Profiling amino acid substitution on proteomic scale unveils novel protein regulation in cancer diseases

Peize Zhao[1,†], Shuaijian Dai[2,†], Shengzhi Lai[3,†], Chen Zhou[3], Ning Li[2,*] and Weichuan Yu[1,3,*]

[1] Individual Interdisciplinary Program,
The Hong Kong University of Science and Technology, Hong Kong, 000000, China
[2] Division of Life Science,
The Hong Kong University of Science and Technology, Hong Kong, 000000, China
[3] Department of Electronic and Computer Engineering,
The Hong Kong University of Science and Technology, Hong Kong, 000000, China
[*] Correspondence authors: N.L. (boningli@ust.hk) and W.Y. (eeyu@ust.hk)

Abstract

Amino acid substitutions (AA subs) play an important role in regulating cellular activities, encompassing complex cell signaling and cycling in both normal and cancer cells. Profiling AA subs at a proteomic level has been a challenging issue in the past. In this study, we analyze a large collection of proteomic data originating from 537 patients across five cancer patient cohorts using a novel mass spectrometry data analysis tool, named PIPI3. The pervasive patterns of AA subs are present in essential proteins like exosome-related proteins, actin filament binding proteins, and proteins in glycolysis and gluconeogenesis pathways of cancer cells. The novel identified AA subs are assumed to be caused during the protein synthesis process, which cannot be identified by traditional genome or transcriptome-based methods. Thus, this proteomic scale analysis provides insights into novel regulation mechanisms of AA subs in carcinogenesis, the outcome of which may be applied in cancer diagnosis and interventions.

1 INTRODUCTION

The utilization of proteomic datasets in tumor research has rapidly improved our comprehension on carcinogenesis and led to crucial progress in cancer interventions.[1] Recent improvements in sample processing techniques coupled with liquid chromatography-tandem mass spectrometry (LC-MS/MS) and tandem mass tag (TMT) technology[2] have made quantification analysis between control and treatment on protein levels possible and made the proteomic analysis more efficient.[3] These technological advancements empower researchers to conduct a deeper quantitative analysis of tumors and provide a more comprehensive understanding of their pathophysiology.[4] Nevertheless, large-scale multi-cancer proteomic data remain inadequately characterized, especially in identifying the multiple types of amino acid substitutions (AA subs), which alter the protein structure, interaction, and activity as well as critically underscore the multifaceted nature of cancers in terms of their molecular and phenotypic heterogeneities.[5,6]

Recent multi-cancer proteomic analyses have been focused on AA subs identification and the influence of AA subs on cancer biology. For example, W>F (tryptophan to phenylalanine) substitutions have been implicated in alternative decoding mechanisms, potentially disrupting gene function and influencing tumor immune responses.[7] Additionally, R>C (arginine to cysteine) substitutions are prominently enriched in lung cancers with mutations in the Kelch-like ECH-associated protein 1 (KEAP1).[8] However, prior studies analyzed only a small number of AA subs. Despite extensive research efforts, the precise functional implications of all types of AA subs in the context of cancer proteomic data remain unknown. The intricate interplay of multiple AA subs and the formation of regulatory networks involving them are still elusive. This knowledge gap is particularly evident when considering the landscape of AA subs in different cancer types. It is well established that AA subs can have significant effects on protein structure, function, and interactions.[9] Therefore, to gain a further understanding of the functional relevance of AA subs in cancer biology and discover novel therapeutic targets, a comprehensive analysis that integrates multi-cancer data and the landscape of all types of AA subs is needed.

Although studying AA subs in cancer proteomics is important, identifying the full range of AA subs in peptides has been an ongoing challenge. Initially, database search methods were limited to identifying peptides with pre-specified post-translational modifications (PTMs), and the number of modifications was pre-specified.[10,11] Such closed search methods can achieve high accuracy. However, this approach faces an exponential growth in database size concerning the number of pre-specified modifications. Consequently, open search methods have been developed to identify peptides with extensive modification databases, such as UNIMOD.[12] Open search methods allow for utilizing a large modification search range but accuracy is compromised. The development of open search methods in recent years has shown an increasing capacity for balancing the search space and search accuracy.[13–15] Therefore, with the help of open search methods, it is now possible to identify all types of AA subs from proteomic datasets.

In this paper, to balance the search range and accuracy of identification, we select PIPI3,[13] an open search method, as the major search engine, and we also use Comet,[11] a closed search method, to validate PIPI3's results. This study provides valuable insights into the extensive AA subs across different cancers for the first time. We uncover the AA subs regulations in exosome-related proteins, actin filament binding proteins, and proteins in glycolysis and gluconeogenesis pathways. Specially, we inspect the effect of F43S on molecular stabilization and the effect of E91D on oxygen affinity in hemoglobin subunit beta (HBB), the change of hydrophobic core by P584T in filamin-A (FLNA), and the influence of substrate binding of A175N in fructose-bisphosphate aldolase B (ALDOB). The comprehensive analysis of AA subs regulations in this study is helpful for further exploration of cancer biology.

2 RESULTS

Profiling AA subs from the cancer proteomic dataset provides new insights to analyze the relationship between crucial proteins and cancer phenotype. In the **Results** section, the proteomic analysis workflow and overview of the multi-cancer AA subs landscape are demonstrated in sub-

section **2.1**. Gene ontology (GO) enrichment analysis[16] and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways enrichment analysis[17] are applied to find crucial proteins. In subsections **2.2**, **2.3**, and **2.4**, AA subs in extracellular exosome-related proteins, actin filament binding-related proteins, and glycolysis/gluconeogenesis pathway-related proteins are investigated in detail, respectively.

2.1 Proteomic analysis workflow and multi-cancer AA subs landscape

Previous studies have found significant protein-based molecular tumor regulation and cancer-specific pathways corresponding to different AA subs.[7,8] However, these studies only focused on a small number of AA subs. The primary objective of this study is to integrate proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC)[18] cohorts and conduct an analysis of all types of AA subs across multiple cancers. The dataset employed in this paper consists of 537 patients diagnosed with five different cancers, glioblastoma (GBM),[19] head and neck squamous cell carcinoma (HNSCC),[20] lung adenocarcinoma (LUAD),[21] lung squamous cell carcinoma (LSCC),[22] and clear cell renal cell carcinoma (ccRCC)[23] (**Figure 1A**). The numbers of patients with each type of cancer are 99, 110, 110, 108, and 110 for GBM, HNSCC, LUAD, LSCC, and ccRCC, respectively. To achieve comprehensive profiling of all types of AA subs, open search methods that can search multiple modifications play a critical role. We implement rigorous and meticulous tests of the false positive control and sensitivity concerning AA subs identification to find the best open search method for the identification task. For the tests, we select three state-of-the-art open search methods for peptide identification, namely PIPI3,[13] Open-pFind,[14] and MODplus.[15] Results show that PIPI3 outperforms the other two methods, therefore, we use PIPI3 as the main search method in this study. For details of the tests, please refer to **Supplementary Notes**.

The workflow of our proteomic data search is shown in **Figure 1B**. All of the spectra from cancer datasets[18] and all types of modifications in UNIMOD[12] are used as the input to PIPI3. Using Comet, we validate the PIPI3's search results with AA subs and retain the validated results for the AA subs profiling investigation. Since Comet cannot work with a large modification search range,

we use PIPI3's identified modifications to guide Comet's search. We find the mass shifts of some AA subs to be the same as those of other modifications (**Supplementary Table S6**). Therefore, to avoid ambiguity, we exclude these AA subs from the following analysis.

Our search pipeline provides comprehensive identification results from the cancer proteomic datasets. **Figure 1C** shows the distribution of the numbers of proteins in PIPI3's identification results. Most proteins appear between 10 and 10^3 times, and the most frequently identified protein appears tens of thousands of times. This phenomenon is consistent among the results of the five cancer cohorts. Please note that the overall distribution of GBM is slightly lower than that of the other cancers because the dataset for GBM is relatively smaller. For the same reason, the number of peptide spectrum matches (PSMs) in PIPI3's results (**Figure 1D**) and the number of PSMs in the validated results (**Figure 1E**) are also lower for GBM than for the other four cancers. In PIPI3's results, after excluding the PSMs that are identified with only TMT labeling mass shift, the ratio of PSMs with modifications ranges from 54% to 58% and the ratio of PSMs that contain more than one AA sub is about 15% (**Figure 1D**). This observation further supports our selection of PIPI3 as the open search software since we do not want to forgo these 15% of PSMs with multiple AA subs.

The spectra with AA subs identified by PIPI3 are then selected and the modification information is used as guidance for the search parameters of Comet. Therefore, every validated PSM contains at least one AA sub. Of the validated PSMs, more than 25% have multiple AA subs (**Figure 1E**). Since carbamidomethyl on cysteine and TMT labeling mass shift on lysine are two fixed modifications, no AA subs can be identified on cysteine or lysine. Therefore, 18 amino acids can be substituted. Among the 18 amino acids, the proportion of each AA sub is not even. After normalization by the total number of AA subs of each cancer, glutamic acid is the most frequently substituted amino acid, with a ratio of about 15%, while tryptophan is the least substituted, with a ratio of only about 0.6% among each cancer cohort (**Figure 1F**). The search results of PIPI3 and the validated results are deposited on Zenodo[24] at https://doi.org/10.5281/zenodo.13383150.

We next perform quantification to analyze the regulation differences of AA subs containing unique

PTM site patterns (UPSPs) between cancer and normal samples (**Figure 1G**). For reference, the regulation differences of UPSPs with all modifications are shown in **Supplementary Figure S6**. The fold change threshold is set as 3 for up-regulations and 1/3 for down-regulations. More quantification descriptions and the threshold for sifting results are detailed in the **Methods** section. The quantification results show that most AA subs-modified UPSPs are not significantly up- or down-regulated. From those significantly regulated UPSPs, about 13% (191 out of 1438) of identified AA subs have been recorded in the protein variant database,[25] the rest 1247 identified AA subs are novel for further AA subs study in cancer biology (**Figure 1H, Supplementary File 1**). Meanwhile, the AA subs are unevenly identified from the proteins in the cancer datasets. The protein identified with the largest normalized ratio of AA subs is HBB, in which 104 AA subs are identified from five cohorts even though the length of HBB is only 147 (**Supplementary Figure S7**). In four cancers (GBM, HNSCC, LUAD, and LSCC), the number of up-regulations is more than the number of down-regulations, with GBM having the most up-regulated UPSPs (795). In the ccRCC quantification results, however, the quantitative relationship is the opposite, with more down-regulated UPSPs (212) than up-regulated (60). The differences in the quantification results between the five cancer datasets are obvious, and we will focus on these significantly regulated UPSPs in the following sections.

