Associations between DPP9 expression, survival and gene expression signature in human hepatocellular carcinoma: Comprehensive in silico analyses

Jiali Carrie Huang¹², Abdullah Al Emran¹², Justine Moreno Endaya¹², Geoffrey W McCaughan¹²³, Mark D Gorrell¹²*, Hui Emma Zhang¹²*

¹ Centenary Institute, Camperdown, New South Wales, Australia; j.huang@centenary.org.au (J.C.H.), a.alemran@centenary.org.au (A.A.E.), g.mccaughan@centenary.org.au (G.W.M.)
² The University of Sydney, Faculty of Medicine and Health, Sydney, New South Wales, 2006, Australia; jend8597@uni.sydney.edu.au (J.M.E.)
³ AW Morrow GE & Liver Centre, Royal Prince Alfred Hospital, Camperdown, New South Wales, 2050, Australia;
* Correspondence: e.zhang@centenary.org.au (H.E.Z.), m.gorrell@centenary.org.au (M.D.G.); Tel.: 61-2-95656156

Simple Summary: This in silico study aimed to investigate associations between dipeptidyl peptidase (DPP) 9 mRNA expression, survival and gene signature in human hepatocellular carcinoma (HCC). DPP9 loss-of-function exonic variants were mostly associated with cancers. In HCC patients, DPP9 and the closely related genes DPP4 and DPP8 were upregulated in liver tumours. High co-expression of genes that were positively correlated with DPP4, DPP8 and DPP9 was associated with poor survival in HCC patients. These findings strongly implicate the DPP4 gene family, especially DPP9, in the pathogenesis of human HCC and therefore encourages future functional studies.

Abstract: Dipeptidyl peptidase (DPP) 9, DPP8, DPP4 and fibroblast activation protein (FAP) are the four enzymatically active members of the S9b protease family. Associations of DPP9 with human liver cancer, exonic single nucleotide polymorphisms (SNPs) in DPP9 and loss of function (LoF) variants have not been explored. Human genomic databases including The Cancer Genome Atlas (TCGA) were interrogated to identify DPP9 LoF variants and associated cancers. Survival and gene signature analyses were performed on hepatocellular carcinoma (HCC) data. We found that DPP9 and DPP8 are intolerant to LoF variants. DPP9 LoF variants were most often associated with uterine carcinoma. Two DPP9 intronic SNPs that have been associated with lung fibrosis and COVID-19 were not associated with liver fibrosis or cancer. All four DPP4-like genes were overexpressed in liver tumours and their joint high expression was associated with poor survival in HCC. Increased DPP9 expression was associated with obesity in HCC patients. High expression of genes that positively correlated with overexpression of DPP4, DPP8, and DPP9 were associated with very poor survival in HCC. Enriched pathways analysis of these positively correlated genes featured Toll-like receptor and SUMOylation pathways. This comprehensive data mining suggests that DPP9 is essential for human survival and the DPP4 protease family is important in cancer pathogenesis.

Keywords: DPP9; SNPs; Hepatocellular carcinoma; Survival; TCGA; DPP4 gene family

1. Introduction

Dipeptidyl peptidase (DPP) 9, DPP4, DPP8 and fibroblast activation protein (FAP) are the enzymatic members of the DPP4 family of serine proteases and have been implicated in cancer pathogenesis [1]. DPP9 is ubiquitously expressed in tissues [2] and has diverse roles in cell behaviours [3], immune regulation [4-6] and cancer [7-9]. In non-small-cell lung cancer, poor 5-year survival has been associated with greater expression...
of DPP9 [7]. Similarly, in colorectal cancer, greater DPP9 expression is associated with poor-prognosis [8]. In contrast, in patients with oral squamous cell carcinoma, lower DPP9 expression correlates with poor survival [9]. These data suggest that DPP9 has different roles in various types of cancers.

Liver is rich in DPP9 enzyme activity [2] and DPP9 mRNA is upregulated in inflamed and fibrotic mouse liver [10]. Downregulating DPP9 or blocking its enzyme activity reduces cell adhesion and migration in the Huh7 human hepatoma cell line [11]. The role of DPP9 in human liver cancer has not been studied, particularly in liver hepatocellular carcinoma (HCC). HCC is the most common primary liver cancer, [12]. The major risk for HCC is chronic liver injury, with causes including non-alcoholic fatty liver disease (NAFLD), obesity, smoking tobacco, chronic hepatitis B or C, and chronic excessive alcohol consumption [13-15]. NAFLD, obesity and associated metabolic diseases are becoming the leading cause of HCC [16].

The DPP4 enzyme family is the S9b subfamily of the prolyl oligopeptidase (S9) family and share 27-60% protein sequence identity with each other [17, 18]. The proteases in the DPP4 enzyme family are able to hydrolyse peptides after a proline residue, and can remove a dipeptide from the N-terminus of substrates [18, 19]. Very few proteases have this capability. DPP4 and FAP are expressed on the cell surface, whereas DPP8 and DPP9 are intracellular [1]. DPP8 and DPP9 share the greatest similarity in protein sequence and crystal structure in this gene family [20]. There are two isoforms of DPP9 with distinct subcellular localisations: the long form (892 amino acids) is in both cytosol and nucleus because it contains a nuclear localisation sequence while the short form (863 amino acids) is only cytosolic [21]. The S9b proteases are active as dimers [22], with each monomer having two domains, the α/β-hydrolase domain and the β-propeller domain [20]. The β-propeller domain binds to protein ligands, and DPP9 has been found to bind to SUMO-1 [23] and filamin A [6] and associate with H-Ras [3], which are probably involved in roles of DPP9 in cell adhesion and migration. Selective inhibitors for DPP9 versus DPP8 are lacking, so individual functions of DPP9 are poorly resolved and derive primarily from genetic manipulations of mice and cell lines [1, 24, 25].

