ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Mutational Impacts on the N and C Terminal Domains of the MUC5B Protein: A Transcriptomics and Structural Biology Study

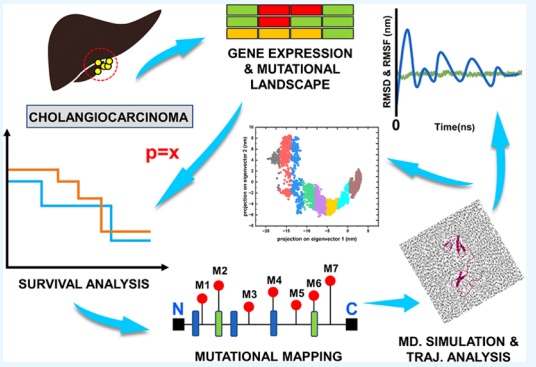
Aamir Mehmood, Sadia Nawab, Yifan Jin, Aman Chandra Kaushik,* and Dong-Qing Wei*

Cite This: *ACS Omega* 2023, 8, 3726−3735

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Cholangiocarcinoma (CCA) involves various epithelial tumors historically linked with poor prognosis because of its aggressive sickness course, delayed diagnosis, and limited efficacy of typical chemotherapy in its advanced stages. In-depth molecular profiling has exposed a varied scenery of genomic alterations as CCA's oncogenic drivers. Previous studies have mainly focused on commonly occurring TP53 and KRAS alterations, but there is limited research conducted to explore other vital genes involved in CCA. We retrieved data from The Cancer Genome Atlas (TCGA) to hunt for additional CCA targets and plotted a mutational landscape, identifying key genes and their frequently expressed variants. Next, we performed a survival analysis for all of the top genes to shortlist the ones with better significance. Among those genes, we observed that MUC5B has the most significant *p*-value of 0.0061. Finally, we chose two missense mutations at different positions in the vicinity of MUC5B N and C terminal domains. These mutations were further subjected to molecular dynamics (MD) simulation, which revealed noticeable impacts on the protein structure. Our study not only reveals one of the highly mutated genes with enhanced significance in CCA but also gives insights into the influence of its variants. We believe these findings are a good asset for understanding CCA from genomics and structural biology perspectives.



1. INTRODUCTION

Cholangiocarcinoma (CCA) is an assorted category of hostile malignancies that may ascend from various sites inside the biliary tree. We classify CCA into intrahepatic (iCCA), perihilar (pCCA), and distal CCA (dCCA) based on their origin, and they vary in etiology, risk factors, prediction, and medical supervision. This group does not include gallbladder cancer and tumors that emerge in the ampulla of Vater. Considering together, the ratio of iCCA and pCCA exceeds 90% of the global CCA cases.[1−4] These tumors have a shocking mortality rate of around 2% of all cancer-associated fatalities globally.[2] This is due to their quiet appearance, extremely aggressive nature, and resistance to treatment. Histological confirmation is required since the present noninvasive methods of diagnosing CCA are insufficient. Furthermore, the effectiveness of currently accessible therapeutic strategies is seriously jeopardized by the heterogeneousness of CCAs both genomically and at the molecular levels. Early-stage CCAs are generally asymptomatic, which makes it challenging to identify the disease until it is well along. This severely limits the possibilities for treatment and leaves patients with a grim prognosis.[5−7] Despite being a rare cancer, CCA has been observed to have an increased incidence (0.3−6 per 100,000 people yearly)[1] and mortality rate (1−6 per 100,000 people annually worldwide,[8] excluding several regions with occurrence >6 per 100,000 habitats like Thailand, China, and South Korea) over the previous decades, signaling a universal health issue. Despite bringing improvements to cause awareness, CCA's understanding, diagnosis, and therapy have not improved satisfactorily in the last ten years. We still face a 5 year survival (7−20%) as well as dismal tumor recurrence rates following resection.[9−15] To enhance patient welfare and results, thorough research on these tumors is urgently required. Given the high variability in CCAs, personalized profiling at the molecular, epigenetic, and genomic levels is a crucial strategy for determining the pathophysiology, opening doors for novel therapeutic methods and precision medicine.[2] Mucins are glycoproteins with high molecular weight distinguished by carbohydrate sugars linked to serine and threonine via the *O*-glycosidic bond. They combine to create a heterogeneous collection of polydisperse, highly glycosylated macromolecules with large molecular masses. Mucins comprise most of the mucus,[16,17] and it is widely known for giving mucin gel its unique characteristics. We may now reach these macromolecules' peptide moiety via immunohistochemistry or in situ hybridization owing to recent developments in cloning

