081	75277725	C/A	0	A	eI	SNP	1KGp3	-
Arg116His	Helix hD	TOL	0,06	BEN	0,286	COSM1240157	75277741	G/A	-	-	eI	Ss	Cos	dbS,Mo,ESP
Arg116His	Helix hD	TOL	0,06	BEN	0,286	rs200265134	75277740	C/A	0	A	eI	SNP	Cos	-
Ser117Leu	Helix hD	DEL	0,02	BEN	0,179	ss1341923083	75277744	C/T	0	T	eI	SNP	1KGp3	-
Val126Met	Loop BTN helix hD-sheet s2A	TOL	0,08	PRD	0,936	COSM932000	75277770	G/A	-	-	eI	Ss	Cos	-
Tyr135His	Sheet s2A	DEL	0	PRD	0,999	rs374115511	75277797	T/C	-	-	eI	SNP	dbS	-
Val140Leu	Loop BTN sheet s2A-helix hE	TOL	0,46	BEN	0,435	rs148088085	75277812	G/C	-	-	eI	SNP	dbS	Fre,ESP
Arg148Ser	Helix hE	TOL	0,47	BEN	0,063	rs61736330	75277836	C/A	-	-	eI	SNP	dbS	Fre,
Ile161Leu	Sheet s1A	TOL	0,24	POD	0,741	rs112083274	75277875	A/C	0	C	eI	SNP	dbS	Mo,ESP
Gly183Ser	Loop BTN helix hF-sheet s3A	DEL	0,01	PRD	1	COSM1638975	75277941	G/A	-	-	eI	Ss	Cos	-
Thr189Ala	Sheet s3A	TOL	0,49	BEN	0,06	rs138784081	75277959	A/G	0,001	G	eI	SNP	dbS	Mo,ESP
Asp191His	Sheet s3A	TOL	0,1	POD	0,884	COSM545258	75277965	G/C	-	-	eI	Ss	Cos	-
Glu193Asp	Sheet s3A	TOL	0,58	BEN	0,074	ss1341923089	75277973	G/C	0	C	eI	SNP	1KGp3	-
Arg194Ser	Sheet s3A	TOL	0,67	BEN	0,103	COSM1235751	75277974	C/A	-	-	eI	Ss	Cos	-
Arg194Ser	Sheet s3A	TOL	0,67	BEN	0,103	rs141721173	75277974	C/A	0	A	eI	SNP	dbS	Mo,ESP