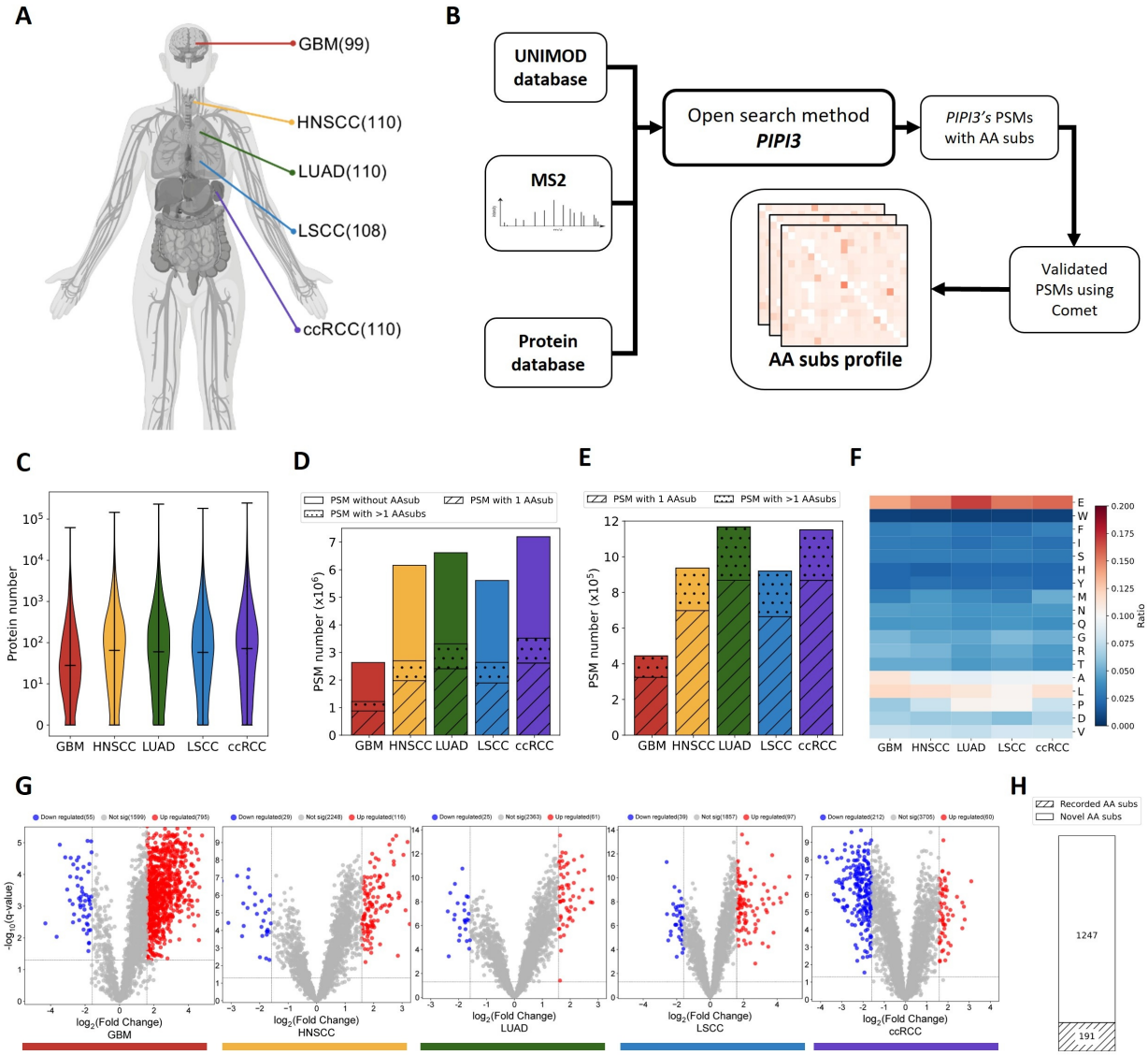
A
GBM(99)
HNSCC(110)
LUAD(110)
LSCC(108)
ccRCC(110)
B
UNIMOD database
MS2
Protein database
Open search method PIPI3
PIPI3's PSMs with AA subs
Validated PSMs using Comet
AA subs profile
C
Protein number
GBM HNSCC LUAD LSCC ccRCC
D
PSM without AAsub
PSM with 1 AAsub
PSM with >1 AAsubs
PSM number (x10^6)
GBM HNSCC LUAD LSCC ccRCC
E
PSM with 1 AAsub
PSM with >1 AAsubs
PSM number (x10^5)
GBM HNSCC LUAD LSCC ccRCC
F
Ratio
GBM HNSCC LUAD LSCC ccRCC
G
-log10(q-value)
log2(Fold Change)
GBM
HNSCC
LUAD
LSCC
ccRCC
H
Recorded AA subs
Novel AA subs
1247
191

Figure 1: Proteomic analysis workflow and multi-cancer AA subs landscape.
(A) Number of patients for each type of cancer: GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(B) The workflow of proteomic datasets analysis.
(C) Violin plot showing the PIPI3-identified protein number of each cancer cohort. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(D) The PSMs in PIPI3's identification results. Clean section: PSMs without any AA subs. Slashed section: PSMs with one AA sub. Dotted section: PSMs with more than one AA subs. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(E) Bar plot showing the number of PSMs in Comet's validated identification results. Slashed section: PSM with one AA sub. Dotted section: PSMs with more than one AA subs. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(F) Normalized ratio of substituted amino acids in each cohort.
(G) Volcano plots of quantification results based on the validated results. Down-regulated UPSPs are denoted by blue. Up-regulated UPSPs are denoted by red. Not significantly regulated UPSPs are denoted by gray.
(H) Bar plot showing 191 out of 1438 identified AA subs are recorded in the protein variant database.

2.2 AA subs in extracellular exosome-related proteins across cancer cohorts

Extracellular exosomes, also known as microvesicles, are a type of cell-excreted extracellular vesicle that is normally less than 100 nanometers in diameter. As the smallest extracellular vesicle, exosomes are crucial in the regulation of physiological processes including immune response, neuronal communication, and cancer development. Because of the endogenous nature of exosomes, they exhibit enhanced biocompatibility and stability when compared to synthetic carriers, making them valuable as delivery vehicles between cells.[26] Some researchers have articulated the relationship between microvesicles and cancers. In the cancer microenvironment, exosomes can take the role of extracellular messengers. For example, exosomes derived from GBM cells have been documented to trigger glioma cell proliferation and enhance cancer growth in a self-sustaining fashion.[27] Elevated exosome release stands out as a critical phenotype of malignant tumors when regulated by tumor extracellular acidity.[28]

In this paper, GO enrichment results show that the extracellular exosome term is enriched in all of the cohorts, both up-regulated and down-regulated (**Supplementary Figure S8**). In the cellular component enrichment results, of those component terms that are up-regulated enriched in at least

four cohorts, four out of eight terms are extracellular-related terms (**Figure 2A**, right side). The enrichment results imply a potential investigation direction examining the relationship between AA subs and the cancer extracellular environment. Besides extracellular exosomes, blood microparticles are another cellular component term that is enriched by up- or down-regulations in at least four cohorts. The numbers of exosome-related genes of each cancer are shown in **Figure 2B**. The number of only up-regulated (86) and both up- and down-regulated genes (14) in GBM outnumber those in the other four cancers, while ccRCC has the largest number of genes (58) that are only found to have down-regulated expression. In LSCC, the genes are either up-regulated or down-regulated.

The number of only up-regulated microvesicle-related proteins and the protein overlap of the five cancer cohorts is displayed in **Figure 2C**. The numbers of only up-regulated proteins are 86, 32, 36, 43, and 11 in GBM, HNSCC, LUAD, LSCC, and ccRCC, respectively. Among all the up-regulated proteins, three are commonly expressed in four out of the five cancers, glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is one of these and is expressed in the GBM, LUAD, LSCC, and ccRCC cohorts. It has been reported that GAPDH is a cytosolic protein found in exosomes and can be released by them.[29] For those significantly only down-regulated gene-expressed peptides, we select out the UPSPs whose fold change is less than 1/3 and show these UPSPs in **Figure 2D**. Among these only down-regulated results, we surprisingly find that the two lung cancer cohorts, LUAD and LSCC, significantly regulate the AA subs in HBB. Out of 23 significant down-regulations in LUAD, 19 belong to HBB, and in the LSCC cohort, this ratio is 14 out of 38. In LSCC, seven UPSPs from hemoglobin subunit alpha (HBA1) are also found to be significantly down-regulated. HBB and HBA1 are not down-regulated in the other three cancer types. These two proteins are both well-known hemoglobin subunits and are primarily responsible for transporting oxygen from the lungs to the body's peripheral organs via the blood. Our discovery of down-regulated AA subs aggregation in these two proteins in the lung cancer cohorts may provide new proteomic-based methods in lung cancer diagnosis and treatment.

In **Figure 2E**, we display the PPI diagram of the 19 exosome-related proteins that are both up- and down-regulated in GBM, HNSCC, LUAD, and ccRCC. The size of the nodes reflects the degree

of each protein. Among the 19 proteins, vimentin (VIM) is the one that is found to have the largest number of AA subs. VIM is known as one of the most abundant types in the intermediate filament (IF) protein family. The results of several studies have shown that microtubules linked with the VIM network within cells exhibit notably enhanced stability.[30,31] Studies have also shown that the VIM-containing microvesicles have an association with promoting tumor metastasis by boosting cancer cell migration and invasion, especially for GBM cancer cells, which express VIM on the cell surface.[32]

The details of the AA sub-profiles of the three prominent proteins, GAPDH, HBB, and VIM, are also demonstrated (**Figure 2F–H**). The profiles show that the AA subs in GAPDH, HBB, and VIM are identified to be only up-regulated, only down-regulated, and bidirectional-regulated, respectively. For GAPDH, we map ten AA subs in the GBM, one in the LUAD, four in the LSCC, and two in the ccRCC cohorts. The one AA sub from LUAD is identified to be the common AA sub from ccRCC at site 174, where leucine is replaced by lysine in fragment 162-VIHDNFGLVEGIMTTVHAITATQK-186, with a 3.12-fold increase in regulation for LUAD and 3.07-fold increase in regulation for ccRCC (**Supplementary File 4 and 6**). This L174K substitution could provide a new angle for analyzing the association between LUAD and ccRCC (**Figure 2F**). In GAPDH, a region from 2 to 148 interacts with WARS1, which is one of the two forms of tryptophanyl-tRNA synthetase. This interaction region of GAPDH is important for the regulation and stimulation of its aminoacylation process.[33] In this region, we locate four GBM up-regulated AA subs, at substitution sites 100, 101, 106, and 119, and two LSCC up-regulated AA subs, at substitution sites 70 and 110. These AA subs probably influence the interaction between GAPDH and WARS1 and the process of aminoacylation.

HBB is one of the proteins with the largest ratio of identified AA subs, which are all discovered from the two lung cancer cohorts and down-regulated. HBB has a globin structural domain that covers the sequence from 3 to 147 and all of the detected AA subs (**Figure 2G**). AA subs are found to occur at 17 substitution sites in LSCC and 11 substitution sites in LUAD. Among these substitution sites, sites 44 and 56 are recorded to have DNA mutation-caused AA subs,[34] the AA subs are E44K and M56I, respectively (**Supplementary File 7**). Substitution sites 69 and 118 are found to have two possible AA subs in LSCC, with site 69 detected to have L69G (0.28-fold) and L69Q (0.30-fold)

substitutions and site 118 detected to have H118N (0.20-fold) and H118E (0.25-fold) substitutions (**Supplementary File 5 and 6**). From these down-regulated AA subs, we notice several common substitutions between the two cohorts, A11Q, G17Q, and F43S. These common AA subs could be investigated to understand the shared mechanism of the two lung cancers.

Among these AA subs and substitution sites on HBB, some have been investigated by other studies. It has been reported that F43S changes the heme pocket and leads to severe hemolytic anemia.[35,36] In HBB, phenylalanine at site 43 is considered an invariant residue, and this internal hydrophobic residue plays a crucial role in maintaining the stability of the hemoglobin molecule. When a less hydrophobic residue like serine replaces this essential phenylalanine residue, it disrupts the native structure and stability of the hemoglobin molecule, leading to molecular instability. In addition, phenylalanine at site 43 is observed to be close to the porphyrin ring of the heme group, maintaining a planar orientation with it. This positioning allows phenylalanine to establish strong van der Waals interactions with the heme group, serving as a vital stabilizing force within the molecule.[36] The replacement of phenylalanine with a non-charged and polar residue serine induces a pronounced destabilization effect. This substitution further disrupts the favorable van der Waals interactions and the stabilizing effect of phenylalanine, leading to increased molecular instability and loss of stabilizing forces within the structure. Two intra-chain hydrogen bonds of F43-G47 and F43-T39 have been discovered.[37] The F43S substitution also affects these two intra-chain hydrogen bonds and intra-molecular stability. Since the porphyrin ring is critical in oxygen transportation,[38] it is hypothesized that the F43S substitution in HBB has a strong relationship with oxygen transport and oxygen use for lung cancer patients.

Another previously studied AA sub in HBB is E91D, which has been found to cause better efficiency in carrying oxygen.[39] Site 91 is situated externally and does not participate in direct contact with the heme pocket. However, it is close to the heme-bound proximal histidine at site 93 and lies between two hydrophobic residues of the heme pocket: leucine at site 89 and leucine at site 92. And, intriguingly, E91D exhibits heightened oxygen affinity, contrasting with the diminished oxygen affinities observed in the other two AA subs: E91L and E91G. Besides, it has been proposed that glutamyl residue at site 91 contributes to raising the pKa value of the histidyl residue at site 147

in the deoxygenated state.[40] This increase in pKa stabilizes the quaternary structure and enhances the oxygen affinity, suggesting a role in modulating the protein's conformational states. And the replacement of aspartic acid is presumably to reinforce the oxygen affinity.[39] Our identification of down-regulated E91D in lung cancer potentially explains the decreased oxygen affinity in lung cancer patients. The remaining untouched AA subs could serve as a valuable asset for further exploration and confirmation.