Large-scale human genetic databases have become increasingly powerful resources to study the impacts of somatic mutation in diseases, especially in cancer. Human gene sequence variants that are predicted to cause loss of function (LoF) of an encoded protein provide valuable data on studying human gene inactivation [26, 27]. Genome-wide association studies (GWAS) have found several risk loci mapping to the DPP9 gene on chromosome 19 (19p13.3) that contribute to human diseases, including alcoholism [28], interstitial lung abnormalities [29], pulmonary fibrosis [29, 30] and adolescent idiopathic scoliosis [31]. The intronic single nucleotide polymorphism (SNP) rs113683471 in DPP9 is associated with alcoholism [28]. Intronic SNPs rs12610495 and rs2109069 in DPP9 are genome-wide significant loci for severe COVID-19, idiopathic pulmonary fibrosis and idiopathic interstitial pneumonia [29, 30, 32, 33]. However, there are limited information on exonic SNPs in DPP9 and DPP9 LoF variants, or their associations with diseases.

This study identifies DPP9 exonic LoF variants in human genomic databases and their associations with survival and cancers. This study also reports the first comprehensive data mining of associations between DPP9 mRNA expression, survival and gene signature in human HCC.

2. Materials and Methods

Protein sequences
The amino acid sequences of the short and long forms of DPP9 were downloaded from the UniProt Consortium (https://www.uniprot.org/uniprot/Q86T12) on 7th April 2020. The long form of DPP9 and its amino acid numbering was used in this study.

Public databases for accessing loss of function variants

Public databases assessed in this study are listed in Table 1.

| Databases | URL | Data obtained date | Citation |
|-----------|-----|--------------------|---------|
| The Genome Aggregation database (gnomAD) | https://gnomad.broadinstitute.org/ | 5th April 2020 | [34] |
| The Cancer Genome Atlas (TCGA) | https://portal.gdc.cancer.gov/ | 20th April 2020 | [35] |
| Catalogue of Somatic Mutation in Cancer (COSMIC) | https://cancer.sanger.ac.uk/cosmic | 15th June 2020 | [36] |
| UniProt Consortium DatabasE of genomiC variation and Phenotype in Humans using Ensembl Resources (DECIPHER) | https://decipher.sanger.ac.uk/ | 7th April 2020 | [37] |
| GeneCards human gene database | https://www.genecards.org/ | 12th May 2020 | [39] |

Loss of function variants

The exonic variants that are predicted to cause LoF of DPP9 were exported from gnomAD, TCGA and COSMIC. Only the LoF variants that are annotated as either stop gained or frameshift were studied. The numbering of amino acid sequence of DPP9 in GnomAD is of the short form of DPP9. Therefore, to use the DPP9 long form, 29 was added to all protein sequence numbers for the gnomAD DPP9 LoF analyses. Variants having sequence number mismatch with the long form of DPP9 were excluded. In addition, if frameshift variants exported from gnomAD and TCGA occurred where there is no corresponding structural residue, these frameshift variants were considered to be unlikely to be DPP9 exonic LoF variants and thus were excluded from this study.

Intronic variants

The two DPP9 intronic variants rs12610495 and rs2109069 have the smallest p values that have been associated with critical illness caused by COVID-19 [33]. More genetic information about these two intronic variants was exported from PhenoScanner V2 [40].

Genetic variant tolerance parameters

The observed to expected ratio (o:e) of LoF, the probability of LoF intolerance (pLI) and haploinsufficiency score were retrieved from gnomAD and DECIPHER [38]. Tolerance percentile and GDI values were retrieved from GeneCards human gene database.

Gene correlation analysis

UALCAN is a comprehensive web database for investigating complete genetic or molecular data of cancers and allows in silico validation of genes of interest [41]. It is available online at http://ualcan.path.uab.edu/index.html. Both positively and negatively
correlated genes of \textit{DPPs} and \textit{FAP} were accessed using liver hepatocellular carcinoma (LIHC; HCC) dataset of TCGA. The expression level of each gene was displayed as Transcripts Per Million (TPM).

\textbf{Survival analysis}

OncoLnc database contains TCGA survival data to mRNA expression level for 8,647 patients from 21 cancer projects [42]. Survival data for 360 patients with HCC and 540 patients with UCEC were exported from OncoLnc. Normal patients’ data (n=50) were exported from UALCAN [41]. HCC patient survival was also analysed based on patients’ BMI (Table 2).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
\textbf{Status} & \textbf{BMI range} \\
\hline
Normal weight & 18.5-25 \\
Overweight & 25-30 \\
Obese & 30-40 \\
Extreme obese & >40 \\
\hline
\end{tabular}
\caption{Categories of BMI}
\end{table}

SurvExpress is an online database for cancer gene expression data using survival analysis, which is able to access the influence of co-expression of multiple genes on survival in a specific cancer type [43]. The genes that correlate in common amongst the DPP4 gene family were imported into SurvExpress and analysed for their associations with HCC patient survival (June 2016 dataset). Patient survival time and status were exported from SurvExpress and Kaplan-Meier curves were plotted.

\textbf{Reactome pathways enrichment analysis}

Reactome enrichment pathways of positively or negatively correlated genes in common between \textit{DPP9} and \textit{DPP8} were analysed from STRING database. The positively correlated genes in common amongst \textit{DPP9}, \textit{DPP8} and \textit{DPP4} were imported into ConsensusPathDB (CPDB) [44] and identified as gene symbols. The analysis was performed under gene set analysis and over-representation analysis.

\textbf{Data and statistical analyses}

Figures and Kaplan-Meier curves were generated in GraphPad Prism (version 8.4.2, San Diego, CA, USA). Data statistics were analysed in GraphPad Prism. Survival analysis was performed in Prism using Log-rank (Mantel–Cox) to assess the significance between the groups. The Mann-Whitney U test was used to compare differences between two independent groups. The Kruskal-Wallis one-way analysis of variance with Tukey’s post-hoc test was used to compare differences amongst multiple groups. Significance was assigned to p-values *<0.05, **<0.01, ****<0.0001.

\section{Results}

\subsection{DPP9 loss of function (LoF) variants and disease associations}

Large-scale human genetic databases provide useful tools to study protein LoF variants and their associations with diseases. \textit{DPP9} LoF has not been explored previously. Therefore, we assessed data from several human genomic databases to identify \textit{DPP9} LoF and potentially associated diseases. Gaining a premature termination codon (Ter), or stop codon, anywhere in a gene of the DPP4 gene family is expected to ablate enzyme activity [45]. We have shown previously that mature termination mutants of \textit{DPP8} are inactive, even if lacking only the C-terminal alpha-helix [45]. Similarly, any frameshift variant is
almost certain to lack activity due to nonsense sequence causing a Ter. Therefore, the exonic gene variants considered to be LoF are very likely true LoF.