doi:10.20944/preprints201811.0280.v1

His209Gln	Loop BTN sheet s3A-helix hF1	TOL	0,14	POD	0,619	ss1341923146	75279780	C/G	0	G	eII	SNP	1KGp3	-
His215Tyr	Loop BTN sheet s3A-helix hF1	DEL	0,03	POD	0,906	rs373526486	75279796	C/T	-	-	eII	SNP	dbS	-
His216Arg	Helix hF1	TOL	0,06	BEN	0,011	COSM131116	75279800	A/G	-	-	eII	Ss	Cos	-
Glu249Asp	Sheet s1B	TOL	0,55	POD	0,551	COSM1357002	75280007	G/A	-	-	eIII	Ss	1KGp3	-
Glu249Lys	Sheet s1B	TOL	0,86	BEN	0,05	COSM1357002	75280007	G/A	-	-	eIII	Ss	Cos	-
Glu251Lys	Sheet s1B	TOL	0,73	BEN	0,266	ss1341923157	75280013	G/A	0	A	eIII	SNP	1KGp3	-
Leu260Met	Sheet s2B	DEL	0,03	PRD	1	COSM1638976	75280040	C/A	-	-	eIII	Ss	Cos	-
Ala261Gly	Sheet s2B	TOL	0,44	BEN	0,007	rs201644679	75280044	C/G	0	G	eIII	SNP	dbS	-
Ser266Asn	Sheet s3B	DEL	0	PRD	0,912	COSM1225309	75280059	G/A	-	-	eIII	Ss	Cos	-
His273Leu	Loop BTN sheet s3B-helix hG	TOL	0,63	BEN	0,033	rs267603192	75280080	A/T	-	-	eIII	SNP	dbS	-
Val275Met	Loop BTN sheet s3B-helix hG	TOL	0,32	BEN	0,122	rs199679249	75280085	G/A	-	-	eIII	SNP	dbS	Fre,
Glu276Ala	Loop BTN sheet s3B-helix hG	DEL	0,01	POD	0,686	rs148613550	75280089	A/C	-	-	eIII	SNP	dbS	Fre,
GLu279Lys	Helix hG	TOL	0,19	BEN	0,37	COSM277225	75280097	G/A	-	-	eIII	Ss	Cos	-
Arg280Cys	Helix hG	DEL	0,03	PRD	0,977	rs200572997	75280100	C/T	0	T	eIII	SNP	dbS	ESP
Arg280His	Helix hG	DEL	0	PRD	0,977	rs370057420	75280101	G/A	0	A	eIII	SNP	dbS	-
Glu288Lys	Helix hH	TOL	0,37	BEN	0,012	rs369751579	75280124	G/A	-	-	eIII	SNP	dbS	-
Met297Ile	Loop BTN helix hH-sheet s2C	DEL	0,03	BEN	0,181	rs200106884	75280153	G/T	0	T	eIII	SNP	dbS	-
Ser305Pro	Sheet s2C	TOL	0,09	PRD	0,999	ss1341923165	75280175	T/C	0	C	eIII	SNP	1KGp3	-
Lys308Arg	Loop BTN sheets s2C-s6A	TOL	0,52	PRD	0,999	rs142663000	75280185	A/G	-	-	eIII	SNP	dbS	Fre,ESP
Leu321Pro	Helix hI	DEL	0	PRD	1	COSM932003	75282833	T/C	-	-	eIV	Ss	Cos	-
Gly323Trp	Helix hI	DEL	0,01	PRD	0,937	COSM326447	75282838	G/T	-	-	eIV	Ss	Cos	-
Gly323Glu	Helix hI	TOL	1	BEN	0,008	COSM326447	75282838	G/T	-	-	eIV	Ss	dbS	Mo,ESP