In VIM, 26 substitution sites are discovered, among which three of the substitution sites are found to be down-regulations (**Figure 2H**). Meanwhile, two up-regulated AA subs are possible at site 251, E251M (16-fold) and E251F (5.54-fold) (**Supplementary File 2**). An IF rod domain ranges from 103 to 411 in VIM, covering 66% of the length of VIM and 22 substitution sites. This IF rod domain encodes the core structure of the long and thin shape of VIM and contributes to the constitution of microtubules.[31] Within this domain is a coiled region ranging from 154–245. To investigate the influence of AA subs on the structural formation of this coiled region, we use AlphaFold3 to predict the structure of the substituted sequence from 144–251 and compare it with a crystal structure of the same length range[41] (**Figure 2I**). Six up-regulated AA subs are located in this sequence range, since the fold change of E251M is larger than that of E251F, we set the amino acid at site 251 as M in the input to AlphaFold3. The TM-score[42] of the predicted and experimental structure is 0.81, indicating a notable structural change given that 6 out of 108 amino acids are substituted. The cumulative AA subs have a chance to destroy the stability of exosomes containing microtubules and affect the tumor metastasis of GBM cancer.

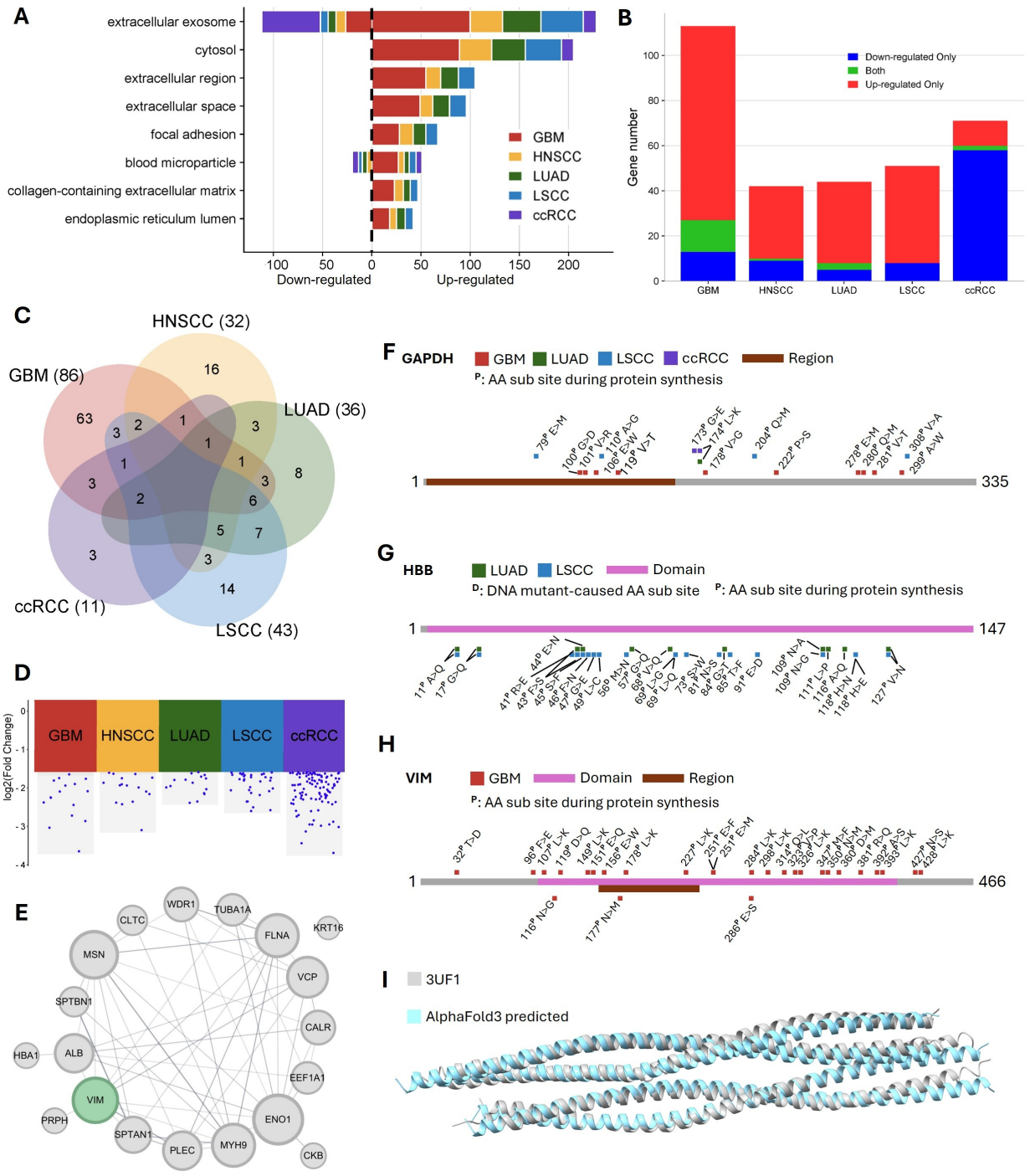
A
extracellular exosome
cytosol
extracellular region
extracellular space
focal adhesion
blood microparticle
collagen-containing extracellular matrix
endoplasmic reticulum lumen
GBM
HNSCC
LUAD
LSCC
ccRCC
Down-regulated
Up-regulated
B
Gene number
Down-regulated Only
Both
Up-regulated Only
GBM
HNSCC
LUAD
LSCC
ccRCC
C
HNSCC (32)
GBM (86)
LUAD (36)
ccRCC (11)
LSCC (43)
D
log2(Fold Change)
GBM
HNSCC
LUAD
LSCC
ccRCC
E
WDR1
TUBA1A
CLTC
FLNA
KRT16
MSN
VCP
SPTBN1
CALR
HBA1
ALB
EEF1A1
VIM
ENO1
PRPH
SPTAN1
PLEC
MYH9
CKB
F
GAPDH
GBM
LUAD
LSCC
ccRCC
Region
P: AA sub site during protein synthesis
G
HBB
LUAD
LSCC
Domain
D: DNA mutant-caused AA sub site
P: AA sub site during protein synthesis
H
VIM
GBM
Domain
Region
P: AA sub site during protein synthesis
I
3UF1
AlphaFold3 predicted

Figure 2: AA subs in extracellular exosome-related proteins across cancer cohorts.
(A) Bar plot showing the cellular component terms of GO enrichment that are expressed in at least four cancer cohorts. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(B) Bar plot representing the number of genes that are related to extracellular exosomes among the five types of cancers. The only up-regulated genes are denoted by the red segment, the only down-regulated genes are denoted by the blue segment, and both up- and down-regulated genes are denoted by the green segment.
(C) Gene overlap of only up-regulated genes. The numbers of only up-regulated genes in each cohort are noted. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(D) Volcano plot presenting the gene expression of UPSPs that are only down-regulated. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), LSCC (blue), and ccRCC (purple).
(E) Diagram representing a network of 19 proteins corresponding to extracellular exosomes that are both up- and down-regulated. The size of the node represents the degree of the protein.
(F) Cumulative AA subs map of GAPDH across different cancer cohorts. The AA subs above the protein sequence are up-regulated. Brown segment: binding and interaction region. ^{P}: AA sub site during protein synthesis. The cohorts are GBM (dark red), LUAD (green), LSCC (blue), and ccRCC (purple).
(G) Cumulative AA subs map of HBB in two types of lung cancers. The AA subs below the protein sequence are down-regulated. Pink segment: structural domain. ^{D}: DNA mutant-caused AA sub site. ^{P}: AA sub site during protein synthesis. The cohorts are LUAD (green) and LSCC (blue).
(H) Cumulative AA subs map of VIM in GBM. The AA subs above the protein sequence are up-regulated, and those below the protein sequence are down-regulated. Brown segment: binding and interaction region. Pink segment: structural domain. ^{P}: AA sub site during protein synthesis. The cohort is GBM (dark red).
(I) Structure comparison between the crystal structure of human vimentin (fragment 144–251) (PDB: 3UF1) and structure prediction with cumulative AA subs. Gray: the structure of 3UF1. Light blue: AlphaFold3-predicted structure with cumulative AA subs in fragments 144–251.

2.3 AA subs analysis of actin filament binding

Actin is the major compositor of cytoskeletal protein in most cells. The polymerization of actin forms the actin filament (microfilament), which is seven nanometers in diameter and several micrometers in length.[43] In cells, the microfilaments are highly organized and form structures of semisolid gels. Various actin-binding proteins rule the assembly and disassembly of actin filaments and are, therefore, important to the actin skeleton. It has been widely found that actin-binding proteins have associations with cancer progress and pave the way to possible therapeutic methods.[44–46]

In our study, GO enrichment analysis shows that, based on the up-regulated UPSPs in the five

cancer cohorts, the proteins are most abundantly enriched in the functions of protein binding and RNA binding. Among the five cancers, four types of cancers except for ccRCC are functionally enriched in actin filament binding. In the functional enrichment results with more than five enriched genes, 13 out of 15 in GBM, 3 out of 6 in HNSCC, 6 out of 7 in LUAD, 5 out of 7 in LSCC, and 1 out of 1 in ccRCC enriched functions are shown to be related to the binding function (**Figure 3A**). The AA subs can alter the primary structure of involved proteins, thereby changing the protein fold and intra-molecular bonds of the amino acid chain. These changes work together to influence the binding sites and functions.[47] This observation deepens the motivation of our AA subs investigation. The GO molecular function enrichment results based on the down-regulated UPSPs are shown in **Supplementary Figure S9**.

The number of actin filament binding-related genes of each cancer cohort and the gene overlap is displayed in **Figure 3B**, where the colors correspond to four of the cancer types, excepting ccRCC. Among the four cancers, 23 proteins are actin filament binding-related and GBM contributes 17 proteins. To demonstrate the PPI relationships between the microfilament binding proteins that are expressed and identified in the cancer cohorts, we show the 23 proteins in **Figure 3C**, where the size reflects the degree of each protein. The three proteins that are up-regulated in all four cancers, namely spectrin alpha chain, non-erythrocytic 1 (SPTAN1), plectin (PLEC), and FLNA, are marked in red. We count the number of significantly regulated UPSPs of the three commonly expressed proteins among the four types of cancers as shown in **Figure 3D–F**. We can observe that the regulated UPSPs of these proteins are more abundant in GBM than in the other three cancers. And GBM is the only cohort that has both up- and down-regulated UPSPs. Notably, the number of GBM-regulated UPSPs in FLNA is 19, where 18 are up-regulated and one is down-regulated, while each of the other three cohorts has merely one up-regulation (**Figure 3F**).

Spectrins represent a group of filament cytoskeletal proteins, which serve as crucial scaffold proteins, playing a vital role in supporting the plasma membrane and arranging internal organelles. Spectrins consist of α and β chains that come together to create tetramers arranged in a head-to-head style. SPTAN1 encodes the α chain of spectrin and is exclusively found in non-erythrocytic cells.[48] With its diverse functions encompassing adhesion and migration, SPTAN1 has both pos-

itive and negative impacts on tumor development and progression, depending on the regulation situation of SPTAN1.[49] In the AA subs analysis results of SPTAN1, we find four AA subs in GBM, two in HNSCC, one in LUAD, and two in LSCC. Among the four AA subs in GBM, three AA subs are discovered to be up-regulated and one AA sub to be down-regulated, while the AA subs for the other three cancers are all up-regulated. Additionally, two common AA subs, I639S substitution and Y644F substitution, are identified among GBM and HNSCC in segment 638-LIDVNHYAKDEVAAR-652 (**Figure 3G, Supplementary File 2 and 3**). The UPSPs with common AA subs show a 3.56-fold and 3.07-fold increase in abundance in GBM and HNSCC, respectively.