Received: August 1, 2022
Accepted: November 18, 2022
Published: January 20, 2023

human mucin genes. To date, clones corresponding to eight apomucins (MUC1−MUC7 and MUC5B) have been recognized,[18−24] and tissue- and cell-specific countenance of these mucin genes is springing, signifying a distinct role of each gene.[25,26] Besides, MUC5B is mutated in other cancers, such as colorectal and renal cancer.[27,28] The presence of MUC5B mutation potentially acts as a predictive marker in lung cancer patients.[29] These mutations are also involved in cell adhesion and can be used for predicting the aggressiveness in papillary thyroid microcarcinomas (PTMCs).[30] Patients with MUC5B get a higher burden of tumor mutations (TMB). According to an examination of the immunological signature, MUC5B mutation is connected to increased genes that control cytolytic immune activity, activated T-cell production, and IFN- release. A potentially unique and practical method for predicting the prognosis of patients with endometrial cancer is the identification of MUC5B mutation through genomic profiling.[31] Based on the expression of this gene in cancers and the importance of missense mutations, we obtained The Cancer Genome Atlas (TCGA)[32−36] data showing MUC5B as a frequently mutated gene in CCA, which is considered further for its mutational studies using molecular dynamic simulations[37,38] to understand the impact of point substitutions on the protein conformation. Our computational structural biology approach provides a detailed analysis of CCA key mutations, and the results obtained could help further understand the role of MUC5B in cancers and the disease in general.

2. MATERIALS AND METHODS

2.1. Data Curation. We retrieved CCA data from the TCGA (https://www.cancer.gov/aboutnci/organization/ccg/research/structural-genomics/tcga), considering the available CCA studies until March 2022. It contained TCGA barcodes, diagnosis, and mutational details.

2.2. Mutational Landscape and Differential Gene Expression. To explore the nature of variants, single nucleotide variants (SNVs), variant classification, and frequently mutated genes in CCA, we used maftools[39] in RStudio.[40] We analyzed the differential gene expression for each series, where the adjusted *p*-value and |logFC| were maintained as <0.01 and >1, respectively. We also examined the biological process (BP), cellular component (CC), and molecular function (MF) during the process of functional enrichment.

2.3. Kaplan−Meier Survival Analysis. To evaluate the prognostic value of shortlisted genes in CCA, we used the "survival" package in RStudio to plot the mRNA expression levels' survival plots. All of the top ten genes were considered for survival analysis to shortlist the one with the most significant *p*-value. The hazard ratio and confidence intervals were calculated, and no threshold was set for the *p*-value to observe the difference between the survival probability of two genes.

2.4. Variant Selection across the Target Protein. The maftools package also helps plot variants' class and their position. We observed lollipop plots for the selected gene (protein) with greater significance and mapped mutations on its protein for further study. Several mutations could be seen at different positions; however, changes in the hotspot regions are more likely to affect the function comparatively. Therefore, we selected only two missense mutations (Y919D and D5551Y) that were harbored inside specific domains.

2.5. Mutational Impacts on Structural Stability.

2.5.1. Target Structure. There is no crystal structure available for Human MUC5B, and the protein length is >5000 amino acids, which makes it computationally intensive to simulate and may not reflect the true impact of a single-point mutation. Therefore, we used an online service known as Robetta[41] to predict the structure of our domains of interest only. Next, we validated the predicted conformations through the Ramachandran plot analysis.[42] Besides, the Chimera software suite allows structure visualization, creation, optimization, and drug development. Therefore, we used it for structure preparation and creating mutant (MT) conformations. Mutants are generated by replacing the native amino acid with a residue of interest and then minimizing the structure that stabilizes and fixes the mutated residue.