doi:10.20944/preprints201811.0280.v1

Gly323Trp	Helix hI	DEL	0,01	PRD	0,937	rs144791057	75282839	G/A	-	-	eIV	SNP	Cos	-
Lys332Asn	Loop BTN helix hI-sheet s5A	TOL	0,13	PRD	0,999	rs147936395	75282867	G/C	-	-	eIV	SNP	dbS	Fre,ESP
Arg339Leu	Loop BTN helix hI-sheet s5A	TOL	0,06	POD	0,597	COSM196204	75282887	G/A	-	-	eIV	Ss	Cos	-
Met340Ile	Loop BTN helix hI-sheet s5A	TOL	1	BEN	0,001	rs201416803	75282891	G/A	-	-	eIV	SNP	dbS	Fre,
Ser350Thr	Sheet s5A	TOL	0,3	BEN	0,439	rs150431930	75282920	G/C	-	-	eIV	SNP	dbS	Fre,ESP
Val351Met	Sheet s5A	TOL	0,06	POD	0,474	rs368336245	75282922	G/A	-	-	eIV	SNP	dbS	ESP
His353Asn	Sheet s5A	DEL	0	PRD	0,997	COSM1357004	75282928	C/A	-	-	eIV	Ss	Cos	-
Ala354Thr	Sheet s5A	TOL	0,2	PRD	0,999	rs369550626	75282931	G/A	-	-	eIV	SNP	dbS	ESP
Glu358Gln	Sheet s5A	DEL	0	PRD	0,999	COSM1298620	75282943	G/C	-	-	eIV	Ss	Cos	-
Pro365Leu	Sheet s4A, within RCL	TOL	0,06	POD	0,697	ss1341923258	75282965	C/T	0	T	eIV	SNP	1KGp3	-
Gly372Arg	Sheet s4A, within RCL	TOL	0,11	PRD	0,962	rs200180052	75282985	G/A	0	A	eIV	SNP	dbS	-
Arg373Pro	Sheet s4A, within RCL	TOL	0,32	POD	0,645	COSM690662	75282989	G/C	-	-	eIV	Ss	Cos	-
Glu375Val	Sheet s4A, within RCL	TOL	0,3	POD	0,846	ss1341923262	75282995	A/T	0	T	eIV	SNP	1KGp3	-
Arg377Cys	Sheet s4A, within RCL	DEL	0,02	PRD	0,916	ss1341923264	75283000	C/T	0	T	eIV	SNP	1KGp3	-
Pro379Leu	Sheet s4A, within RCL	DEL	0	PRD	1	rs267603193	75283007	C/T	-	-	eIV	SNP	dbS	-
Phe382Leu	Sheet s1C	DEL	0	PRD	0,995	COSM932005	75283017	C/A	-	-	eIV	Ss	Cos	-
Ala384Thr	Loop BTN sheets s1C-s4B	DEL	0,01	POD	0,699	rs200974428	75283021	G/A	-	-	eIV	SNP	dbS	Fre,
Arg393Trp	Sheet s4B	TOL	0,06	BEN	0,004	ss1341923267	75283048	C/T	0	T	eIV	SNP	1KGp3	-
Ser399Phe	Sheet s5B	DEL	0	PRD	0,999	rs376520307	75283067	C/T	-	-	eIV	SNP	dbS	ESP
Leu401Val	Sheet s5B	DEL	0	POD	0,849	ss1341923268	75283072	C/G	0	G	eIV	SNP	1KGp3	-
Ile403Thr	Sheet s5B	TOL	0,15	BEN	0,062	rs201566218	75283079	T/C	-	-	eIV	SNP	dbS	Mo,Fre,ESP
Leu406Val	Sheet s5B	TOL	0,17	BEN	0,087	COSM72604	75283087	C/G	-	-	eIV	Ss	Cos	-

doi:10.20944/preprints201811.0280.v1

Val407Leu	Sheet s5B	TOL	0,23	POD	0,686	rs138241050	75283090	G/C	-	-	eIV	SNP	dbS	Mo,ESP
Arg408Gln	Sheet s5B	TOL	0,19	BEN	0,152	rs371699925	75283094	G/A	-	-	eIV	SNP	dbS	ESP
Asp412Glu	C-terminal end	TOL	0,84	BEN	0,015	rs200334001	75283107	C/G	0	G	eIV	SNP	dbS	-

*bold - deleterious by both SIFT and Polyphen V2; italics - deleterious by Polyphen V2 only; underline - deleterious by SIFT only

1KGp3 - 1KG_phase3; BEN - benign; BTN - between; Cos - COSMIC database; dbS -dbSNP; DEL - deleterious, SIFT score <0.06; Fre - Frequency; gMAF - global MAF; Mo - Multiple_observations; NtgMAT - Nucleotide for global MAF is given; POD - possibly damaging, Polyphen V2 score > 0.45; PosChr11 - Position on the Chr. 11; PRD - probably damaging, Polyphen V2 score > 0.90; sS - somatic_SNV; TOL – tolerated; UNK - Unknown

doi:10.20944/preprints201811.0280.v1

Table 2. Overview of HSP47 missense variants from cBioPortal.