PLEC, a long-sequence protein present in almost all mammalian cells, serves as a crucial connection point among the primary components of the cytoskeleton proteins, including actin microfilaments.[50] In PLEC, more AA subs are identified from the four cancer cohorts. We point out eight up-regulated AA subs in GBM, two in HNSCC, one in LUAD, and four in LSCC, and only one down-regulated AA sub, F758S in GBM. Among these AA subs sites, sites 2525, 2836, and 3484 are three sites recorded to have AA subs caused by DNA or mRNA mutation. R2525C is found to be caused by DNA mutation, while A2836T and V3484M are caused by both DNA and mRNA mutation, according to OncoDB[34] (**Supplementary File 7**). Two binding and interaction regions in the PLEC sequence cover most of the identified AA subs (**Figure 3H**). The first region ranges from 175–400, which is an actin-binding region,[51] and in the LSCC cohort we find an F330L substitution, which replaces an aromatic amino acid with a non-aromatic amino acid, in this region. The second region includes the sequence of 964–4574, which covers about 77% of PLEC's length. This region has been previously discovered to interact with IF proteins.[52] It includes seven AA subs in GBM, two AA subs in HNSCC, and one in LSCC, and all of the covered AA subs are up-regulated substitutions. The interaction between PLEC and IF proteins might be influenced by the AA subs crowding in this region.

FLNA is closely involved with the proliferation and invasion of GBM,[53,54] and has been studied together with the mechanistic target of rapamycin complex two activity as a signaling pathway in the motility of GBM cells.[55] Meanwhile, FLNA is an essential protein in linking actin filaments

to integrate molecules within cell membranes, crucially contributing to maintaining cell shape and signaling processes.[56] Probing the AA subs in GBM and GBM-regulated UPSPs potentially creates a new angle to uncover the mechanism of migration and invasion of GBM cells. The number of identified AA subs sites of FLNA in the GBM cohort is 21, of which two are down-regulated. Among these substitution sites G776V, E798K, and V1480M have been recorded as DNA mutation-caused AA subs, while A366T and G776S have been recorded to be caused by both DNA mutation and mRNA mutation (**Supplementary File 7**). The rest of the identified AA subs should take place after protein translation. Interestingly, substitution sites 1804 and 2544 are identified to have two possible AA subs in GBM (**Figure 3I**). Site 1804 is resident in segment 1802-GEITGEVR-1809, with a 4.50-fold increase in abundance for the I1804S substitution and 4.29-fold for the I1804D substitution. Meanwhile, site 2544 is in segment 2541-ATCAPQHGAPGPGPADASK-2559, the A2544W substitution has a 3.97-fold up-regulation and the A2544S substitution has a 5.98-fold up-regulation (**Supplementary File 2**). The number of substitutions for the other three cohorts is just one, which is compatible with **Figure 3F**. Some AA subs are located in the regions of binding and interaction. In the actin-binding region from 2 to 274, there is a V122A substitution. In the region of interaction with furin from sequence 1490 to 1607, which UniProt annotates,[25] two AA subs, L1493K and H1543Y, are identified.

The FLNA dimer structure has 24 immunoglobulin-like (Ig) domains, among which domain 4 is stabilized by interaction with domain 5 and the interface between domain 4 and domain 5 is governed by several hydrophobic β-sheet interactions.[57] Additionally, two amino acid substitutions in domains 4 and 5, P637E and V711D, respectively, have been previously reported to cause familial cardiac valvular dystrophy because these two substitutions likely change the hydrophobic core and destabilize the individual β-sheet domain formulation.[58] Notably, we find P584T in domain 4 of FLNA, which alters a hydrophobic amino acid to become hydrophilic, with a 4-fold increase in abundance (**Supplementary File 2**). Although the amino acid is not situated at the domain-domain interface, the side chain is oriented towards the hydrophobic core formed by the β-sheets (**Figure 3J**). Consequently, we hypothesize that substituting proline with threonine would significantly alter the hydrophobic core, potentially destabilizing the fold of the individual β-sheet domain. Further, the β-sheet augmentation interaction between FLNA and integrin is disordered

due to the destabilized domain 4. It has been reported that integrin is activated to propel the cells after being released by FLNA,[59] therefore, this substitution could explain the enhancing function of mutated FLNA on proliferation and invasion of GBM. To compare the structure change before and after the P584T substitution, we use AlphaFold3[60] to predict the structure with the substitution and compare it with the crystal structure of FLNA domains 3–5 (**Figure 3J**).[57]

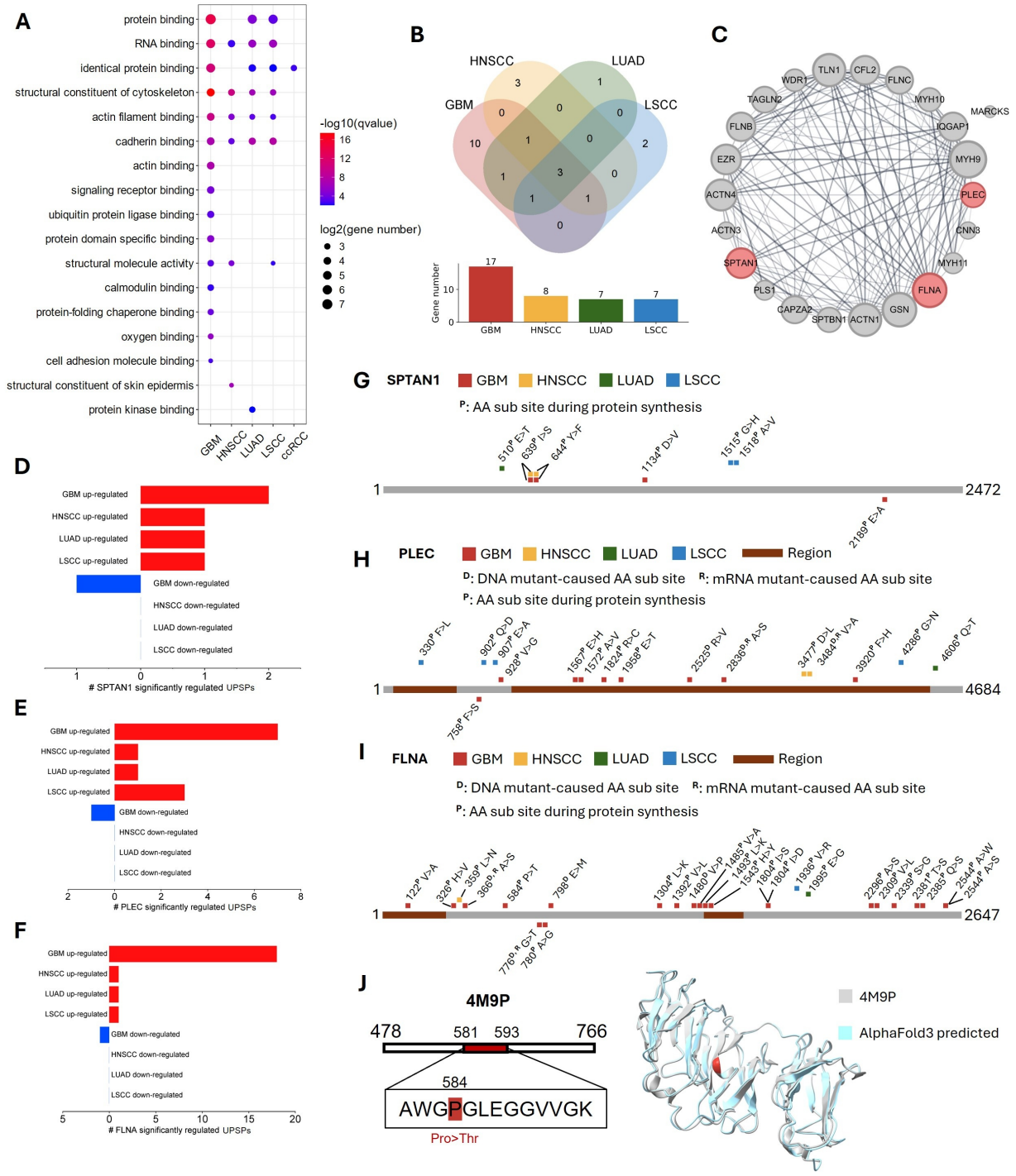
A
protein binding
RNA binding
identical protein binding
structural constituent of cytoskeleton
actin filament binding
cadherin binding
actin binding
signaling receptor binding
ubiquitin protein ligase binding
protein domain specific binding
structural molecule activity
calmodulin binding
protein-folding chaperone binding
oxygen binding
cell adhesion molecule binding
structural constituent of skin epidermis
protein kinase binding
GBM
HNSCC
LUAD
LSCC
ccRCC
-log10(qvalue)
16
12
8
4
log2(gene number)
3
4
5
6
7
B
HNSCC
LUAD
GBM
LSCC
Gene number
17
8
7
7
C
TLN1
CFL2
WDR1
FLNC
TAGLN2
MYH10
MARCKS
FLNB
IQGAP1
EZR
MYH9
ACTN4
PLEC
ACTN3
CNN3
SPTAN1
MYH11
PLS1
FLNA
CAPZA2
GSN
SPTBN1
ACTN1
D
GBM up-regulated
HNSCC up-regulated
LUAD up-regulated
LSCC up-regulated
GBM down-regulated
HNSCC down-regulated
LUAD down-regulated
LSCC down-regulated
SPTAN1 significantly regulated UPSPs
E
PLEC significantly regulated UPSPs
F
FLNA significantly regulated UPSPs
G
SPTAN1
GBM
HNSCC
LUAD
LSCC
P: AA sub site during protein synthesis
510P E>T
639P I>S
644P Y>F
1134P D>V
1515P G>H
1518P A>V
2189P E>A
1
2472
H
PLEC
Region
D: DNA mutant-caused AA sub site
R: mRNA mutant-caused AA sub site
P: AA sub site during protein synthesis
330P F>L
902P Q>D
907P E>A
928P V>G
1567P E>H
1572P A>V
1824P R>C
1958P E>T
2525P R>V
2836D,R A>S
3477P D>L
3484D,R V>A
3920P F>H
4286P G>N
4606P Q>T
758P F>S
4684
I
FLNA
122P V>A
326P H>V
359P L>N
366D,R A>S
584P P>T
798P E>M
1304P L>K
1392P V>L
1480P V>P
1485P V>A
1493P L>K
1543P H>Y
1804P I>S
1804P I>D
1936P V>R
1995P E>G
2298P A>S
2309P V>L
2339P S>G
2391P T>S
2385P Q>S
2544P A>W
2544P A>S
778D,R G>T
780P A>G
2647
J
4M9P
478
581
593
766
584
AWGPGLEGGVVGK
Pro>Thr
4M9P
AlphaFold3 predicted

Figure 3: AA subs analysis of actin filament binding.
(A) Bubble plot showing the molecular function terms of GO enrichment based on the up-regulations in the five cancer cohorts. The size of the bubble represents the base two logarithm of the enriched gene number, and the color of the bubble shows the minus base 10 logarithm of the q-value.
(B) Venn plot (top) and bar plot (bottom) demonstrating the number of genes related to actin filament binding in four cohorts of cancers. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), and LSCC (blue).
(C) Diagram representing a network of 23 proteins corresponding to actin filament binding that are significantly up-regulated. The size of the node represents the degree of the protein.
(D) Barplot showing the number of SPTAN1 significantly regulated UPSPs. Red denotes the up-regulations, and blue denotes the down-regulations.
(E) Barplot showing the number of PLEC significantly regulated UPSPs. Red denotes the up-regulations, and blue denotes the down-regulations.
(F) Barplot showing the number of FLNA significantly regulated UPSPs. Red denotes the up-regulations, and blue denotes the down-regulations.
(G) Cumulative AA subs map of PLEC across different cancer cohorts. The AA subs above the protein sequence are up-regulated, and those below the protein sequence are down-regulated. ^{P}: AA sub site during protein synthesis. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), and LSCC (blue).
(H) Cumulative AA subs map of PLEC across different cancer cohorts. The AA subs above the protein sequence are up-regulated, and those below the protein sequence are down-regulated. Brown segment: binding and interaction region. ^{D}: DNA mutant-caused AA sub site. ^{R}: mRNA mutant-caused AA sub site. ^{P}: AA sub site during protein synthesis. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), and LSCC (blue).
(I) Cumulative AA subs map of FLNA across different cancer cohorts. The AA subs above the protein sequence are up-regulated, and those below the protein sequence are down-regulated. Brown segment: binding and interaction region. ^{D}: DNA mutant-caused AA sub site. ^{R}: mRNA mutant-caused AA sub site. ^{P}: AA sub site during protein synthesis. The cohorts are GBM (dark red), HNSCC (yellow), LUAD (green), and LSCC (blue).
(J) Structure comparison between the crystal structure of FLNA Ig-like domains 3–5 (PDB: 4M9P) and structure prediction with AA sub of P584T from the GBM cohort. Left: the site of proline to threonine (P584T) substitution in the protein sequence. Right: the structure of 4M9P (gray) and AlphaFold3-predicted structure after P584T substitution (light blue).