In gnomAD, there were 18 exonic variants of DPP9 that were tagged as LoF. However, examining the sequence of long-form DPP9 in the UniProt Consortium database revealed that some variants were mismatched with UniProt and thus were excluded. Only exonic LoF variants that were annotated as either stop-gained or frameshift variants were included in this study. We identified three frameshift variants and seven stop gained variants that were exonic in gnomAD (Table 3). There was no record of associated diseases for each mutation variant. In TCGA, we identified eight DPP9 LoF variants (Table 4), and none of these variants were the same as found in gnomAD. Four DPP9 variants were designated as stop gained, with three of them associated with primary uterine corpus endometrial carcinoma (UCEC). Another four DPP9 variants were designated as frameshift mutation, with half of them found in patients diagnosed with stomach adenocarcinoma (Table 4). None of the DPP9 variants in TCGA were found to be associated with HCC. There were many more DPP8 than DPP9 LoF gene variants in TCGA, found in diverse types of cancer, mostly in UCEC and one in HCC (Supplementary Table 1). In a separate database, COSMIC, eleven DPP9 variants had been tagged as nonsense mutation and were substitution mutations resulting in Ter codons (Table 5). These variants were associated with different types of carcinoma, most often lung carcinoma. All variants that were likely to cause LoF for DPP8 or DPP9 have a Ter. All exonic DPP9 variants found in each of the three databases were mapped to the DPP9 protein structure (Supplementary Figure 1).

**Table 3. DPP9 LoF variants from gnomAD**

| Position | rsID       | Reference | Alternate | Protein consequence | Annotation            | Allele frequency |
|----------|------------|-----------|-----------|---------------------|-----------------------|------------------|
| 4682801  | rs766329416| CCT       | C         | p.Gly794ValTer3     | frameshift variant    | 7.89E-6          |
| 4694820  | rs1360075526| AGAAG     | A         | p.Phe456IleTer34    | frameshift variant    | 4.02E-6          |
| 4714106  | rs1162979581| GC        | G         | p.Arg100ProTer58    | frameshift variant    | 4.1E-6           |
| 4683566  | rs1199678192| G         | A         | p.Arg752Ter         | Stop gained           | 4.02E-6          |
| 4684713  | rs1165785000| G         | A         | p.Arg714Ter         | Stop gained           | 4.1E-6           |
| 4676614  | rs1479652400| C         | A         | p.Glu881Ter         | Stop gained           | 1.25E-5          |
| 4694808  | rs1365336720| G         | A         | p.Gln461Ter1        | Stop gained           | 4.02E-6          |
| 4694825  | rs1418069794| A         | ATCTTC-TATC| p.Asp454Ile455insArgTer-Lys | Stop gained | 4.02E-6        |
| 4702672  | rs1466943995| G         | A         | p.Gln276Ter         | Stop gained           | 7.53E-6          |
| 4705965  | rs779854175 | G         | A         | p.Arg111Ter1        | Stop gained           | 1.2E-5           |

1 Gln461 and Arg111: no corresponding residue in the primary structure of DPP9.
Table 4. DPP9 LoF variants in TCGA. Simple somatic mutation (SSM) affected frequency was calculated as the number of cases affected by a specific mutation in a TCGA disease project divided by the number of cases tested for SSM in that disease project in TCGA. *Termination codon (Ter).

| Position       | Reference | Alternate | Protein consequence | Annotation | Disease                        | #SSM affected cases and frequency |
|----------------|-----------|-----------|--------------------|------------|--------------------------------|-----------------------------------|
| 4685761_4685762insG |           | p.Asp63 3Gly*51 | Frameshift         | Stomach Adenocarcinoma | 1/440 (0.23%)                     |
| 4684670delT     |           | p.Asn72 4Thr*27 | Frameshift         | Stomach Adenocarcinoma | 1/440 (0.23%)                     |
| 4679893delA     |           | p.Phe84 3Ser*10 | Frameshift         | Colon Adenocarcinoma   | 1/400 (0.25%)                     |
| 4676645_4676646insT |       | p.Asn86 6Lys*16 | Frameshift         | Uterine Corpus Endometrial Carcinoma | 1/530 (0.19%)                     |
| 4702132         | G         | A         | p.Arg30 3*         | Stop gained          | Uterine Corpus Endometrial Carcinoma | 2/530 (0.38%)                     |
| 4704271         | C         | A         | p.Glu15 4*         | Stop gained          | Lung adenocarcinoma              | 1/567 (0.18%)                     |
| 4695519         | C         | T         | p.Trp40 4*         | Stop gained          | Uterine Corpus Endometrial Carcinoma | 1/530 (0.19%)                     |
| 4685745         | C         | A         | p.Glu63 8*         | Stop gained          | Uterine Corpus Endometrial Carcinoma | 1/530 (0.19%)                     |

1 Glu154 does not have a corresponding residue in the primary structure of DPP9.

Table 5. DPP9 LoF variants in COSMIC. Nonsense mutation that is a substitution mutation resulting in a termination codon (*). CDS= coding sequence; AA= amino acid; SSM= simple somatic mutation.