2.5.2. MD Simulation. We subjected the selected MUC5B mutations to MD simulation to examine their effects on atoms' topography and the general domain structure. We used Chimera for the initial structure preparation of proteins, such as energy minimization of the MT structures, to reduce the chances of atomic clashes and unwanted errors. We conducted MD simulations using the GROMACS force field (GROMOS96 43a1).[43] By including explicit flexible SPC molecules of water implanted in a cubic box with edges located ≥10 Å from all of the protein atoms, four apo simulations for native and mutant proteins were performed. The box angles on either side were kept at 90°, and the box dimensions and vectors were maintained at 4.256 × 4.061 × 4.142 and 6.7 × 6.7 × 6.7 nm^3, respectively. Next, we introduced counterions of Na^+ that balanced the system's net charge to create a neutral system. The solvated structures were reduced for 50,000 steps using the steepest descent minimization. This stops when the total peak force is <1000 kJ/mol/nm. A constant 300 K temperature, regular 1 bar of pressure, and an even 2 fs of time step were kept to reach an equilibrium state. Under the position restraint circumstances for the heavyweight atoms, the LINCS (LINear Constraint SolVer)[44] constraints and nonbonded pair list were modified every ten steps. Using a particle mesh Ewald technique, electrostatic interactions were estimated.[45] The temperature within the box was kept constant using the v-rescale (modified Berendsen thermostat) temperature coupling method.[46] The final step was to run four apo simulations (MUC5B Y919D, D5551Y, and their native structures) lasting 100 ns.

2.5.3. Postsimulation Study. The physical characteristics as a time function, for instance, the root-mean-square deviation (RMSD), root-mean-square fluctuation (RMSF), radius of gyration (R_g), principal component analysis (PCA), and free energy landscape (FEL), were calculated to track atomic drifts and deviations in the MUC5B MT structures, concluding the differences between MT and wild-type (WT). The protein chain's local changes are observed using RMSF. The R_g plotted against time provides a measure of the folding compactness. Throughout the simulation, PCA successfully caught the large-amplitude oscillations in the protein structure. First, using the given GROMACS tool, the covariance matrix was calculated in view of backbone carbon atoms. Next, the predetermined covariance was diagonalized to score the eigenvectors and eigenvalues.

3. RESULTS

3.1. Mutational Landscape Analysis. The mutational landscape for CCA has been visualized, containing variant type,

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

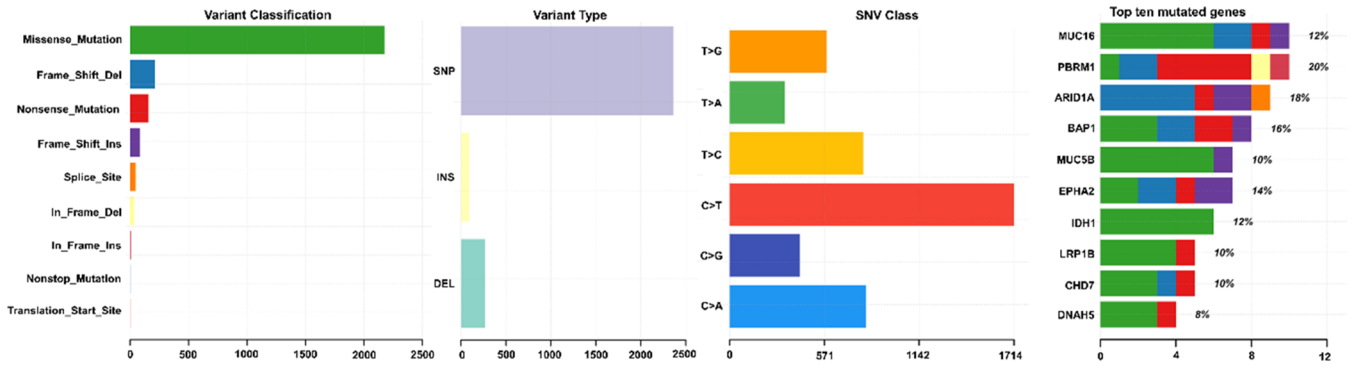


Figure 1. Mutational landscape. The variant type and classification, SNV class, and top ten frequently mutated genes in cholangiocarcinoma are visually presented. Each color corresponds to a specific variant type or class.