2.4 Cancer-regulated AA subs expression in glycolysis and gluconeogenesis

Glycolysis and gluconeogenesis are two reciprocally regulated pathways. Glucose is broken down to pyruvate in the glycolysis pathway and a little adenosine triphosphate (ATP) is made in this process. Gluconeogenesis is similar to the reverse process of glycolysis, producing glucose de novo from precursors, including pyruvate or alanine. The pathway of gluconeogenesis is closed when

glycolysis is stimulated and vice versa.[61] It has been revealed that many cancer cells take glycolysis as an essential source of energy, because of metabolic and mitochondrial imperfection. Additionally, the speed of ATP production by glycolysis is faster than mitochondrial oxidative phosphorylation. Therefore, it is more suitable for the rapid energy demands of cancer cell proliferation.[62] At the same time, the simultaneous activation or suppression of glycolysis and gluconeogenesis could result in metabolic stress and lead to metabolic interference.[63] Therefore, the study of these pathways could provide therapeutic approaches to cancers due to their pivotal role in cancer cell proliferation and viability. KEGG pathways enrichment[17] among the five cancer cohorts shows that the AA subs-containing proteins are enriched in the pathways of glycolysis and gluconeogenesis. For the up-regulations, four of the cancer cohorts, excepting HNSCC, are enriched in these pathways (**Figure 4A**, indicated by the arrow), while in the down-regulations, the results of ccRCC are enriched (**Figure 4B**, indicated by the arrow).

The numbers of glycolysis and gluconeogenesis pathways-related genes are displayed in **Figure 4C**. Among the four cohorts, GBM has seven up-regulated genes, which is more than the other cancer types, while ccRCC has 12 down-regulated genes and one bidirectionally regulated gene, and is the only cancer that has down-regulated AA subs. The regulation situation of the significantly regulated UPSPs of the four cohorts is demonstrated in **Figure 4D**. We observe that the L-lactate dehydrogenase B chain (LDHB) is mutually expressed by the four cancers (GBM, LUAD, LSCC, and ccRCC) in our analysis results. In the glycolysis pathway, LDHB holds a vital position because it is a crucial glycolytic enzyme responsible for catalyzing the conversion of pyruvate to lactate and the backward conversion.[64] And since the enzymatic activity and direction of LDHB dictate the pace of glycolysis, LDHB is under active regulation in cancer cells.[65] While the critical involvement of LDHB in the advancement of different cancers has been increasingly documented,[66,67] the multi-cancer AA subs analysis of LDHB remains inadequate.

In this paper, we display the PPI diagram of the 13 down-regulated genes of ccRCC in **Figure 4E**. Of these proteins, ALDOB is the one that is modified by the largest number of AA subs. ALDOB is a type of glycolytic enzyme that facilitates the transformation of fructose 1,6-diphosphate into glyceraldehyde 3-phosphate and dihydroxyacetone phosphate within the glycolysis metabolic

pathway.[68] As a possible prediagnosis indicator for ccRCC patients, the down-regulated expression of ALDOB has been strongly linked to clinicopathological characteristics of ccRCC patients.[69] The literature validates the findings in this study of down-regulated AA subs in ALDOB from the ccRCC cohort.

We continue to analyze the profile of the AA subs for the AA subs-modified, glycolysis and gluconeogenesis pathways-enriched proteins LDHB and ALDOB. Among the identified AA subs and substitution sites in LDHB and ALDOB, some AA subs sites have been found to have DNA mutation-caused AA subs.[34] Specifically, in LDHB, the DNA mutation-caused AA sub is V136M, and in ALDOB, the DNA mutation-caused AA subs are P231R and A239S (**Supplementary File 7**). The up-regulated AA subs identified in the four cohorts are annotated in **Figure 4F**, showing two AA subs for each cancer cohort. In the two lung cancers, LUAD and LSCC, two common AA subs, V136L and Y146H are found, and the UPSP in LUAD is 4.92-fold, in LSCC is 3.29-fold (**Supplementary File 4 and 5**). In ccRCC, a V136L substitution is also detected, and another AA sub is T148S with a 3.07-fold regulation (**Supplementary File 6**). The AA subs of LUAD, LSCC, and ccRCC all occur in segment 134-LLVVSNPVDLLTYVTWK-150 (**Supplementary File 4–6**). In the GBM cohort, Q31M, and G42A are discovered in PSM 24-ITVVGVGQVGMACAISILGK-43 with an up-regulation of 4.29-fold (**Supplementary File 2**). The two AA subs of LDHB in the GBM results reside in the binding region of the oxidized form of nicotinamide adenine dinucleotide (NAD^+), whose sequence ranges from 31–53. LDHB can catalyze the reversible transformation of lactate and NAD^+ into pyruvate and NADH (H for hydrogen).[70] The two AA subs could interfere with the conversion process and the glycolysis and gluconeogenesis pathways and affect the energy generation in cancer cells.

Another frequently substituted protein is ALDOB, for which all of the AA subs are from the renal cancer cohort and are down-regulated. The down-regulated expression of ALDOB has been investigated and found to be involved with pathological features of ccRCC patients, as previously mentioned.[69] In ALDOB, 14 substitution sites are detected to be substitution sites (**Figure 4G**), with site 190 identified to have two AA subs, E190F (0.23-fold) and E190M (0.27-fold) (**Supplementary File 6**). In this paper, we identify an A175N substitution. Site 175 has been reported as frequently

substituted,[71–73] and the alanine at site 175 is conserved across all aldolase isozymes. From a crystal structure,[74] the alanine 175 residue is situated within an α-helical region close to the substrate-binding residues. Substituting alanine 175 alters the charge of amino acid and potentially disrupts the spatial arrangement of neighboring residues on the β-strand of the ALDOB substrate-binding pocket for ATP generation. Besides, replacement by asparagine alters the intra-chain hydrogen bond between site 175 and site 179 and converts the hydrophilicity of site 175,[75] therefore influencing the molecular stability. These confounding changes possibly affect the glycolysis and gluconeogenesis pathways, thereby, the energy generation of cancer cells.

Recently, the experimental structure of human kidney ALDOB was proposed (PDB: 8D44). To exhibit the influence of the AA subs in this study on protein structure, we present a structural comparison between the experimental structure and the AlphaFold3-predicted structure with cumulative AA sub modifications (**Figure 4G**). The TM-score between the two structures is 0.53. Molecular analysis of these newly discovered AA subs and function analysis of the AA subs-modified ALDOB remain to be performed.

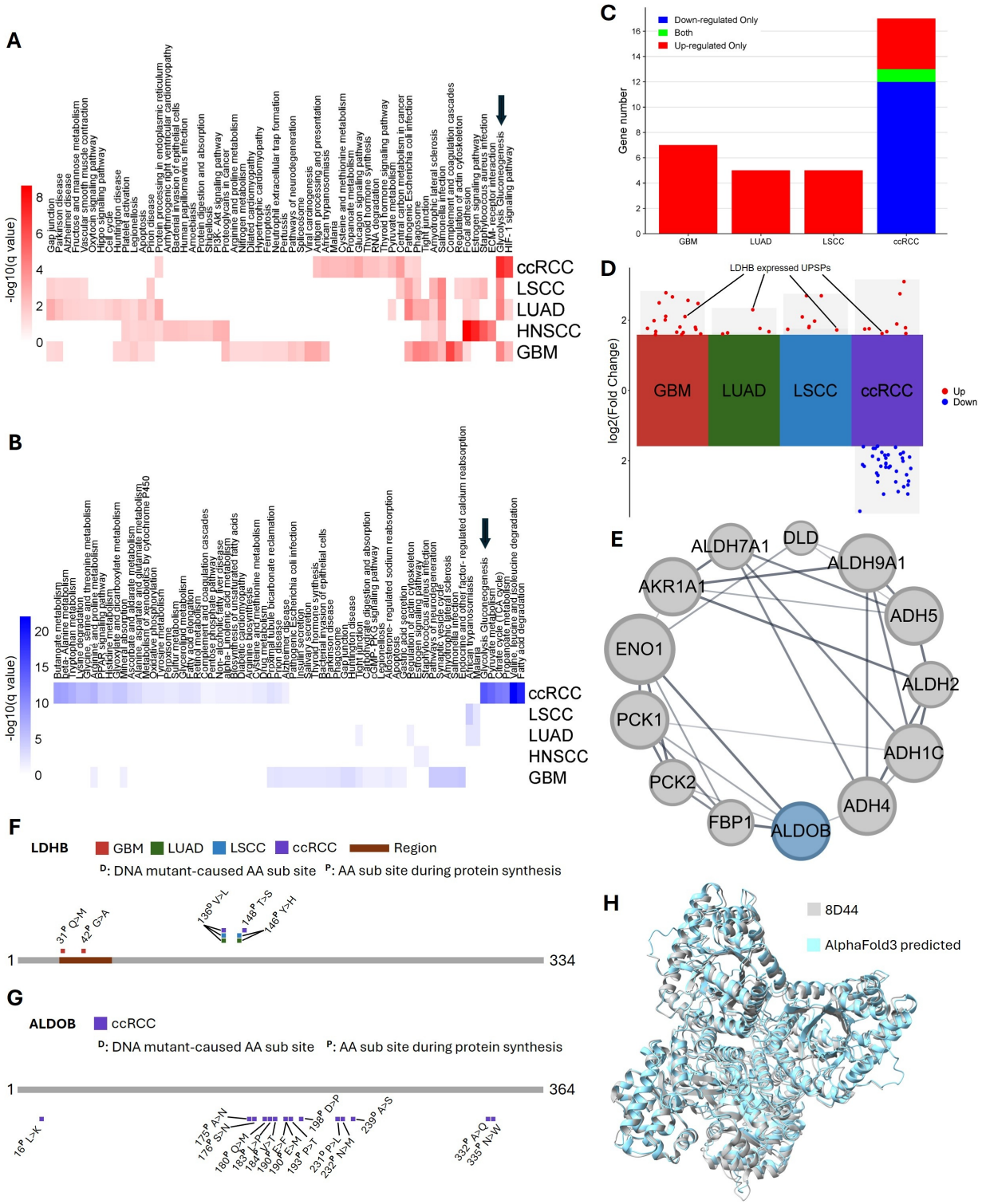
A
-log10(q value)
ccRCC
LSCC
LUAD
HNSCC
GBM
B
-log10(q value)
ccRCC
LSCC
LUAD
HNSCC
GBM
C
Down-regulated Only
Both
Up-regulated Only
Gene number
GBM
LUAD
LSCC
ccRCC
D
LDHB expressed UPSPs
log2(Fold Change)
GBM
LUAD
LSCC
ccRCC
Up
Down
E
ALDH7A1
DLD
ALDH9A1
AKR1A1
ADH5
ENO1
ALDH2
PCK1
ADH1C
PCK2
ADH4
FBP1
ALDOB
F
LDHB
GBM
LUAD
LSCC
ccRCC
Region
D: DNA mutant-caused AA sub site
P: AA sub site during protein synthesis
1
334
G
ALDOB
ccRCC
D: DNA mutant-caused AA sub site
P: AA sub site during protein synthesis
1
364
H
8D44
AlphaFold3 predicted

Figure 4: Cancer-regulated AA subs expression in glycolysis and gluconeogenesis.
(A) Heatmaps showing KEGG enrichment results of up-regulations among five cancer cohorts. The color represents the minus base 10 logarithm of the q-value; the darker the red, the smaller the q-value.
(B) Heatmaps showing KEGG enrichment results of down-regulations among five cancer cohorts. The color represents the minus base 10 logarithm of the q-value; the darker the blue, the smaller the q-value.
(C) Bar plot showing the number of genes related to glycolysis and gluconeogenesis among four types of cancers. The up-regulated-only genes are denoted by the red segment, the down-regulated-only genes are denoted by the blue segment, and the both up- and down-regulated genes are denoted by the green segment.
(D) Volcano plot presenting the gene expression of glycolysis and gluconeogenesis corresponding UPSPs. The cohorts are GBM (dark red), LUAD (green), LSCC (blue), and ccRCC (purple).
(E) Diagram representing a network of 13 proteins corresponding to glycolysis and gluconeogenesis that are down-regulated in the ccRCC cohort. The size of the node represents the degree of the protein.
(F) Cumulative AA subs map of LDHB in four types of cancers. The AA subs above the protein sequence are up-regulated. Brown segment: binding and interaction region. ^{D}: DNA mutant-caused AA sub site. ^{P}: AA sub site during protein synthesis. The cohorts are GBM (dark red), LUAD (green), LSCC (blue), and ccRCC (purple).
(G) Cumulative AA subs map of ALDOB in ccRCC. The AA subs below the protein sequence are down-regulated. ^{D}: DNA mutant-caused AA sub site. ^{P}: AA sub site during protein synthesis. The cohort is ccRCC (purple).
(H) The structure comparison between cryo-electron microscopy structure of human renal ALDOB (PDB: 8D44) and structure prediction with cumulative AA subs from ccRCC. Gray: the structure of 8D44. Light blue: AlphaFold3-predicted structure with cumulative AA subs. Purple: antibody binding sequence with cumulative AA subs.