| CDS mutation | AA mutation | Legacy mutation ID | Type         | Disease                                   | #SSM affected cases |
|--------------|-------------|--------------------|--------------|-------------------------------------------|--------------------|
| c.43G>T      | p.Gly15*    | COSM3960 221       | Nonsense     | Lung squamous cell carcinoma              | 2                  |
| c.180G>A     | p.Trp60*    | COSM9150 037       | Nonsense     | Plasma cell myeloma                       | 1                  |
| c.306C>G     | p.Tyr10 2*  | COSM4982 488       | Nonsense     | Oesophagus squamous cell carcinoma         | 3                  |
| c.460G>T     | p.Glu15 4*  | COSM7761 095       | Nonsense     | Lung adenocarcinoma                       | 1                  |
| c.864C>A     | p.Cys28 8*  | COSM8582 330       | Nonsense     | Stomach carcinoma                         | 1                  |
| c.907C>T     | p.Arg30 3*  | COSM9984 42        | Nonsense     | Endometrioid carcinoma                    | 2                  |
| c.976G>T     | p.Glu32 6*  | COSM7365 414       | Nonsense     | Lung adenocarcinoma                       | 1                  |
| c.1090G>T   | p.Glu36 4*       | COSM8089 586 | Nonsense | Malignant melanoma | 1 |
| c.1385C>G   | p.Ser462*       | COSM7365 412 | Nonsense | Lung adenocarcinoma | 1 |
| c.2584C>T   | p.Gln86 2*      | COSM7322 837 | Nonsense | Biliary tract adenocarcinoma | 1 |
| c.2668G>T   | p.Glu89 0*      | COSM6866 899 | Nonsense | Skin squamous cell carcinoma | 1 |

1 Gly15, Glu154 and Ser462: do not have corresponding residue in the primary structure of DPP9.

There were 22 DPP8 nonsense mutations in COSMIC, most often in colon adenocarcinoma and one in HCC (Supplementary Table 2). DPP9 variants Glutamic acid (Glu) 154 to Ter and Arginine (Arg) 303 to Ter were both reported in both TCGA and COSMIC, but with different nucleotide substitutions. Another anomaly was that in COSMIC, both endometrioid carcinoma and colon adenocarcinoma were associated with the Arg303* variant, but in TCGA only endometrioid carcinoma was recorded.

3.2. Intronic SNPs in DPP9

Intronic SNPs rs12610495 and rs2109069 in DPP9 are genome-wide significant loci for severe COVID-19, idiopathic pulmonary fibrosis and idiopathic interstitial pneumonia [29, 30, 32, 33]. These SNPs are located 5' to the open reading frame, at nucleotides 4717672 and 4719443 of chromosome 19. They are important DPP9 SNPs because they are relatively common, with allelic frequencies of 0.706 to 0.828 for rs12610495 and 0.186 to 0.321 for rs2109069 (Supplementary Table 3). These SNPs have been associated with lower DPP9 expression levels in human lung [33]. We found that these SNPs are not associated with liver fibrosis or cancer in the PhenoScanner database.

3.3. Genetic variation tolerance in DPP9 and the DPP4 gene family

Some databases incorporate scoring systems that determine tolerance towards genetic variations. We analysed genetic variation tolerance in the DPP4 gene family using five scores: observed to expected ratio (o:e), probability of LoF intolerance (pLI), haploinsufficiency score, Tolerance Percentile, and gene damage index (GDI) scores. When o:e is approximately 0.1, the gene is haploinsufficient and heterozygous LoF is very probably not tolerated; when o:e is between 0.1 and 0.5, the heterozygous LoF may be tolerated; when o:e is close to 1, LoF is tolerated [46]. DPP8 and DPP9 had low o:e, little more than 0.1, DPP4 had a moderate o:e less than 0.5 and FAP has a high o:e (Table 6). The probability of LoF intolerance (pLI) is the probability of a given gene being haploinsufficient; genes with pLI ≥ 0.9 are extremely intolerant to LoF variation and genes with pLI ≤ 0.1 are LoF tolerant [46]. The pLI scores for DPP9, DPP8, DPP4 and FAP (Table 6) suggested that DPP8 and DPP9 LoF variants are poorly tolerated. Haploinsufficient genes only contain a single copy of a diploid genome that changes the organism phenotype from normal to a disease state, and many of the LoF associated haploinsufficient genes contribute to metabolic disorders and tumorigenesis [47]. The haploinsufficiency score was greater for DPP9 than for DPP8, DPP4 or FAP, and, like the o:e and pLI, indicated that DPP9 is likely to exhibit haploinsufficiency whereas DPP4 and FAP are unlikely to exhibit haploinsufficiency. The gene damage index (GDI) scores for these four enzymes were low (Table 6), suggesting that LoF variants are likely to be disease-causing [48]. Genes in the 25th Tolerance Percentile and below are considered relatively intolerant to variation. Unlike the other haploinsufficiency indications, GDI and Tolerance Percentile did not discriminate amongst the four genes (Table 6).

Table 6. Genetic variation tolerance in the DPP4 gene family. The o:e ratio, pLI and haploinsufficiency score were retrieved from gnomAD and DECIPHER. Tolerance percentile and GDI values
were retrieved from the GeneCards human gene database. o:e, the observed to expected ratio. pLI, the probability of LoF intolerance. GDI, gene damage index score.

| Gene | O:e ratio of LoF | pLI | Haploinsufficiency score | Tolerance percentile | GDI |
|------|------------------|-----|--------------------------|----------------------|-----|
| DPP9 | 0.17 (0.1-0.3)   | 0.99| 74.03                    | 7.37%                | 3.14|
| DPP8 | 0.14 (0.08-0.27) | 1   | 24.90                    | 26.2%                | 1.79|
| DPP4 | 0.44 (0.31-0.62) | 0   | 8.9                      | 15%                  | 2.81|
| FAP  | 0.86 (0.68-1.1)  | 0   | 8.57                     | 10.2%                | 2.77|

3.4. Association of DPP9 expression with patient survival in HCC and UCEC

Previous studies reported associations of DPP9 expression with patient survival in non-small-cell lung cancer [7], colorectal cancer [8] and oral squamous cell carcinoma [9]. Many DPP8 and DPP9 variants were identified in patients with UCEC and HCC in our study (Table 4 and Supplementary Tables 1-2). Therefore, we analysed whether DPPs expression is associated with patient survival in HCC and UCEC. Clinical parameters and mRNA expression of DPPs in patients with HCC and UCEC were exported from TCGA. Patient demographics are presented in Supplementary Table 4. Kaplan-Meier analysis for DPP9, DPP8, DPP4 and FAP mRNA suggested that expression of none of these four genes was associated with HCC overall survival (OS) (Figure 1A-D). In contrast, patients with low DPP9 mRNA expression had significantly shorter survival time for UCEC (Figure 1E). Less than 15% of UCEC patients with low DPP9 expression had survived for 10 years after diagnosis. Unlike DPP9, DPP8, DPP4 and FAP expression did not associate with OS in UCEC (Figure 1F-H). This data indicates that DPP9 might have an important role in UCEC.
Figure 1. Kaplan-Meier curves for DPP9, DPP8, DPP4 and FAP in liver tumours (hepatocellular carcinoma, HCC; A-D) and uterine corpus endometrial carcinoma (UCEC; E-H). High or low was defined as the mRNA expression values in the upper or lower quartile, respectively. P values were calculated by Log-rank (Mantel-Cox) test. **p<0.01.