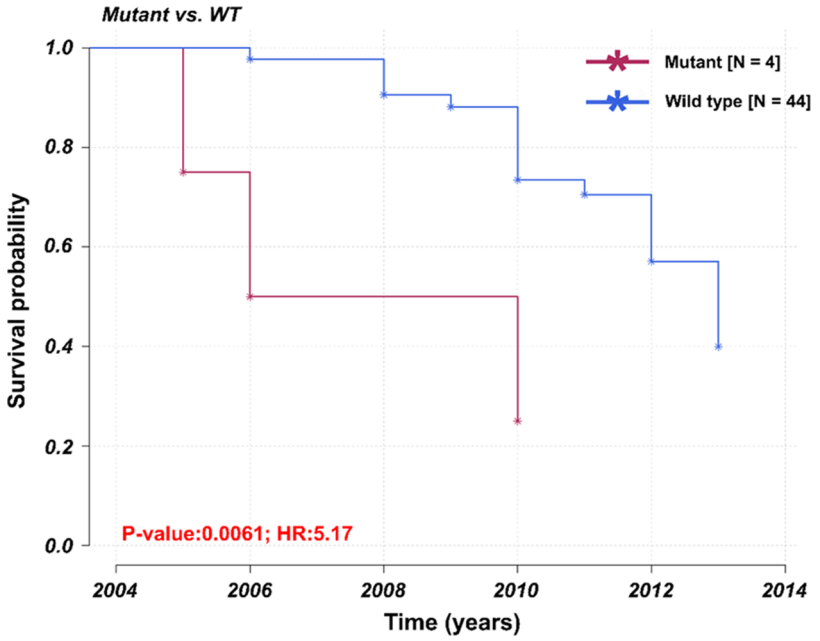


Figure 2. Survival analysis of the human MUC5B gene with the most significant *p*-value. This gene frequently mutates in cholangiocarcinoma patients and is observed to harbor only missense and frameshift insertions.

classification, SNVs, and frequently mutated genes (Figure 1). Most of the variant types in CCA are single nucleotide polymorphisms (SNPs), followed by deletions, and the lowest ratio is of insertions. The classification chart shows that missense mutations are reported in most cases, and their ratio is much higher than any other form of variant. The C > T and C > A variations are highly reported and bear great importance to be considered, as can be seen in the SNV class graph. Besides the MUC16, PBRM1, AR1D1, and BAP1, the MUC5B is frequently mutated though its ratio is lower than the aforementioned genes. We also plotted the mutation type in each cancer to get a clearer idea of CCA mutations. This plot confirms that most of the mutations reported are missense, nonsense, and frameshift deletions. The variant types are mostly SNPs, and the C > T conversion ratio is far higher than any other mutation, as observed in the SNV class panel. Additionally, the top ten mutated genes are plotted in ascending order, among which we can see that PBRM1 is the highest and DNAH5 is the least frequently mutated gene. However, all of these top ten genes are considered for further evaluation to shortlist a gene with higher significance.

The differential expression analysis revealed that among 40, most genes are underexpressed, which may be caused by mutational impacts (Figure S1). There is no MUC5B listed in these genes, which signifies that mutations in this gene did not significantly impact its gene expression. However, exploring the protein would be more helpful in understanding the consequences. The biological process (BP), cellular components (CCs), molecular function (MF), and general pathways considered during the functional enrichment are given in Figure S2.

3.2. Target Gene Selection. We plotted the survival analysis for all of the top ten mutated genes to shortlist the ones with the most significant *p*-values. Of all of the genes, we observed that MUC5B has a *p*-value of 0.0061, which is far more significant than the rest of the shortlisted genes (Figure 2). Survival plots for the remaining nine genes have been provided in the Supporting Material (Figure S3).

3.3. Selection of Target Variants. Since MUC5B has the most significant *p*-value and is considered for further evaluation, we examined the lollipop plot for selecting important mutations in this gene's protein. It can be seen that Y919D and D5551Y are the two missense mutations