Mutation	Structural Elements	Cancer types (#samples)	SIFT status*	Polyphen V2 status**
Arg2His	N-terminal end	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor (1)	TLC	B
Ala9Thr	N-terminal end	Uterine Endometrioid Carcinoma (1)	T	B
Ala19Thr	N-terminal end	Pancreatic Adenocarcinoma (1); Prostate Adenocarcinoma (1)	T	U
Lys22Asn	N-terminal end	Breast Invasive Ductal Carcinoma (1)	DLC	POD
Ala40Val	N-terminal end	Colorectal Adenocarcinoma (1)	D	PRD
Thr42Met	N-terminal end	Diffuse Large B-Cell Lymphoma, NOS (1)	T	B
Ala44Thr	Helix hA	Esophageal Squamous Cell Carcinoma (3)	D	PRD
Ala59Gly	Helix hA	Leiomyosarcoma (1)	T	POD
Ser76Leu	Helix hB	Cutaneous Melanoma (1); Uterine Endometrioid Carcinoma (1)	D	POD
Ser77Leu	Helix hB	Diffuse Large B-Cell Lymphoma, NOS (2)	D	POD
Gly79Val	Helix hB	Cutaneous Melanoma (1)	D	PRD
Leu83Val	Helix hB	Lung Adenocarcinoma (1)	D	B
Gly85Asp	Loop between hB-hC	Uterine Endometrioid Carcinoma (1)	D	PRD
Ala90Thr	Helix hC	Uterine Endometrioid Carcinoma (3)	D	POD
Ser91Leu	Helix hC	Bladder Urothelial Carcinoma (4)	D	POD
Ala99Thr	Loop B between hC-hD	Colorectal Adenocarcinoma (1)	T	POD
Glu100Lys	Loop between hC-hD	Uterine Endometrioid Carcinoma (3)	T	B

doi:10.20944/preprints201811.0280.v1

Arg103Cys	Loop between hC-hD	Stomach Adenocarcinoma (2)	D	POD
Arg103His	Loop between hC-hD	Small Cell Lung Cancer (1)	T	POD
Glu106Lys	Loop between hC-hD	Cutaneous Melanoma (1)	T	B
Arg116Ser	Helix hD	Germinal Center B-Cell Type (1)	T	B
Arg116His	Helix hD	Esophageal Adenocarcinoma (1); Stomach Adenocarcinoma (2)	T	B
Val126Met	Loop between hD-s2A	Uterine Endometrioid Carcinoma (4)	T	POD
Arg133Gln	Sheet s2A	Uterine Endometrioid Carcinoma (1)	D	PRD
Ser139Pro	Loop between s2A-hE	Hepatocellular Carcinoma (1)	D	PRD
Ser141Ile	Loop between s2A-hE	Leiomyosarcoma (1)	D	B
Val147Glu	Helix hE	Tubular Stomach Adenocarcinoma (1)	D	PRD
Arg148His	Helix hE	Colon Adenocarcinoma (1)	D	POD
Ser150Arg	Helix hE	Hepatocellular Adenoma (1)	D	PRD
His153Tyr	Helix hE	Mucinous Adenocarcinoma of the Colon and Rectum (1)	D	PRD
Ser159Cys	Sheet s1A	Colorectal Adenocarcinoma (1)	D	PRD
Asp165Asn	Helix hF	Colorectal Adenocarcinoma (2); Glioblastoma Multiforme (1)	T	PRD
Arg167His	Helix hF	Colorectal Adenocarcinoma (1)	D	PRD
Gln171Leu	Helix hF	Lung Squamous Cell Carcinoma (2)	T	B
Ileu173Val	Helix hF	Esophagogastric Adenocarcinoma (1)	D	POD
Glu175Lys	Helix hF	Diffuse Large B-Cell Lymphoma (2); Rectal Adenocarcinoma (1)	T	POD
Arg178Val	Helix hF	Multiple Myeloma (1); Bladder Urothelial Carcinoma (2); Cervical Squamous Cell Carcinoma (1); Ampullary Carcinoma (1)	T	POD
Asp182Asn	Helix hF	Colorectal Adenocarcinoma (1)	T	B
Gly183Ser	Loop between hF-s3A	Stomach Adenocarcinoma (1)	D	PRD