3.5. Association of DPP9 expression with clinical parameters and survival outcome in HCC

We next explored closely the association of DPP9 with clinical parameters relevant to HCC, including gender, age, BMI and cancer stages. In HCC patients, all four DPP genes were significantly upregulated in tumours compared to normal tissue (Figure 2A-D). Significantly more expression of DPP9 was detected in female than male (Figure 2E). DPP9 expression did not differ significantly amongst six age groups (Figure 2F) and different cancer stages (Figure 2G).
Figure 2. The relative mRNA expression level of DPP9, DPP8, DPP4 and FAP in liver tumours (A-D) and association of DPP9 expression with gender (E), age (F) and cancer stages (G) in HCC patients. P values were calculated by Mann-Whitney U test (A-E) and Kruskal-Wallis one-way analysis of variance with Tukey’s post-hoc test (F-G). *p<0.05, ****p<0.0001.

Obesity is a well-known risk factor for HCC [16], so we examined DPP9 expression with HCC patient survival based on body mass index (BMI; BMI categories in Table 2). We stratified HCC patients based on BMI and found that DPP9 mRNA was significantly upregulated in the obese and extreme obese patients compared to patients with normal body weight (Figure 3A). Survival analysis suggested that DPP9 was not significantly associated with patient survival when patients are overweight (Figure 3B), or obese/extreme obese (Figure 3C), however there was trend to poor prognosis in the overweight group.
Figure 3. The relative mRNA expression and survival analysis of DPP9 in HCC based on patient BMI. (A) The mRNA expression of DPP9 based on HCC patient BMI compared to non-diseased subjects. P values were calculated by Kruskal-Wallis one-way analysis of variance with Tukey’s post-hoc test, with *p<0.05, **p<0.01, ****p<0.0001. (B) Kaplan-Meier curve for DPP9 expression in overweight HCC patients. (C) Kaplan-Meier curve for DPP9 expression in obese and extreme obese HCC patients. The high (red) and low (green) mRNA expression levels of genes in liver tumours were stratified based on median expression value. P values and hazardous ratio (HR) with confidence interval were calculated by Log-rank (Mantel-Cox) test.

3.6. Association of expression of DPPs with poor survival of patients with HCC

To further investigate DPP4 gene family expression in liver cancer prognosis, we stratified DPP9, DPP8 and DPP4 based on median expression in HCC. High co-expression of DPP9, DPP8 and DPP4 was not associated with OS (Figure 4A). However, high co-expression of DPP9, DPP8, DPP4 and FAP in the liver tumour was significantly associated with poor survival in HCC patients (Figure 4B). This data strengthens evidence that FAP might contribute to poor outcomes, possibly due to its role in liver cirrhosis and steatosis, which is strongly associated with poor liver function and liver cancer [49, 50].
Figure 4. Survival analysis of high co-expression of DPP9, DPP8, DPP4 and FAP in HCC patients. (A) Kaplan-Meier curve for high co-expression of DPP9, DPP8 and DPP4 in HCC patients. (B) Kaplan-Meier curve for high co-expression of DPP9, DPP8, DPP4 and FAP in HCC patients. The high (red) and low (green) mRNA expression levels of genes in liver tumours were stratified based on median expression value. P values were calculated by Log-rank (Mantel-Cox) test. *p<0.05.

3.7. Genes correlated with DPPs significantly associated with poor prognosis in HCC and multiple oncogenic and epigenetic pathways

Having shown that DPP9, DPP8, DPP4 and FAP were overexpressed in liver tumours (Figure 2A-D), and high co-expression of the DPP4 gene family members was significantly associated with shorter survival time in HCC (Figure 4B), we next assessed the gene expression signature correlated with the DPP4 gene family in HCC patients to gain mechanistic insights. Firstly, we analysed the two homologs DPP8 and DPP9 and found that 661 genes were positively correlated in common between DPP8 and DPP9 (Figure 5A). Reactome pathway enrichment analysis revealed that these genes were associated with tumour suppressor TP53 regulation and with chromatin remodeling pathways in HCC (Figure 5B). In contrast, only 10 negatively correlated genes in common between DPP8 and DPP9 were identified and they were associated with metabolism related pathways (Supplementary figure 2B).
Figure 5. Enriched Reactome pathways of genes that were positively correlated with DPP9 and DPP8. (A) Venn diagram of the genes that were positively correlated with DPP9 and DPP8. “Pos cor” refers to positive correlation. (B) Enriched Reactome pathways associated with positively correlated genes in common between DPP9 and DPP8.

A similar approach was undertaken to study the gene expression signature correlated with all three genes, DPP9, DPP8 and DPP4. We found that 268 genes were positively correlated in common amongst DPP9, DPP8 and DPP4 in HCC patients (Figure 6A). Most importantly, high expression of these positively correlated gene signatures was significantly associated with poor prognosis in HCC patients (Figure 6B, Logrank ****p<0.0001).
Figure 6. Survival analysis and enriched Reactome pathways analysis on genes that were positively correlated with DPP9, DPP8 and DPP4. (A) Venn diagram of the genes that were positively correlated with DPP9, DPP8 and DPP4. The gene in the blue section is TPRA1. “Pos cor” refers to positive correlation. (B) Kaplan-Meier curve for the 268 genes that were positively correlated and in-common among DPP9, DPP8 and DPP4 in HCC patients. The high (red) and low (green) mRNA expression levels of genes in liver tumours were stratified based on median expression value. P values were calculated by Log-rank (Mantel-Cox) test. ****p<0.0001. (C) Enriched Reactome pathways of genes that were positively correlated in-common amongst DPP9, DPP8 and DPP4.