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

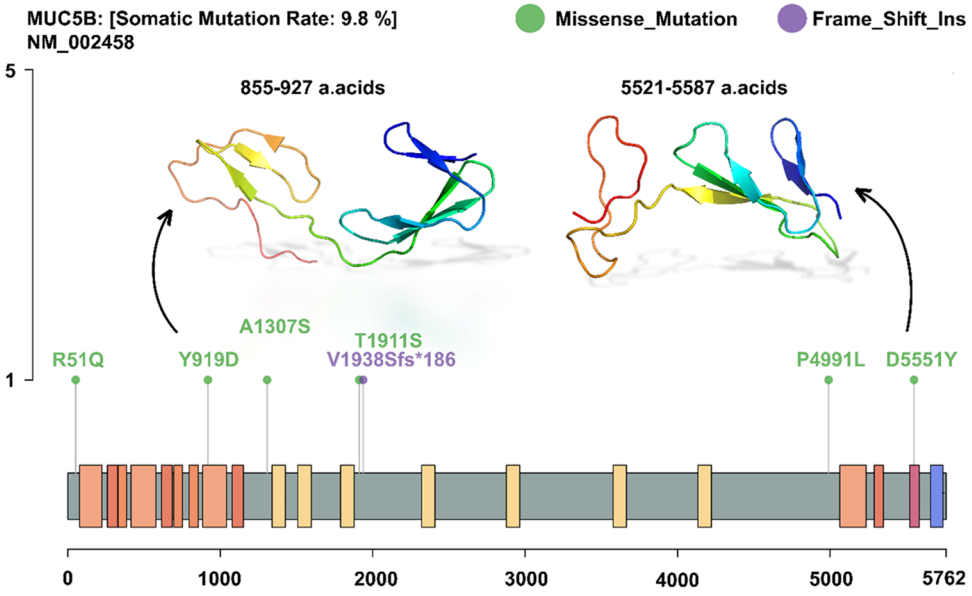


Figure 3. Mapping mutations on the MUC5B protein. Structures of the targeted regions are given concerning their mutation position. These given structures are wild-type proteins and are rendered before the MD simulation.

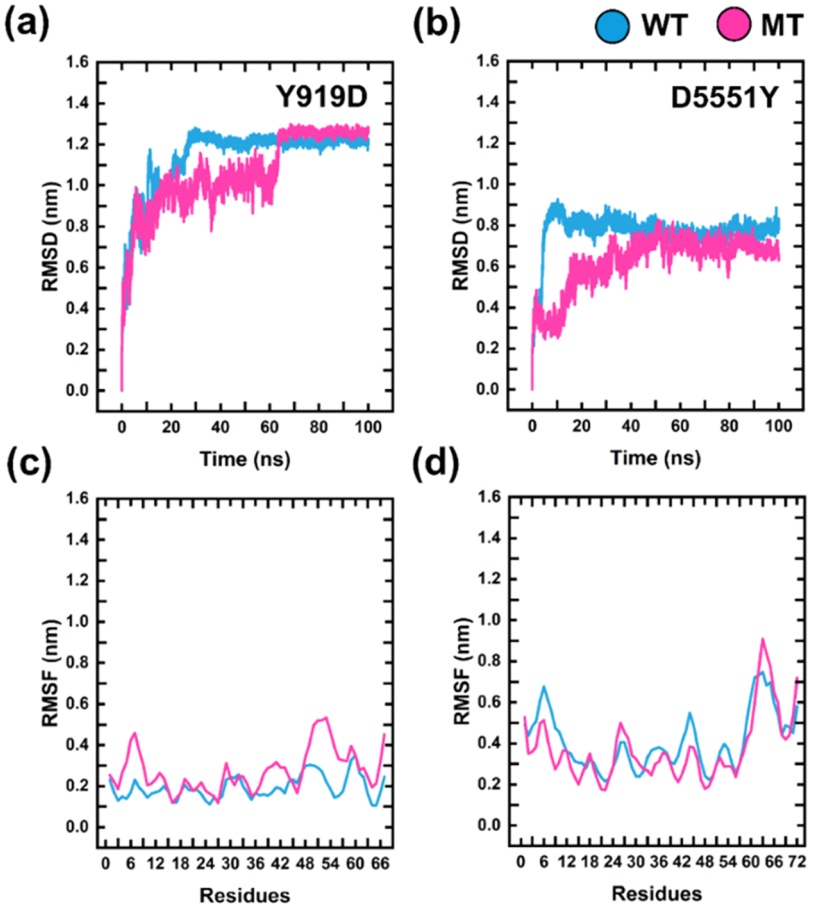


Figure 4. RMSD and RMSF. (a, b) RMSD of the WT and MT of both targeted domains. (c, d) RMSF of both domains of the MUC5B protein. Deviations can be observed in all cases.

inside the specific domains, while the rest of the variants are off those regions (Figure 3). We selected only these two missense mutations because domains (particularly hotspot regions) are vital for a protein's function, and any variations in these regions could directly affect the function. The lollipop plots for the remaining genes are given in Figure S4.

3.4. Structure Prediction and Target Domains. Since the human MUC5B protein has more than 5000 amino acids, we only considered the domains where the mutations exist.