doi:10.20944/preprints201811.0280.v1

Lys184Gln	Loop between hF-s3A	Breast Invasive Ductal Carcinoma (1)	D	PRD
Glu187Lys	Loop between hF-s3A	Breast Invasive Ductal Carcinoma (1)	D	POD
Asp191His	Sheet s3A	Lung Adenocarcinoma (2)	D	POD
Arg194His	Sheet s3A	Colorectal Adenocarcinoma (1)	T	POD
Thr195Met	Sheet s3A	Diffuse Glioma (1); Anaplastic Oligoastrocytoma (1); Oligodendroglioma (1)	T	B
Arg198Thr	Sheet s3A	Papillary Thyroid Cancer (3)	T	PRD
Leu199Met	Sheet s3A	Cutaneous Melanoma (1)	T	B
Arg203Thr	Sheet s3A	B-Lymphoblastic Leukemia/Lymphoma (1)	D	B
Met204Val	Sheet s3A	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma (1)	D	B
Lys217Asn	Helix hF1	Thymoma (1)	T	B
Arg228Trp	Loop between s4C-s3C	Cutaneous Squamous Cell Carcinoma (1)	D	PRD
Arg228Leu	Loop between s4C-s3C	Cutaneous Melanoma (1)	T	PRD
Ser229Pro	Loop between s4C-s3C	Hepatocellular Carcinoma (1)	D	PRD
Ser229Phe	Loop between s4C-s3C	Cutaneous Melanoma (2)	T	B
Thr231Ile	Sheet s3C	Cutaneous Melanoma	D	PRD
Met235Ile	Sheet s3C	Bladder Urothelial Carcinoma (1)	T	B
Arg239Gln	Sheet s3C	Diffuse Large B-Cell Lymphoma (1); Colon Adenocarcinoma (1)	T	B
Gly241Asp	Sheet s3C	Uterine Serous Carcinoma/Uterine Papillary Serous Carcinoma (1)	T	PRD
Glu249Lys	Sheet s1B	Colon Adenocarcinoma (1)	T	B
Val256Met	Sheet s2B	Breast Invasive Ductal Carcinoma (1)	D	POD
Leu260Met	Sheet s2B	Stomach Adenocarcinoma (1)	D	PRD
Ser266Asn	Sheet s3B	Colorectal Adenocarcinoma (1)	D	PRD

doi:10.20944/preprints201811.0280.v1

Glu279Lys	Helix hG	Colorectal Adenocarcinoma (1); Mucinous Adenocarcinoma of the Colon and Rectum (2)	T	B
Arg280Cys	Helix hG	Head and Neck Squamous Cell Carcinoma (3); Ampullary Carcinoma (1)	D	PRD
Leu281Ile	Helix hG	Uterine Mixed Endometrial Carcinoma (1)	T	POD
Lys283Asn	Helix hG	Hepatocellular Adenoma (1)	D	B
Lys291Asn	Helix hH	Uterine Endometrioid Carcinoma (1)	T	B
Arg303Thr	Sheet s2C	Invasive Breast Carcinoma (1); Breast Invasive Ductal Carcinoma (1); Uterine Endometrioid Carcinoma (1)	T	B
Lys308Glu	Loop between s2C-s6A	Hepatocellular Carcinoma (2)	D	PRD
Leu321Pro	Helix hI	Uterine Endometrioid Carcinoma (3)	D	PRD
Gly323Trp	Helix hI	Small Cell Lung Cancer (1)	D	POD
Leu324Met	Helix hI	Bladder Urothelial Carcinoma (2)	T	POD
Leu326Arg	Loop between hI-hI1	Uterine Carcinosarcoma/Uterine Malignant Mixed Mullerian Tumor (2)	D	PRD
Arg329Thr	Helix hl1	Uterine Endometrioid Carcinoma (1)	D	PRD
Ile330Val	Loop between hI1-s5A	Cutaneous Melanoma (2)	T	B
Arg339His	Loop between hI1-s5A	Stomach Adenocarcinoma (2); Colorectal Adenocarcinoma (1); Mucinous Adenocarcinoma of the Colon and Rectum (1)	T	POD
Lys343Thr	Loop between hI1-s5A	Cutaneous Melanoma (2)	D	PRD
Val351Met	Sheet s5A	Hepatocellular Carcinoma (1)	T	POD
His353Asn	Sheet s5A	Colon Adenocarcinoma (1)	D	PRD
Arg356Thr	Sheet s5A	Stomach Adenocarcinoma (6); Intestinal Type Stomach Adenocarcinoma (1); Uterine Endometrioid Carcinoma (1)	T	B
Glu358Gln	Sheet s5A	Bladder Urothelial Carcinoma (4)	D	PRD
Asp360Gly	Sheet s4A	Uterine Endometrioid Carcinoma (1)	T	POD