To gain a better understanding of how DPPs may influence survival, we analysed the biological pathways that were enriched in genes that were positively correlated with DPPs. There were 266 significantly enriched pathways according to Reactome annotations for those 268 genes (Supplementary Table 5). Amongst the 15 top-ranked pathways (Figure 6C), Toll-like receptors (TLRs) cascade, especially TLR9 and TLR7/8 cascade, MyD88 dependent cascade initiated on endosome, N-glycosylation and SUMOylation pathways are dominant. When FAP was added to this analysis, only 7 in-common genes emerged, most of which are the zinc finger gene regulators LIMS1, PHF20L1, ZEB1 and ZNFX1 (Supplementary Figure 3).

4. Discussion

This is the first analysis of DPP9 exonic variants and LoF variants in human genomic databases and their associations with survival and cancers. There were 27 unique DPP9 LoF variants found from our analyses of gnomAD, TCGA and COSMIC. Notably, patients with DPP9 LoF variants were most commonly diagnosed with cancers. This differs from the intronic SNPs of DPP9, which lower DPP9 expression [33], but have been associated with pulmonary fibrosis and severe COVID [29, 30, 32, 33]. Uterine corpus endometrial carcinoma (UCEC) and lung carcinoma were the most common diagnosed cancers in
patients with DPP9 LoF variants and low DPP9 expression was associated with poor survival in UCEC. Therefore, DPP9 might be protective in UCEC. Our datamining evidence that DPP9 LoF is associated with lung carcinoma does not support a previous study that linked high DPP9 expression with poor survival in human lung cancer [7]. Perhaps the dominant role of DPP9 differs in different cancer types and differs depending upon whether high or low DPP9 expression is a recent response or is life-long.

We observed that intrahepatic DPP9 expression is greater in females and that DPP8 and DPP9 LoF variants are more commonly diagnosed with a gynaecological cancer, such as UCEC, cervical cancer or ovarian cancer. These data might suggest that DPP9 is a suppressor of gynaecological cancers, which would align with the recently discovered role of DPP9 in BRCA2 homeostasis in DNA repair [51]. However, in both TCGA and COSMIC, there were limited clinical data concerning patient outcomes, especially in COSMIC where many cases were filed as unknown gender, age and diseases. This is a common limitation of in-silico investigations.

Our investigation found that DPP9 and DPP8 are very likely essential for life as they are LoF intolerant in human. Genetic variation tolerance parameters universally suggested that DPP9 was LoF intolerant (Table 7). LoF intolerance has been associated with successful targets of medicinal therapies for cancer [26]. The necessity of DPP9 is conserved in mice, where constitutive homozygous ablation of DPP9 enzyme activity is neonate lethal [25, 52-54]. People who are homozygous for a DPP9 LoF variant would not survive to be included in the searched databases and any that do survive possibly possess a defect in a currently unknown [25] molecular pathway that causes death in DPP9 deficiency. Another implication of this information is that the LoF DPP9 variants in genomic databases are expected to be heterozygous and would therefore have lifelong low DPP9 expression throughout the body.

The databases contained some data mismatches about DPP9 sequences, presumably due to automation of data importation. There were more LoF variants in gnomAD than in COSMIC, but some of the listed LoF locations were not found in either long form or short form DPP9. Moreover, in both TCGA and COSMIC, variants of Arg303* and Glu154* were reported, but the mutations of nucleotides differed between the two databases. Also, in patients with the Arg303* variant, COSMIC reported both endometrial carcinoma and large intestinal adenocarcinoma, while TCGA only reported endometrial carcinoma. Hence, using a single database can be misleading and unmatched information is inevitable when using multiple databases. Therefore, database derived information needs to be carefully and critically assessed.

Our TCGA datamining on the DPP4 gene family showed that in HCC patients, DPP9, DPP8, DPP4 and FAP were all overexpressed in liver tumours. Most notably, high expression of all four genes showed poorer survival in HCC patients (Figure 4B). Therefore, DPP targeted therapy might need to target all four enzymes together. Supporting this concept, VbP (Talabostat; PT-100; BXCL-701), which exerts potent inhibition of DPP9 and DPP8 and moderate inhibition of DPP4 and FAP, has been shown to lessen tumour burden in lung cancer and sarcoma models [55-57].

Enriched biochemical pathways analysis provided insights into differentially expressed genes correlated with DPPs in human HCC. Positively correlated genes in common between DPP8 and DPP9 were associated with oncogenic and chromatin remodeling pathways in HCC, whilst 10 negatively correlated genes were associated with metabolism related pathways (Supplementary figure 2A). Mutation of the tumour suppressor TP53 is associated with poor outcome for HCC patients [58]. The top ranking TP53 regulation pathway in our analyses indicates that DPP8 and DPP9 overexpression could be oncogenic in human HCC. The downregulation of metabolism genes with DPP8 and DPP9
upregulation concords with previous data showing that DPP9 deficiency dysregulates metabolic pathways in neonatal liver and gut [53], and that DPP9 hydrolyses metabolism related substrates including adenylate kinase 2 [59, 60] and nucleobindin1 [61] and regulates preadipocyte differentiation [62]. Moreover, people with obesity are more likely to develop HCC [63, 64]. DPP9 expression was significantly upregulated in HCC patients who were obese and extreme obese compared to those with normal weight. Taken together, this information suggests that DPP9 has an important role in liver manifestations of the metabolic syndrome, which drives progression to NASH, fibrosis and HCC.

The high expression of positively correlated gene signatures in-common amongst DPPs were associated with considerably shorter survival time in HCC patients. Amongst the identified Reactome pathways, TLR cascade, glycosylation and SUMOylation were the most frequent pathways enriched with increased DPP9 mRNA expression. No association between DPP9 and glycosylation is known. SUMOylation is mostly associated with SUMOylation of chromatin organization proteins, and SUMOylation of DNA damage response and repair proteins, which could have roles in cancer development. SUMO1 is a DPP9 interacting protein, acting as an allosteric regulator of DPP9 and thereby influencing its enzymatic activity [23]. TLR7/8 and TLR9 signalling pathways have not been linked to DPP9 previously. The expression of TLRs is low in healthy liver [65]. TLRs are upregulated in HCC tissues [66], and inhibition of TLR7 and TLR9 eliminates the proliferation of liver cancer cells [66, 67]. Therefore, overexpression of DPPs might activate these TLR cascades, thereby inducing overactivation of the innate immune system and pathological changes.