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

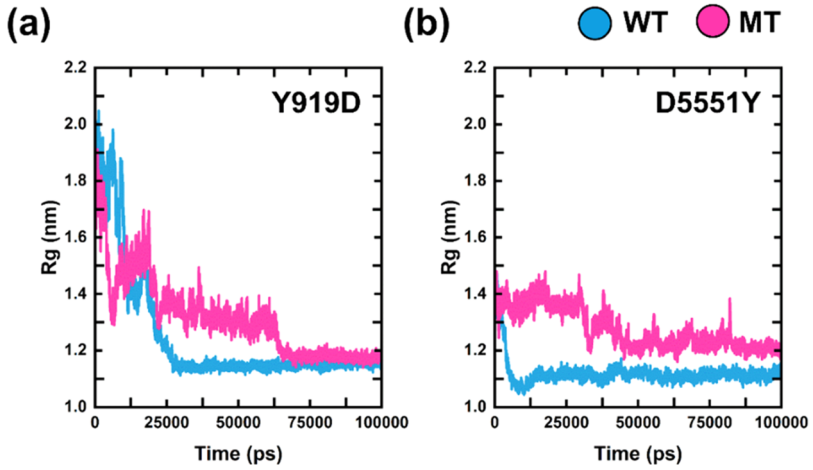


Figure 5. Radius of gyration for (a) N and (b) C domains revealed significant fluctuations compared to their native protein structure.

Therefore, we selected two domains with lengths of 72 and 67 amino acids in the vicinity of the N and C terminal domains, respectively. Structures for these domains are provided in Figure 3. The selected parts range correspondingly from 855 to 927 and from 5521 to 5587 amino acids.

3.5. Trajectory Analysis. *3.5.1. Backbone RMSD Calculation.* We obtained MD simulation trajectories of the targeted MUC5B WT and MT domains after a period of 100 ns. The backbone RMSD of each WT and MT was computed using the GROMACS default function g_rms.[47] The mutations in targeted structures, let us call them the N and C domains, showed slight deviations relative to their WT protein (Figure 4). The WT protein shows minor fluctuations in the initial 25 ns and then stabilizes until the last frame of the simulation, maintaining an RMSD value of 1.3 nm. On the other hand, the mutant Y919D fluctuates until 65 ns and then stabilizes until the last frame as the WT (Figure 4a). Here, we can make two conclusions. First, the WT's RMSD is much higher than the typical RMSD values of 0.4/5 nm; this is because our simulated structure is not a complete protein in its compact form but only a piece of the protein that is naturally expected to behave unstably. Another reason for this higher RMSD is its loop structure, which logically produces higher fluctuations in the system. Compared to the WT protein, we maintain that Y919D impacts the protein's structure. We can confirm that it fluctuated significantly in more than 60% of the simulation. Even if it stabilizes at the end, but remains slightly higher than the WT though this change could be considered negligible in some cases, the fact is that Y919D affected the structural stability noticeably.

Similarly, the RMSD analysis of the C domain remains relatively stable throughout the simulation, excluding the first 10−13 ns. Minor fluctuations can be observed at 50 ns and around 80 ns, but the overall system stays calm, maintaining a backbone RMSD value of 0.8 nm, much lower than the N domain. However, the MT C domain (D5551Y) shows apparent differences in the RMSDs (Figure 4b). It starts from 0.2 nm and gradually elevates, reaching 0.8 nm until 50 ns, from where it stays close to the WT, exhibiting minor and major fluctuations until 90 ns, from where it appears to deviate again until the last frame.

3.5.2. Structural Flexibility Evaluation. To investigate changes in the dynamics of residues brought on by the Y919D and D5551Y mutations, MUC5B N and C domains' backbone RMSF values were estimated. Figure 4 shows the change in amino acid variations compared to the WT, revealing notable changes. Until the simulation's last frame, both MTs show higher RMSF scores compared to the WT, coupled with noticeable fluctuations, particularly close to the altered residues. Deviations from the WT protein by Y919D and D5551Y variants are prominent compared to those observed in RMSD values. This means that the overall system did not experience high impacts, but residual fluctuations revealed the alterations caused by these N and C domain variants. In the case of the N domain (Figure 4c), the WT RMSF stays up to 0.6 nm but falls to 0.3 nm and takes long leaps until the end of the simulation. However, compared with the MT N domain (Y919D), the RMSF keeps fluctuating until the first 24 residues and then goes higher than the WT, reaching up to 0.5 nm, and then keeps changing, taking a sharp leap between residue 56 and residue 58. Overall, both the wild and MT have a competing RMSF value; the MT is still observed to deviate quite prominently. In the case of the C domain (Figure 4d), the RMSF values stay lower, with a maximum peak of 0.3 nm in the case of WT and 0.6 in the case of MT. There is noted a clear difference between the two proteins, reflecting the impression of D5551Y on the C domain's conformation.