doi:10.20944/preprints201811.0280.v1

Phe366Leu	Sheet s4A, within RCL	Rectal Adenocarcinoma (1)	T	B
Arg373Cys	Sheet s4A, within RCL	Pancreatic Adenocarcinoma (1)	T	POD
Arg373Pro	Sheet s4A, within RCL	Lung Squamous Cell Carcinoma (4)	T	B
Arg377Cys	Sheet s4A, within RCL	Colorectal Adenocarcinoma (1); B-Lymphoblastic Leukemia/Lymphoma (1)	D	PRD
Phe382Leu	Sheet s1C	Bladder Urothelial Carcinoma (2); Uterine Endometrioid Carcinoma (3)	D	PRD
Pro387Ser	Loop between s1C-s4B	Cutaneous Melanoma (1)	D	PRD
Thr395Ile	Sheet s4B	Uterine Endometrioid Carcinoma (1)	D	B
Ser399Phe	Sheet s5B	Cutaneous Squamous Cell Carcinoma (1); Cutaneous Melanoma (1)	D	PRD
Leu406Val	Sheet s5B	High-Grade Serous Ovarian Cancer (1); Serous Ovarian Cancer (1)	T	B
Gly411Cys	C-terminal end	Cutaneous Melanoma (1)	D	PRD

BTN – Between; *SIFT status types: D - deleterious;DLC - deleterious_low_confidence;T- tolerated;TLC - tolerated_low_confidence. **Polyphen-2 status types; B - benign;POD - possibly_damaging;PRD - probably_damaging;U - unknown

doi:10.20944/preprints201811.0280.v1

Table 3: Germline missense variants of HSP47 found in colorectal cancer patients, derived from CanVar database

Missense Mutation	Structural Location	Transcript Change	Chrom	Position	RSID	Reference	Alternate	Allele Count	Allele Number	EXAC_AF	Allele Frequency	CADD score
Ile161Leu	Sheet s1A	c.481A>C	11	75277875	rs112083274	A	C	1	1216	9.5e-05	8.224	23.4
Gly183Arg	Loop between hF-s3A	c.547G>C	11	75277941	.	G	C	1	1654	1.92e-05	6.046	31

4

5 **Table 4. Top protein-protein interaction partners of HSP47.**

Protein	Protein name	Chromosomal Location*	Cytoplasmic band*	Gene ID**	ENSEMBL ID***	Uniprot ID	Protein Length	Gene Synonyms	OMIM ****
CREBBP	CREB binding protein	16: 3,725,054-3,880,726	16p13.3	1387	ENSG00000005339	Q92793	2442	CBP, KAT3A, RSTS, RSTS1	600140
EP300	E1A binding protein p300	22: 41,091,786-41,180,079	22q13.2	2033	ENSG00000100393	Q09472	2414	p300, KAT3B, RSTS2	602700
CRTAP	Cartilage associated protein	3: 33,113,979-33,147,773	3p22.3	10491	ENSG00000170275	O75718	401	CASP, LEPREL3, OI7, P3H5	605497
DNAJB1	DnaJ (Hsp40) homolog subfamily B member 1	19: 14,514,770-14,529,770	19p13.12	3337	ENSG00000132002	P25685, Q6FHS4	340	HSPF1, Hdj1, Hsp40, RSPH16B, Sis1	604572
DNAJB6	DnaJ (Hsp40) homolog subfamily B member 6:	7: 157,335,381-157,417,439	7q36.3	10049	ENSG00000105993	O75190	326	DJ4, DnaJ, HHDJ1, HSJ-2, HSJ2, LGMD1D, LGMD1E, MRJ, MSJ-1	611332
FKBP	FK506-binding protein 4	6: 35,573,585-35,728,583	6p21.31	2289	ENSG00000096060	Q13451	457	AIG61, FKBP54, P54, PPIase, Ptg-10, FKBP5	602623
HSPA6	Heat shock 70kDa protein 6	1: 161,524,540-161,526,910	1q23.3	3310	ENSG00000173110	P17066	643	HSP70B'	140555
HSPA8	Heat shock 70kDa protein 8	11: 123,057,489-123,063,230	11q24.1	3312	ENSG00000109971	P11142, V9HW22	646	HEL-33, HEL-S-72p, HSC54, HSC70, HSC71, HSP71, HSP73, HSPA10, LAP-1, LAP1, NIP71	600816