5. Conclusions

This study reports the first comprehensive data mining of associations between DPP9 expression, survival and gene expression signature in human HCC. DPP9 was found to be very LoF intolerant and its LoF exonic variants were mostly associated with cancers. Two intronic variants that lower DPP9 expression and have been linked to lung fibrosis and COVID-19 were not associated with HCC or liver fibrosis. In HCC patients, DPP9 and many closely related genes were significantly upregulated in liver tumours. High co-expression of genes that were positively correlated in-common amongst DPP4, DPP8 and DPP9 were associated with poor survival of HCC patients, and included Toll-like receptors (TLRs), glycosylation and SUMOylation. These findings strongly implicate the DPP4 gene family, especially DPP9, in the pathogenesis of human HCC and therefore requires future functional studies.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Table S1: DPP8 LoF variants in TCGA, Table S2: DPP8 LoF variants in COSMIC, Table S3: Genome-wide significant loci for severe COVID-19: Intrinsic DPP9 variants rs12610495 and rs2109069, Table S4: The LIHC/HCC (n=360) and UCEC (n=540) patient demographics. Figure S1: DPP9 variants mapping to the DPP9 protein structure, Figure S2: Enriched Reactome pathways of genes that were negatively correlated and in-common between DPP9 and DPP8, Figure S3: Survival analysis on genes that were positively correlated and in-common amongst DPP9, DPP8, DPP4 and FAP in HCC.

Author Contributions: Conceptualization, J.C.H., M.D.G., H.E.Z.; methodology, J.C.H., A.A.E.; formal analysis, J.C.H., A.A.E., J.M.E.; investigation, J.C.H., A.A.E., M.D.G., H.E.Z.; data curation, J.C.H., A.A.E., J.M.E.; writing—original draft preparation, J.C.H., M.D.G., H.E.Z.; writing—review and editing, J.C.H., A.A.E., J.M.E., G.W.M., M.D.G., H.E.Z.; supervision, M.D.G., H.E.Z.; funding acquisition, M.D.G., H.E.Z., G.W.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Australian National Health and Medical Research Council Grant 1113842, a Perpetual Trustees IMPACT Philanthropy grant IPAP2019/1295, and a Kickstart grant from The University of Sydney.
Acknowledgments: We thank Dr. Ulf Schmitz and Dr. Richard D. Bagnall from Centenary Institute for kindly providing expert guidance. We thank the many people involved in making the TCGA, COSMIC and LIHC data available for study.

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

References

1. Zhang, H.E., Y. Chen, F.M. Keane, and M.D. Gorrell, Advances in understanding the expression and function of dipeptidyl peptidase 8 and 9. Molecular Medicine Research, 2013. 11(12): p. 1487-1496.

2. Yu, D.M.T., K. Ajami, M.G. Gall, J. Park, C.S. Lee, K.A. Evans, E.A. McLaughlin, M.R. Pitman, C.A. Abbott, G.W. McLaughan, et al., The In Vivo Expression of Dipeptidyl Peptidases 8 and 9. Journal of Histochemistry & Cytochemistry, 2009. 57(11): p. 1025-1040.

3. Yao, T.-W., W.-S. Kim, D.M.T. Yu, G. Sharbeen, G.W. McLaughan, K.-Y. Choi, P. Xia, and M.D. Gorrell, A Novel Role of Dipeptidyl Peptidase 9 in Epidermal Growth Factor Signaling. Molecular Cancer Research, 2011. 9(7): p. 948.

4. Okondo, M.C., S.D. Rao, C.Y. Taabazuing, A.J. Chui, S.E. Poplawski, D.C. Johnson, and D.A. Bachovchin, Inhibition of Dpp8/9 Activates the Nlrp1b Inflammasome. Cell Chemical Biology, 2018. 25(3): p. 262-267.e5.

5. Okondo, M.C., D.C. Johnson, R. Sridharan, E.B. Go, A.J. Chui, M.S. Wang, S.E. Poplawski, W. Wu, Y. Liu, J.H. Lai, et al., DPP8 and DPP9 inhibition induces pro-caspase-1-dependent monocyte and macrophage pyroptosis. Nature Chemical Biology, 2017. 13(1): p. 46-53.

6. Justa-Schuch, D., M. Silva-Garcia, E. Pilla, M. Engkelke, M. Kilisich, C. Lenz, U. Möller, F. Nakamura, H. Urlaub, and R. Geiss-Friedlander, DPP9 is a novel component of the N-end rule pathway targeting the tyrosine kinase Syk. Elife, 2016. 5: e16370.

7. Tang, Z., J. Li, Q. Shen, J. Feng, H. Liu, W. Wang, L. Xu, G. Shi, X. Ye, M. Ge, et al., Contribution of upregulated dipeptidyl peptidase 9 (DPP9) in promoting tumorigenesis and cell biology and the prediction of poor prognosis in non-small cell lung cancer (NSCLC). International Journal of Cancer, 2017. 140(7): p. 1620-1632.

8. Yokobori, T., Editorial Comment on: DPP9 Increases Chemoresistance and is an Indicator of Poor Prognosis in Colorectal Cancer. Annals of Surgical Oncology, 2020. 27: p. 4084-4085.

9. Wu, Q.-Q., M. Zhao, G.-Z. Huang, Z.-N. Zheng, Y. Chen, W.-S. Zeng, and X.-Z. Lv, Fibroblast Activation Protein (FAP) Overexpression Induces Epithelial-Mesenchymal Transition (EMT) in Oral Squamous Cell Carcinoma by Down-Regulating Dipeptidyl Peptidase 9 (DPP9). Oncotargets and Therapy, 2020. 13: p. 2599-2611.

10. Choudhary, S., Y. Chen, T.-W. Yao, K. Ajami, X.M. Wang, Y. Popov, D. Schuppan, P. Bertolino, G.W. McLaughan, D.M. Yu, et al., Regulation of dipeptidyl peptidase 8 and 9 expression in activated lymphocytes and injured liver. World Journal of Gastroenterology, 2013. 19(19): p. 2883-93.