3.5.3. Analyzing the Structural Compactness (R_g). Over the duration of a 100 ns simulation, studying the R_g plots for the MUC5B protein, MTs revealed several noteworthy discrepancies (Figure 6). Contrary to its MT, the N domain WT shows a high R_g value of 2.0 nm in the first 30 ns with significant variations before dropping abruptly to 1.1 nm and remaining unchanged until the simulation's end. The MT starts from 1.9 nm and fluctuates down to 1.3 nm in the first 10 ns. Here, it jumps up, reaching up to more than 1.6 nm, and then lowers to 1.3 nm at 20 ns. It then keeps fluctuating until 60 ns, from where it reduces more and almost becomes equal to the WT. We can see a major deviation in the structural compactness in more than 60% of the simulation in the case of the Y919D variant (Figure 5a). Contrary to the N domain variant (Figure 5b), clear differences can be seen in the case of the C domain variant (D5551Y), which stays higher than the WT protein throughout the simulation and exhibits major fluctuations, especially in the first 50 ns, unlike its native structure, which remains quite steady until the last moment of simulation with no major fluctuations. Based on these

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

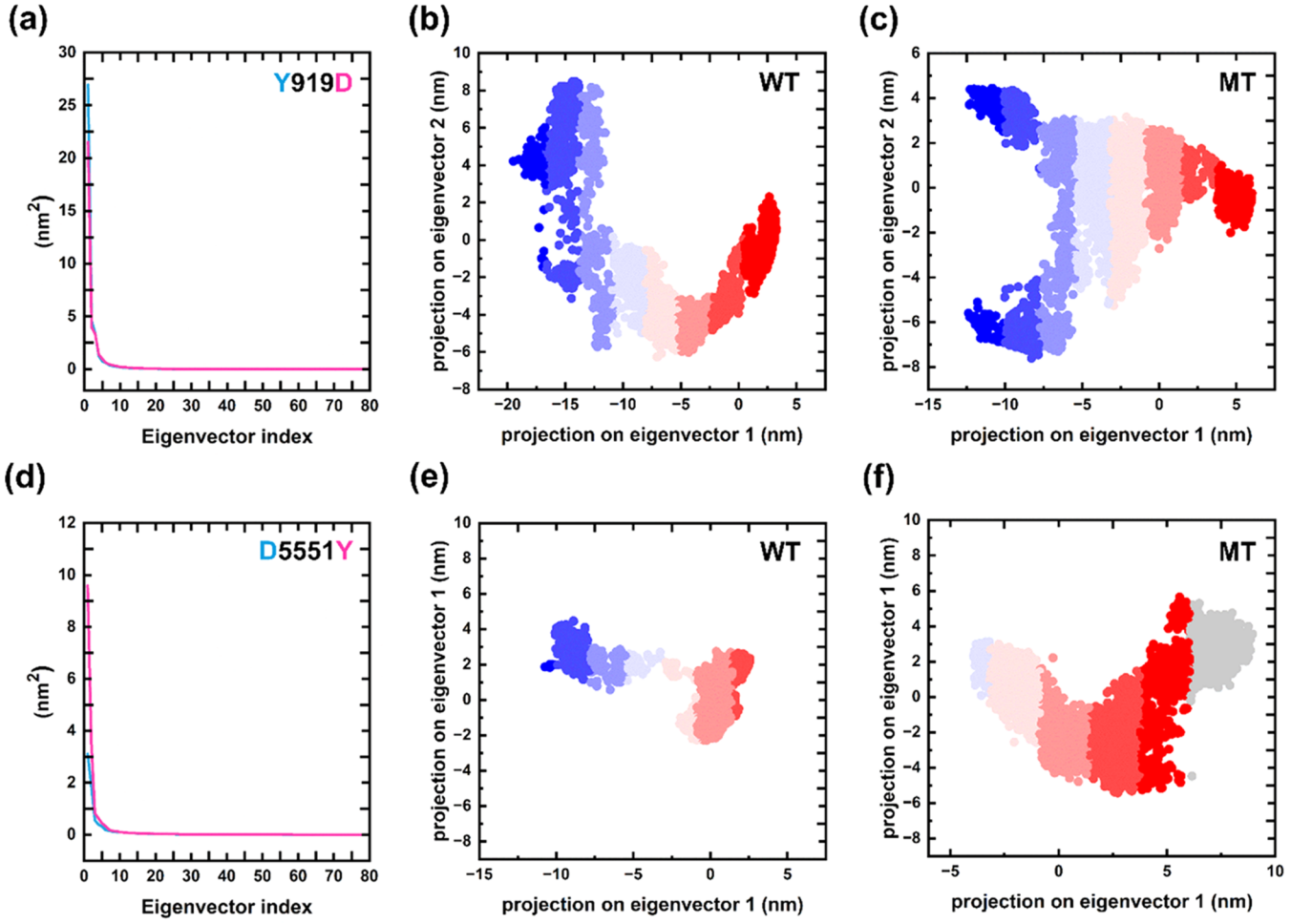


Figure 6. PCA analysis. (a) Essential dynamics of the N domain WT and MT. Two-dimensional (2D) projections of the N domain WT (b) and MT (c). Essential dynamics of the C domain. Two-dimensional projections of the N domain WT (e) and MT (f).

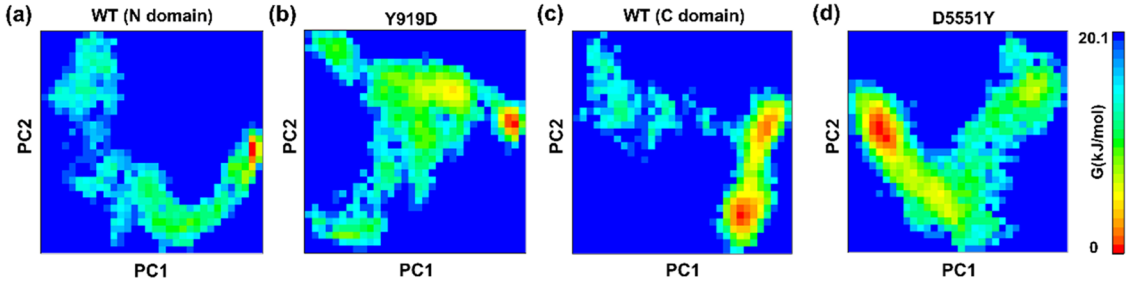


Figure 7. Gibbs free energy calculation of MUC5B. The FELs of MUC5B N domain WT (a) and MT (b) plus C domain WT (c) and MT (d) systems were determined using PC1 and PC2, the first two major components. To compute FEL, an entire trajectory with 5000 snaps was used. The overall energy is limited to 20.1 kJ/mol.