doi:10.20944/preprints201811.0280.v1

HSPB1	heat shock protein beta-1	7: 76,302,544-76,304,295	7q11.23	3315	ENSG00000106211	P04792, V9HW43	205	CMT2F, HEL-S-102, HMN2B, HS.76067, HSP27, HSP28, Hsp25, SRP27	602195
HSPB2	heat shock protein beta-2	11: 111,912,242-111,914,093	11q23.1	3316	ENSG00000170276	Q16082	182	MKBP, HSP27, Hs.78846, LOH11CR1K	602179
HSPH1	Heat shock 105kDa/110kDa protein 1	13: 31,134,974-31,162,388	13q12.3	10808	ENSG00000120694	Q92598	858	HSP105, HSP105A, HSP105B, NY-CO-25	610703
HSBP1	Heat shock factor binding protein 1	16: 83,807,843-83,819,737	16q23.3	3281	ENSG00000230989	O75506	76	NPC-A-13	604553
HSP90AA1	Heat shock protein Hsp 90-alph(cytosolic), class A member 1	14: 102,080,738-102,139,699	14q32.31	3320	ENSG00000080824	P07900	854	EL52, HEL-S-65p, HSP86, HSP89A, HSP90A, HSP90N, HSPC1, HSPCA, HSPCAL1, HSPCAL4, HSPN, Hsp103, Hsp89, Hsp90, LAP-2, LAP2	140571
HSP90AB1	Heat shock protein Hsp 90-alpha (cytosolic), class B member 1	6: 44,246,166-44,253,888	6p21.1	3326	ENSG00000096384	P08238	724	D6S182, HSP84, HSP90B, HSPC2, HSPCB	140572
LEPRE1	Leucine proline-enriched proteoglycan (leprecan) 1	1: 42,746,335-42,767,084	1p34.2	64175	ENSG00000117385	Q32P28	736	GROS1, OI8, P3H1,	610339

doi:10.20944/preprints201811.0280.v1

KLF13	Kruppel-like factor 13	15: 31,326,855-31,435,665	15q13.3	51621	ENSG00000169926	Q9Y2Y9	288	BTEB3, FKLF2, NSLP1, RFLAT-1, RFLAT1	605328
P4HB	Prolyl 4-hydroxylase beta subunit	17: 81,843,159-81,860,694	17q25.3	5034	ENSG00000185624	P07237	508	CLCRP1, DSI, ERBA2L, GIT, P4Hbeta, PDI, PDIA1, PHDB, PO4DB, PO4HB, PROHB	176790
PIPB	Peptidyl-prolyl isomerase B	15: 64,155,812-64,163,205	15q22.31	5479	ENSG00000166794	P23284	216	B, CYP-S1, CYPB, HEL-S-39, OI9, SCYLP	123841
UBB	Ubiquitin B	17: 16,380,798-16,382,745	17p11.2	7314	ENSG00000170315	P0CG47, Q5U5U6	229	HEL-S-50	191339
UBC	Ubiquitin C	12: 124,911,604-124,917,368	12q24.31	7316	ENSG00000150991	P0CG48	685	HMG20	191340

*Genome assembly: GRCh38.p12 (GCA_000001405.27); **NCBI gene ID; ***Ensemb

doi:10.20944/preprints201811.0280.v1