3 METHODS

3.1 Proteomic datasets of five cancer cohorts and data preprocessing

In this study, proteomic data generated by CPTAC were downloaded from the Proteomic Data Commons.[76] The exploited data included 537 patients from five cancer cohorts: 11 TMT-11 plex experiments from 99 patients with GBM,[19] 20 TMT-11 plex experiments from 110 patients with HNSCC,[20] 25 TMT-10 plex experiments from 110 patients with LUAD,[21] 22 TMT-11 plex experiments from 108 patients with LSCC,[22] and 23 TMT-10 plex experiments from 110 patients with ccRCC.[23] All of the non-empty MS2 spectra were extracted and converted to mgf files using ProteomeWizard MSConvert.[77]

3.2 Analysis of proteomic data

The human protein FASTA file was downloaded from UniProt[25] (downloaded January 2024). PIPI3 was used as the open search method with the following parameters: the MS1 tolerance was 10 ppm; the MS2 tolerance was 0.01 Da; the max missed cleavage sites was 2; the enzyme was Trypsin; the peptide length was from 6 to 50; the fixed modifications were carbamidomethyl on cysteine and TMT labeling mass shift on lysine; modifications search range was ±250 Da; the variable modifications were from the UNIMOD,[12] including TMT labeling mass shift on serine and n-term and all types of AA subs; isoleucine and leucine were treated as identical. Decoy proteins were added for FDR control with the target-decoy strategy. The remaining settings were kept as the default. Spectra were selected for the Comet search if PIPI3 identified more than two peptides with AA subs in them. The variable modifications for the Comet search were determined by PIPI3's search results for the spectra, and other search parameters of Comet were set the same as those of PIPI3. The validated PSMs, which were the same in the identification results of PIPI3 and Comet, were kept for further bioinformatic analysis.

3.3 Quantification details

In the quantification step, we only kept the PSMs that were detected with TMT labeling mass shift. And the peptides with multiple PSMs were kept as quantifiable peptides. The intensities of reporter ions for each PSM were obtained from the initial MS2 spectra. Every sample was assumed to encompass a collection of unregulated peptides that had comparable abundances to those of the common reference (CR) sample. Therefore, in the normalized sample, these unregulated peptides should demonstrate a log TMT ratio that is centered to zero. As a result, we implemented a two-component normalization operation for the intensities of the TMT labels from the CR intensity.[22]

Log ratios at the PSM level were computed as the average log values across patients within the corresponding PSM. Peptides sharing identical modification patterns were clustered into a UPSP. The average value of all PSM-level TMT ratios in each UPSP was used as the UPSP ratio. Each

cancer cohort was composed of a set of distinct experiments. Therefore, the distinct average log ratios for each UPSP in each experiment were calculated. And we only selected those UPSPs that appeared in over 3/5 of the experiments for further analysis and quantification. One-sample t-test and BH-FDR correction were employed to determine the p-value and q-value for each UPSP, respectively.

3.4 GO and KEGG enrichment analysis

Proteins that had quantified (in quantification results, q-value<0.05 and fold change>3 or fold change<1/3) UPSPs with AA subs modifications were subjected to GO enrichment and KEGG enrichment analyses where the background was the human genome database. Two-tailed Fisher exact test and BH-FDR correction were used to measure the statistical significance of the enrichment analyses. The visualizations in this study only included the enrichment terms with q-values less than 0.01 in GO and KEGG analysis.

The script for GO enrichment analysis was developed using GOATOOLS[78] by Python, while the KEGG pathway enrichment analysis was conducted using GSEApy[79] by Python.

3.5 Terminology and construction of protein-protein interaction network

We incorporated PPI data sourced from the STRING database,[80] which aggregates interaction information from various sources such as text mining, experiments, databases, co-expression, neighborhood, gene fusion, and co-occurrence. The PPI network was visualized using Cytoscape (version 3.10.2).[81] In this network representation, each pair of interacting proteins and the link between them were respectively denoted as nodes and edges.

3.6 Protein structure modeling, comparing and measuring

The protein structures were aligned by the Matchmaker function of ChimeraX (version 1.4).[82] The structures were modeled by ChimeraX. The structure of FLNA Ig-like domains 3–5 was obtained by X-ray diffraction from Escherichia coli expression (PDB: 4M9P). The crystal structure of vimentin (fragment 144–251) was from humans (PDB: 3UF1). The structure of human kidney fructose-bisphosphate ALDOB was obtained by cryo-electron microscopy (PDB: 8D44). AlphaFold3 built the predicted structures with cumulative AA subs with its web service (https://www.alphafoldserver.com/). The provided sequences were the original sequences, in which the "I" was not replaced by "L", and the amino acids were substituted as the identified AA subs. When there was more than one possible AA sub identified on the same site, we preferred the AA sub with the higher regulation fold change. The predicted structure with the best pLDDT score was used for structure comparison and similarity measurement. The similarity metric between the experimental structure and the predicted structure was the TM-score.[42]

4 DISCUSSION

In this study, AA sub-profiling is performed based on the mass spectrometry-based proteomic search results. CPTAC provides proteomic datasets from five cancer types: GBM, HNSCC, LUAD, LSCC, and ccRCC. These five cancers are representative and located in different human body tissues. Therefore, the proteomic-based analysis of these five cancers helps explore pervasive AA sub patterns among different types of cancers. As for the proteomic search method, we implement a novel search pipeline in this paper, which balances the search range and search accuracy. To the best of our knowledge, empowered by PIPI3, this pipeline is the first one to cover all types of AA subs without the numerical limitation of modifications in each PSM. Although PIPI3's performance surpasses all other competitors through several rigorous tests, especially the test of identifying multiple AA subs in one peptide, the search results of PIPI3 are not guaranteed to be error-free. Therefore, to make the bioinformatic results as correct as possible, we use Comet as the

validation step, following the PIPI3 search step. By incorporating the two search methods in our pipeline, the search range and search accuracy of identification results are balanced.

The protein level regulations of AA subs are investigated based on the quantification results. We compare the significantly regulated AA subs with the AA subs that are recorded in the protein variant database of UniProt.[25] This comparison shows that the AA subs analyzed in this research are not arbitrarily determined, since there are 13% overlapping with prior studies (**Supplementary File 1**). Besides, to investigate the source of the AA subs in this research, we compare the significantly regulated AA subs with the AA subs database of OncoDB,[34] which is an oncology database of DNA or mRNA mutant-cause AA subs. The comparison shows that most of the discovered AA subs in this research are not deposited in the AA subs database of OncoDB (**Supplementary File 7**), which indicates that most of these significantly regulated AA subs are caused during the protein synthesis process. AA subs are widely recognized for their significant impact on the exploration of biological functions and pathways. Typically, these substitutions are thought to originate from three primary pathways: DNA mutations, mRNA mutations, and AA subs occurring during the protein synthesis process.[34] Constrained by the proteomic-based identification methods, notably the absence of open search methods capable of detecting modifications on a broad scale. Conventional high-throughput approaches rooted in genomics and transcriptomics primarily capture amino acid mutations stemming from DNA and mRNA mutations. Consequently, mutations occurring during protein synthesis have not been thoroughly scrutinized. Nevertheless, in this article, we introduce a robust proteomics-based pipeline that enables the identification of AA subs occurring during protein synthesis in a high-throughput manner. This pipeline holds significant promise for broader application across diverse disease datasets, offering valuable insights to the community in exploring not only cancer but also various other diseases.

In this study, we select several crucial proteins with multiple significantly regulated AA subs. These proteins are all closely involved with cancer biology. However, restricted by the accessibility of patients and wet lab experiment facilities, we cannot finish the follow-up wet lab experiments to validate our discoveries. Besides, there are hundreds of types of AA subs, and each amino acid in proteins has a chance to be substituted by other amino acids. Therefore, the combinatorial

scenarios of AA subs in these proteins are out of number, and there is not enough literature to support every discovered AA sub in this study. Limited by these factors, it is hard for us to determine whether these significantly regulated AA subs are the cause of cancer or the effect of cancer. This would be our future research direction. Moreover, the mass spectrometer can only take a snapshot of integral proteins in the input sample at a specific time point. As a result of the characteristics of mass spectrometry, all AA subs identified in this study represent average states on proteins, implying that these substitutions may not necessarily be localized on the same protein molecule. Instead, these substitutions are probably dispersed across multiple protein molecules. However, with current analytical methods, accurately distinguishing these individual AA subs is challenging. The visualizations presented in this study illustrate the cumulative AA subs and the protein structural alterations caused by these cumulative substitutions. How to differentiate these AA subs is another future challenge for the community.

5 CONCLUSION

AA subs are crucial for regulating cellular activities in normal and cancer cells, impacting complex signaling and physiology. The CPTAC-released rich cancer-related proteomic datasets enabled researchers to pay attention to the importance of AA subs in multi-cancer analyses. Nonetheless, prior studies have so far scrutinized a restricted number of AA subs while efficient open-search peptide identification tools have been lacking. To fill the research gap in profiling AA subs in cancer analysis, in this paper, we used PIPI3[13] to analyze all types of AA subs across diverse cancers. Among the identified AA subs and AA substitution sites, some of them have been recorded to be caused by DNA or mRNA mutation in OncoDB,[34] while most of the identified AA subs are not recorded. These unrecorded AA subs are hypothesized to arise post-translation, underscoring the efficacy of our proteomic search pipeline.

Extracellular exosomes, as the smallest extracellular vesicles, take a crucial position in regulating various physiological functions such as immune responses, neuronal signaling, and the development

of cancer.[26] Both up- and down-regulated UPSPs across four cancer cohorts were enriched in the extracellular exosomes term. Three exosome-related proteins were selected for AA sub-profiling analysis. Within the protein GAPDH, we identified several AA subs from the GBM and LSCC cohorts in the interaction region for regulating and enhancing the aminoacylation process.[33] Meanwhile, in HBB, all of the significantly regulated AA subs were discovered from the two lung cancer cohorts and were down-regulated. Several common AA subs were found in both cohorts, including A11Q, G17Q, and F43S. These common amino acid substitutions could be explored to elucidate the common mechanisms underlying these two types of lung cancer. AA sub-modified VIM was also profiled in this study. A total of 26 substitution sites were identified in VIM, where some up-regulated sites were located in a coiled region ranging from 154 to 245. TM-score measurement showed that the predicted structure with the six cumulative AA subs differed from the experimentally verified structure (PDB: 3UF1).[41] These AA subs can potentially disrupt the stability of microtubules and influence the metastasis of GBM cancer when encapsulated within exosomes.