11. Zhang, H.E., Y. Chen, C. Wadham, G.W. McLaughan, F.M. Keane, and M.D. Gorrell, Dipeptidyl peptidase 9 subcellular localization and a role in cell adhesion involving focal adhesion kinase and paxillin. Biochimica et Biophysica Acta - Molecular Cell Research, 2015. 1853(2): p. 470-480.

12. Badvie, S., Hepatocellular carcinoma. Postgraduate Medical Journal, 2000. 76(891): p. 4.

13. Forner, A., M. Reig, and J. Bruix, Hepatocellular carcinoma. The Lancet, 2018. 391(10127): p. 1301-1314.

14. El-Serag, H.B., Hepatocellular Carcinoma: An Epidemiologic View. Journal of Clinical Gastroenterology, 2002. 35(Suppl 2): p. S72-S78.

15. El-Serag, H.B. and K.L. Rudolph, Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. Gastroenterology, 2007. 132(7): p. 2557-2576.

16. Ray, K., NAFLD—the next global epidemic. Nature Reviews Gastroenterology & Hepatology, 2013. 10(11): p. 621-621.

17. Ajami, K., C.A. Abbott, G.W. McLaughan, and M.D. Gorrell, Dipeptidyl peptidase 9 has two forms, a broad tissue distribution, cytoplasmic localization and DPIP-like peptidase activity. Biochimica et Biophysica Acta - Gene Structure and Expression, 2004. 1679(1): p. 18-28.

18. Yu, D.M.T., T.-W. Yao, S. Chakraborti and N.S. Dhal, Physiological Aspects of Proteases, S. Chakraborti and N.S. Dhal, Editors. 2017, Springer: Singapore. p. 23-46.

19. Gall, M.G. and M.D. Gorrell, The multifunctional post-proline dipeptidyl peptidase, DPP9. in mice, cell biology and immunity, in Pathophysiological Aspects of Proteases, S. Chakraborti and N.S. Dhal, Editors. 2017, Springer: Singapore: p. 23-46.

20. Choudhary, S., Y. Chen, T.W. Yao, K. Ajami, X.M. Wang, Y. Popov, D. Schuppan, P. Bertolino, G.W. McLaughan, D.M. Yu, et al., Regulation of dipeptidyl peptidase 8 and 9 expression in activated lymphocytes and injured liver. World Journal of Gastroenterology, 2013. 19(19): p. 2883-93.

21. Zhang, H.E., Y. Chen, C. Wadham, G.W. McLaughan, F.M. Keane, and M.D. Gorrell, Dipeptidyl peptidase 9 subcellular localization and a role in cell adhesion involving focal adhesion kinase and paxillin. Biochimica et Biophysica Acta - Molecular Cell Research, 2015. 1853(2): p. 470-480.

22. Badvie, S., Hepatocellular carcinoma. Postgraduate Medical Journal, 2000. 76(891): p. 4.

23. Forner, A., M. Reig, and J. Bruix, Hepatocellular carcinoma. The Lancet, 2018. 391(10127): p. 1301-1314.

24. El-Serag, H.B., Hepatocellular Carcinoma: An Epidemiologic View. Journal of Clinical Gastroenterology, 2002. 35(Suppl 2): p. S72-S78.

25. El-Serag, H.B. and K.L. Rudolph, Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. Gastroenterology, 2007. 132(7): p. 2557-2576.

26. Ray, K., NAFLD—the next global epidemic. Nature Reviews Gastroenterology & Hepatology, 2013. 10(11): p. 621-621.

27. Ajami, K., C.A. Abbott, G.W. McLaughan, and M.D. Gorrell, Dipeptidyl peptidase 9 has two forms, a broad tissue distribution, cytoplasmic localization and DPIP-like peptidase activity. Biochimica et Biophysica Acta - Gene Structure and Expression, 2004. 1679(1): p. 18-28.

28. Gall, M.G. and M.D. Gorrell, The multifunctional post-proline dipeptidyl peptidase, DPP9. in mice, cell biology and immunity, in Pathophysiological Aspects of Proteases, S. Chakraborti and N.S. Dhal, Editors. 2017, Springer: Singapore: p. 23-46.
61. Zhang, H., S. Maqsudi, A. Rainczuk, N. Duffield, J. Lawrence, F.M. Keane, D. Justa-Schuch, R. Geiss-Friedlander, M.D. Gorrell, and A.N. Stephens, Identification of novel dipeptidyl peptidase 9 substrates by two-dimensional differential in-gel electrophoresis. FEBS Journal, 2015. 282(19): p. 3737-57.
62. Han, R., X. Wang, W. Bachovchin, Z. Zukowska, and J.W. Osborn, Inhibition of dipeptidyl peptidase 8/9 impairs preadipocyte differentiation. Scientific Reports, 2015. 5: p. 12348.
63. Wolk, A., G. Gridley, M. Svensson, O. Nyrén, J.K. McLaughlin, J.F. Fraumeni, and H.-O. Adami, A prospective study of obesity and cancer risk (Sweden). Cancer Causes & Control, 2001. 12(1): p. 13-21.
64. Calle, E.E., C. Rodriguez, K. Walker-Thurmond, and M.J. Thun, Overweight, obesity, and mortality from cancer in a prospectively studied cohort of U.S. adults. New England Journal of Medicine, 2003. 348(17): p. 1625-38.
65. Sun, L., J.J. Dai, W.F. Hu, and J. Wang, Expression of toll-like receptors in hepatic cirrhosis and hepatocellular carcinoma. Genetics and Molecular Research, 2016. 15(2).
66. Mohamed, F.E., R.M. Al-Jehani, S.S. Minogue, F. Andreola, A. Winstanley, S.W. Olde Damink, A. Habtesion, M. Malagó, N. Davies, T.V. Luong, et al., Effect of toll-like receptor 7 and 9 targeted therapy to prevent the development of hepatocellular carcinoma. Liver International, 2015. 35(3): p. 1063-76.
67. Seki, E., E. Park, and J. Fujimoto, Toll-like receptor signaling in liver regeneration, fibrosis and carcinogenesis. Hepatology Research, 2011. 41(7): p. 597-610.