The analysis results showed that the up-regulated UPSPs in the cancer cohorts were functionally enriched in actin filament binding. SPTAN1, PLEC, and FLNA, which are key actin filament-binding proteins, were highlighted as frequently AA subs-modified proteins in different cancers. In PLEC, we identified an F330L substitution in the LSCC cohort, replacing an aromatic amino acid with a non-aromatic amino acid, in an actin-binding region.[51] We also noticed that most of the AA subs in PLEC were contained in the IF protein interaction region, and all of these AA subs were up-regulated. Domain 4 in FLNA is stabilized by interaction with domain 5, and the connection between these two domains is regulated by multiple hydrophobic β-sheets (PDB: 4M9P).[57] We identified a P584T substitution in domain 4 of FLNA, resulting in the change of a hydrophobic amino acid into a hydrophilic one. Similar AA subs in domain 4 have been discovered to cause familial cardiac valvular dystrophy.[58] Both our findings and the literature imply that the P584T substitution in domain 4 might regulate the function of FLNA.

We also found AA subs were enriched in the glycolysis and gluconeogenesis pathways. These pathways are interdependent, with one being active while the other is suppressed. Cancer cells often rely on glycolysis for energy due to metabolic inefficiencies.[62] Analysis of KEGG enrichment showed

an abundance of UPSPs in the glycolysis and gluconeogenesis pathways. Among the enriched proteins, we analyzed the AA subs profile of two proteins, LDHB and ALDOB. Within LDHB, in addition to the common AA subs across the LUAD, LSCC, and ccRCC cancer cohorts, we also located two AA subs from the GBM cohort in LDHB's binding region of the NAD^+. Considering the catalyzing function of LDHB in the reversible transformation of lactate and NAD^+ into pyruvate and NADH,[70] the two AA subs may disrupt the conversion process and impact the glycolysis and gluconeogenesis pathways. Further, the structure of ALDOB with cumulative down-regulated AA subs from the ccRCC cohort was predicted. Structure comparison between the prediction and the cryo-electron microscopy structure of human renal ALDOB (PDB: 8D44) showed an obvious structural change. The remaining AA subs also presented the potential to conduct molecular analysis and explore the functional implications of glycolysis and gluconeogenesis pathway-enriched proteins.

In summary, AA subs are important in regulating protein function and tumor physiology. Utilizing the recently proposed PIPI3, we provide a new angle to study and discover the protein level regulations of AA subs from different components, functions, and pathways in cancer proteomic data. The findings in this research provide insights into the AA subs study in multiple cancers for the first time and exhibit possible biomarkers for cancer diagnosis.

6 COMPETING INTERESTS

No competing interests are declared.

7 ACKNOWLEDGEMENTS

This work is partially supported by ITC grant MHP/033/20, RGC grants 16102422, R4012-18, C7015-23G, and T12-101/23N from the Hong Kong SAR government, Hetao Cooperation Zone

Project HzQB-KCZYB-2020083, and internal grants 3030_009, BG.001.2023, CSSET24SC01, and Z1056 from HKUST. We appreciate Tania Wilmshurst's help in proofreading the manuscript. We thank the Generic Diagramming Platform (https://biogdp.com/) for offering the schematic diagram.

References

1. L. Ding, M. H. Bailey, E. Porta-Pardo, V. Thorsson, A. Colaprico, D. Bertrand, D. L. Gibbs, A. Weerasinghe, K.-l. Huang, C. Tokheim, *et al.*, "Perspective on oncogenic processes at the end of the beginning of cancer genomics," *Cell*, vol. 173, no. 2, pp. 305–320, 2018.

2. A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon, "Tandem mass tags: A novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS," *Analytical Chemistry*, vol. 75, no. 8, pp. 1895–1904, 2003.

3. D. B. Doroshow and J. H. Doroshow, "Genomics and the history of precision oncology," *Surgical Oncology Clinics*, vol. 29, no. 1, pp. 35–49, 2020.

4. H. Rodriguez, J. C. Zenklusen, L. M. Staudt, J. H. Doroshow, and D. R. Lowy, "The next horizon in precision oncology: Proteogenomics to inform cancer diagnosis and treatment," *Cell*, vol. 184, no. 7, pp. 1661–1670, 2021.

5. M. Mann and O. N. Jensen, "Proteomic analysis of post-translational modifications," *Nature Biotechnology*, vol. 21, no. 3, pp. 255–261, 2003.

6. H. A. Shihab, J. Gough, D. N. Cooper, I. N. Day, and T. R. Gaunt, "Predicting the functional consequences of cancer-associated amino acid substitutions," *Bioinformatics*, vol. 29, no. 12, pp. 1504–1510, 2013.

7. A. Pataskar, J. Champagne, R. Nagel, J. Kenski, M. Laos, J. Michaux, H. S. Pak, O. B. Bleijerveld, K. Mordente, J. M. Navarro, *et al.*, "Tryptophan depletion results in tryptophan-to-phenylalanine substitutants," *Nature*, vol. 603, no. 7902, pp. 721–727, 2022.

8. C. Yang, A. Pataskar, X. Feng, J. M. Navarro, I. Paniagua, J. J. Jacobs, E. A. Zaal, C. R. Berkers, O. B. Bleijerveld, and R. Agami, "Arginine deprivation enriches lung cancer proteomes with cysteine by inducing arginine-to-cysteine substitutants," *Molecular Cell*, vol. 84, no. 10, pp. 1904–1916, 2024.

9. P. C. Ng and S. Henikoff, "Predicting the effects of amino acid substitutions on protein function," *Annu. Rev. Genomics Hum. Genet.*, vol. 7, no. 1, pp. 61–80, 2006.

10. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis: An International Journal*, vol. 20, no. 18, pp. 3551–3567, 1999.

11. J. K. Eng, T. A. Jahan, and M. R. Hoopmann, "Comet: An open-source MS/MS sequence database search tool," *Proteomics*, vol. 13, no. 1, pp. 22–24, 2013.

12. D. M. Creasy and J. S. Cottrell, "Unimod: Protein modifications for mass spectrometry," *Proteomics*, vol. 4, no. 6, pp. 1534–1536, 2004.

13. S. Lai, S. Dai, P. Zhao, C. Zhou, N. Li, and W. Yu, "Identifying crosstalks among post-translational modifications in lung cancer proteomic data," *bioRxiv*, 2024.

14. H. Chi, C. Liu, H. Yang, W.-F. Zeng, L. Wu, W.-J. Zhou, R.-M. Wang, X.-N. Niu, Y.-H. Ding, Y. Zhang, *et al.*, "Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine," *Nature Biotechnology*, vol. 36, no. 11, pp. 1059–1061, 2018.

15. S. Na, J. Kim, and E. Paek, "MODplus: Robust and unrestrictive identification of post-translational modifications using mass spectrometry," *Analytical Chemistry*, vol. 91, no. 17, pp. 11324–11333, 2019.

16. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, *et al.*, "Gene ontology: Tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

17. M. Kanehisa and S. Goto, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

18. Y. Li, Y. Dou, F. D. V. Leprevost, Y. Geffen, A. P. Calinawan, F. Aguet, Y. Akiyama, S. Anand, C. Birger, S. Cao, *et al.*, "Proteogenomic data and resources for pan-cancer analysis," *Cancer cell*, vol. 41, no. 8, pp. 1397–1406, 2023.

19. L.-B. Wang, A. Karpova, M. A. Gritsenko, J. E. Kyle, S. Cao, Y. Li, D. Rykunov, A. Colaprico, J. H. Rothstein, R. Hong, *et al.*, "Proteogenomic and metabolomic characterization of human glioblastoma," *Cancer Cell*, vol. 39, no. 4, pp. 509–528, 2021.

20. C. Huang, L. Chen, S. R. Savage, R. V. Eguez, Y. Dou, Y. Li, F. da Veiga Leprevost, E. J. Jaehnig, J. T. Lei, B. Wen, *et al.*, "Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma," *Cancer Cell*, vol. 39, no. 3, pp. 361–379, 2021.

21. M. A. Gillette, S. Satpathy, S. Cao, S. M. Dhanasekaran, S. V. Vasaikar, K. Krug, F. Petralia, Y. Li, W.-W. Liang, B. Reva, *et al.*, "Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma," *Cell*, vol. 182, no. 1, pp. 200–225, 2020.

22. S. Satpathy, K. Krug, P. M. J. Beltran, S. R. Savage, F. Petralia, C. Kumar-Sinha, Y. Dou, B. Reva, M. H. Kane, S. C. Avanessian, *et al.*, "A proteogenomic portrait of lung squamous cell carcinoma," *Cell*, vol. 184, no. 16, pp. 4348–4371, 2021.

23. D. J. Clark, S. M. Dhanasekaran, F. Petralia, J. Pan, X. Song, Y. Hu, F. da Veiga Leprevost, B. Reva, T.-S. M. Lih, H.-Y. Chang, *et al.*, "Integrated proteogenomic characterization of clear cell renal cell carcinoma," *Cell*, vol. 179, no. 4, pp. 964–983, 2019.

24. European Organization For Nuclear Research and OpenAIRE, "Zenodo," 2013.

25. U. Consortium, "Uniprot: A worldwide hub of protein knowledge," *Nucleic Acids Research*, vol. 47, no. D1, pp. D506–D515, 2019.

26. M. A. Di Bella, "Overview and update on extracellular vesicles: Considerations on exosomes and their application in modern medicine," *Biology*, vol. 11, no. 6, p. 804, 2022.

27. J. Skog, T. Würdinger, S. Van Rijn, D. H. Meijer, L. Gainche, W. T. Curry Jr, B. S. Carter, A. M. Krichevsky, and X. O. Breakefield, "Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers," *Nature Cell Biology*, vol. 10, no. 12, pp. 1470–1476, 2008.

28. M. Logozzi, E. Spugnini, D. Mizzoni, R. Di Raimo, and S. Fais, "Extracellular acidity and increased exosome release as key phenotypes of malignant tumors," *Cancer and Metastasis Reviews*, vol. 38, pp. 93–101, 2019.

29. Z. A. Malik, K. S. Kott, A. J. Poe, T. Kuo, L. Chen, K. W. Ferrara, and A. A. Knowlton, "Cardiac myocyte exosomes: Stability, HSP60, and proteomics," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 304, no. 7, pp. H954–H965, 2013.

30. G. Gurland and G. G. Gundersen, "Stable, detyrosinated microtubules function to localize vimentin intermediate filaments in fibroblasts.," *The Journal of Cell Biology*, vol. 131, no. 5, pp. 1275–1290, 1995.

31. L. Schaedel, C. Lorenz, A. V. Schepers, S. Klumpp, and S. Köster, "Vimentin intermediate filaments stabilize dynamic microtubules by direct interactions," *Nature Communications*, vol. 12, no. 1, p. 3799, 2021.

32. A. Tabatabaee, B. Nafari, A. Farhang, A. Hariri, A. Khosravi, A. Zarrabi, and M. Mirian, "Targeting vimentin: A multifaceted approach to combatting cancer metastasis and drug resistance," *Cancer and Metastasis Reviews*, vol. 43, no. 1, pp. 363–377, 2024.

33. K. Wakasugi, T. Nakano, and I. Morishima, "Oxidative stress-responsive intracellular regulation specific for the angiostatic form of human tryptophanyl-trna synthetase," *Biochemistry*, vol. 44, no. 1, pp. 225–232, 2005.

34. G. Tang, X. Liu, M. Cho, Y. Li, D.-H. Tran, and X. Wang, "Pan-cancer discovery of somatic mutations from RNA sequencing data," *Communications Biology*, vol. 7, no. 1, p. 619, 2024.

35. J. Dacie, N. Shinton, P. Gaffney, R. Carrell, and H. Lehmann, "Haemoglobin Hammersmith (β 42 (CD1) Phe→Ser)," *Nature*, vol. 216, no. 5116, pp. 663–665, 1967.

36. S. Rahbar, R. Feagler, and E. Beutler, "Hemoglobin Hammersmith (β42 (CDI) Phe→Ser) Associated with severe Hemolytic Anemia," *Hemoglobin*, vol. 5, no. 1, pp. 97–105, 1981.

37. O. De Bei, M. Marchetti, L. Ronda, E. Gianquinto, L. Lazzarato, D. Y. Chirgadze, S. W. Hardwick, L. R. Cooper, F. Spyrakis, B. F. Luisi, *et al.*, "Cryo-EM structures of staphylococcal

IsdB bound to human hemoglobin reveal the process of heme extraction," *Proceedings of the National Academy of Sciences*, vol. 119, no. 14, p. 119, 2022.

38. G. Layer, J. Reichelt, D. Jahn, and D. W. Heinz, "Structure and function of enzymes in heme biosynthesis," *Protein Science*, vol. 19, no. 6, pp. 1137–1161, 2010.

39. F. Baklouti, Y. Giraud, A. Francina, G. Richard, J. Favre-Gilly, and J. Delaunay, "Hehemoglobin Pierre-Bénite [β90 (F6) Glu→Asp], a new high affinity variant found in a French family," *Hemoglobin*, vol. 12, no. 2, pp. 171–177, 1988.

40. H. Wajcman, J. A. i Bascompte, D. Labie, C. Poyart, and B. Bohn, "Structural and functional studies of hemoglobin Barcelona ($\alpha 2\beta 2$ 94 Asp (FG1)→His): Consequences of altering an important intrachain salt bridge involved in the alkaline Bohr effect," *Journal of Molecular Biology*, vol. 156, no. 1, pp. 185–202, 1982.

41. A. Aziz, J. F. Hess, M. S. Budamagunta, J. C. Voss, A. P. Kuzin, Y. J. Huang, R. Xiao, G. T. Montelione, P. G. Fitzgerald, and J. F. Hunt, "The structure of vimentin linker 1 and rod 1B domains characterized by site-directed spin-labeling electron paramagnetic resonance (SDSL-EPR) and X-ray crystallography," *Journal of Biological Chemistry*, vol. 287, no. 34, pp. 28349–28361, 2012.

42. Y. Zhang and J. Skolnick, "Scoring function for automated assessment of protein structure template quality," *Proteins: Structure, Function, and Bioinformatics*, vol. 57, no. 4, pp. 702–710, 2004.

43. G. Cooper and K. Adams, *The cell: A molecular approach.* Oxford University Press, 2022.

44. M. Izdebska, W. Zielińska, D. Grzanka, and M. Gagat, "The role of actin dynamics and actin-binding proteins expression in epithelial-to-mesenchymal transition and its association with cancer progression and evaluation of possible therapeutic targets," *BioMed Research International*, vol. 2018, no. 1, p. 373, 2018.

45. J. R. Stehn, G. Schevzov, G. M. O'Neill, and P. W. Gunning, "Specialisation of the tropomyosin composition of actin filaments provides new potential targets for chemotherapy," *Current Cancer Drug Targets*, vol. 6, no. 3, pp. 245–256, 2006.

46. M. Tafazzoli-Shadpour, E. Mohammadi, and E. Torkashvand, "Mechanics of actin filaments in cancer onset and progress," in *International Review of Cell and Molecular Biology*, vol. 355, pp. 205–243, Elsevier, 2020.

47. H. O. Villar and L. M. Kauvar, "Amino acid preferences at protein binding sites," *FEBS Letters*, vol. 349, no. 1, pp. 125–130, 1994.

48. L. H. Teliska and M. N. Rasband, "Spectrins," *Current Biology*, vol. 31, no. 10, pp. R504–R506, 2021.

49. A. Ackermann and A. Brieger, "The role of nonerythroid spectrin αii in cancer," *Journal of Oncology*, vol. 2019, no. 1, p. 604, 2019.

50. T. M. Svitkina, A. B. Verkhovsky, and G. G. Borisy, "Plectin sidearms mediate interaction of intermediate filaments with microtubules and other components of the cytoskeleton," *The Journal of Cell Biology*, vol. 135, no. 4, pp. 991–1007, 1996.

51. B. Garcıa-Alvarez, A. Bobkov, A. Sonnenberg, and J. M. de Pereda, "Structural and functional analysis of the actin binding domain of plectin suggests alternative mechanisms for binding to F-actin and integrin β4," *Structure*, vol. 11, no. 6, pp. 615–625, 2003.

52. J.-E. Bouameur, B. Favre, L. Fontao, P. Lingasamy, N. Begré, and L. Borradori, "Interaction of plectin with keratins 5 and 14: Dependence on several plectin domains and keratin quaternary structure," *Journal of Investigative Dermatology*, vol. 134, no. 11, pp. 2776–2783, 2014.

53. A. A. Ketebo, C. Park, J. Kim, M. Jun, and S. Park, "Probing mechanobiological role of filamin A in migration and invasion of human U87 glioblastoma cells using submicron soft pillars," *Nano Convergence*, vol. 8, no. 1, p. 19, 2021.

54. L. Cai, Q. Li, W. Li, C. Wang, M. Tu, Z. Zhu, Z. Su, and X. Lu, "Calpain suppresses cell growth and invasion of glioblastoma multiforme by producing the cleavage of filamin A," *International Journal of Clinical Oncology*, vol. 25, pp. 1055–1066, 2020.

55. N. Chantaravisoot, P. Wongkongkathep, J. A. Loo, P. S. Mischel, and F. Tamanoi, "Significance of filamin A in mTORC2 function in glioblastoma," *Molecular Cancer*, vol. 14, pp. 1–14, 2015.

56. D. V. Iwamoto, A. Huehn, B. Simon, C. Huet-Calderwood, M. Baldassarre, C. V. Sindelar, and D. A. Calderwood, "Structural basis of the filamin A actin-binding domain interaction with F-actin," *Nature Structural & Molecular Biology*, vol. 25, no. 10, pp. 918–927, 2018.

57. R. Sethi, J. Seppälä, H. Tossavainen, M. Ylilauri, S. Ruskamo, O. T. Pentikäinen, U. Pentikäinen, P. Permi, and J. Ylänne, "A novel structural unit in the N-terminal region of filamins," *Journal of Biological Chemistry*, vol. 289, no. 12, pp. 8588–8598, 2014.

58. F. Kyndt, J.-P. Gueffet, V. Probst, P. Jaafar, A. Legendre, F. Le Bouffant, C. Toquet, E. Roy, L. McGregor, S. A. Lynch, *et al.*, "Mutations in the gene encoding filamin A as a cause for familial cardiac valvular dystrophy," *Circulation*, vol. 115, no. 1, pp. 40–49, 2007.

59. R. M. Savoy and P. M. Ghosh, "The dual role of filamin A in cancer: Can't live with (too much of) it, can't live without it," *Endocrine-related cancer*, vol. 20, no. 6, pp. R341–R356, 2013.

60. J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, *et al.*, "Accurate structure prediction of biomolecular interactions with Alphafold 3," *Nature*, pp. 1–3, 2024.

61. R. Gupta, N. Gupta, R. Gupta, and N. Gupta, "Glycolysis and gluconeogenesis," *Fundamentals of Bacterial Physiology and Metabolism*, pp. 267–287, 2021.

62. J. Zheng, "Energy metabolism of cancer: Glycolysis versus oxidative phosphorylation," *Oncology Letters*, vol. 4, no. 6, pp. 1151–1157, 2012.

63. R. Ma, W. Zhang, K. Tang, H. Zhang, Y. Zhang, D. Li, Y. Li, P. Xu, S. Luo, W. Cai, *et al.*, "Switch of glycolysis to gluconeogenesis by dexamethasone for treatment of hepatocarcinoma," *Nature Communications*, vol. 4, no. 1, p. 2508, 2013.

64. A. Cheng, P. Zhang, B. Wang, D. Yang, X. Duan, Y. Jiang, T. Xu, Y. Jiang, J. Shi, C. Ding, *et al.*, "Aurora-A mediated phosphorylation of LDHB promotes glycolysis and tumor progression by relieving the substrate-inhibition effect," *Nature Communications*, vol. 10, no. 1, p. 5566, 2019.

65. M. G. Vander Heiden, L. C. Cantley, and C. B. Thompson, "Understanding the warburg effect: The metabolic requirements of cell proliferation," *Science*, vol. 324, no. 5930, pp. 1029–1033, 2009.

66. M. L. McCleland, A. S. Adler, L. Deming, E. Cosino, L. Lee, E. M. Blackwood, M. Solon, J. Tao, L. Li, D. Shames, *et al.*, "Lactate dehydrogenase B is required for the growth of KRAS-dependent lung adenocarcinomas," *Clinical Cancer Research*, vol. 19, no. 4, pp. 773–784, 2013.

67. L. Brisson, P. Bański, M. Sboarina, C. Dethier, P. Danhier, M.-J. Fontenille, V. F. Van Hée, T. Vazeille, M. Tardy, J. Falces, *et al.*, "Lactate dehydrogenase B controls lysosome activity and autophagy in cancer," *Cancer Cell*, vol. 30, no. 3, pp. 418–431, 2016.

68. W. M. Haschek, C. G. Rousseaux, M. A. Wallig, and B. Bolon, *Haschek and Rousseaux's Handbook of Toxicologic Pathology, Volume 1: Principles and Practice of Toxicologic Pathology.* Academic Press, 2021.

69. Y. Shao, B. Wu, Z. Yang, Z. Liu, Y. Ma, H. Huang, Y. Liu, Z. Wang, W. Hu, Y. Wang, *et al.*, "ALDOB represents a potential prognostic biomarker for patients with clear cell renal cell carcinoma," *Translational Andrology and Urology*, vol. 12, no. 4, p. 549, 2023.

70. C. J. Valvona, H. L. Fillmore, P. B. Nunn, and G. J. Pilkington, "The regulation and function of lactate dehydrogenase a: Therapeutic potential in brain tumor," *Brain Pathology*, vol. 26, no. 1, pp. 3–17, 2016.

71. N. Cross, T. Cox, R. De Franchis, G. Sebastio, C. Dazzo, D. Tolan, C. Grégori, M. Odievre, M. Vidailhet, V. Romano, *et al.*, "Molecular analysis of aldolase B genes in hereditary fructose intolerance," *The Lancet*, vol. 335, no. 8685, pp. 306–309, 1990.

72. G. Esposito, M. R. Imperato, L. Ieno, R. Sorvillo, V. Benigno, G. Parenti, R. Parini, L. Vitagliano, A. Zagari, and F. Salvatore, "Hereditary fructose intolerance: Functional study of two novel ALDOB natural variants and characterization of a partial gene deletion," *Human mutation*, vol. 31, no. 12, pp. 1294–1303, 2010.

73. J. Lau and D. Tolan, "Screening for hereditary fructose intolerance mutations by reverse dot-blot," *Molecular and cellular probes*, vol. 13, no. 1, pp. 35–40, 1999.

74. J. Sygusch, D. Beaudry, and M. Allaire, "Molecular architecture of rabbit skeletal muscle aldolase at 2.7-Å resolution.," *Proceedings of the National Academy of Sciences*, vol. 84, no. 22, pp. 7846–7850, 1987.

75. S. Kwon, J. Safer, D. T. Nguyen, D. Hoksza, P. May, J. A. Arbesfeld, A. F. Rubin, A. J. Campbell, A. Burgin, and S. Iqbal, "Genomics 2 Proteins portal: A resource and discovery tool for linking genetic screening outputs to protein sequences and structures," *Nature Methods*, pp. 1–11, 2024.

76. R. R. Thangudu, P. A. Rudnick, M. Holck, D. Singhal, M. J. MacCoss, N. J. Edwards, K. A. Ketchum, C. R. Kinsinger, E. Kim, and A. Basu, "Abstract LB-242: Proteomic Data Commons: A resource for proteogenomic analysis," *Cancer Research*, vol. 80, no. 16, p. 242, 2020.

77. M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, *et al.*, "A cross-platform toolkit for mass spectrometry and proteomics," *Nature Biotechnology*, vol. 30, no. 10, pp. 918–920, 2012.

78. D. Klopfenstein, L. Zhang, B. S. Pedersen, F. Ramírez, A. Warwick Vesztrocy, A. Naldi, C. J. Mungall, J. M. Yunes, O. Botvinnik, M. Weigel, *et al.*, "GOATOOLS: A Python library for Gene Ontology analyses," *Scientific Reports*, vol. 8, no. 1, pp. 1–17, 2018.

79. Z. Fang, X. Liu, and G. Peltz, "GSEApy: A comprehensive package for performing gene set enrichment analysis in Python," *Bioinformatics*, vol. 39, no. 1, p. 757, 2023.

80. D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, *et al.*, "The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic Acids Research*, vol. 49, no. D1, pp. D605–D612, 2021.

81. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

82. E. F. Pettersen, T. D. Goddard, C. C. Huang, E. C. Meng, G. S. Couch, T. I. Croll, J. H.

Morris, and T. E. Ferrin, "UCSF Chimerax: Structure visualization for researchers, educators, and developers," *Protein Science*, vol. 30, no. 1, pp. 70–82, 2021.