deviations, we conclude that these variants influence conformational firmness meaningfully, causing protein dispersion.

3.5.4. Principal Motions of the MUC5B N and C Domains. The essential dynamics (ED) depiction (Figure 6a) could be used to explain how the MT and WT trajectories vary. The size of the eigenvalues represents the degree of MT variations that may be caused by changes in conformation. The other eigenvectors represent more concentrated fluctuations in a single structure, but the top three eigenvectors represent maximal oscillation. The WT has a very low variance during the whole motion compared to the N and C domain variations.

To comprehend the overall frame-to-frame shifts and compactness, we also showed the 2D projections of the eigenvectors. The three-dimensional (3D) structures of the MUC5B MTs were altered, with each system exhibiting substantial modifications with various amplitudes brought about by a specific mutation that PCA discovered. Compared to the WT, all mutant structures showed deviating motion. The projections have been established for the principal components (PCs) of WT and MTs. The N domain is observed to have a uniform frame transition but suddenly scatters around the mutated regions, unlike its WT protein, which shows more compactness and solidity. The WT N

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

domain and its MT (Y919D) are given in Figure 6b and 6c, respectively. Consistent with our previous analysis, ED revealed significant differences in the case of D5551Y (Figure 6d−f). The C domain WT is much more compact, while the MT is scattered with unclear transitioning among the frames. It is noticeable that both variants extended longer than the WT on eigenvector 2, ranging from 7 nm up to almost 10 nm.

3.5.5. Gibbs Free Energy Landscape. The relationship between a protein's conformation and function could be investigated using FEL, which measures the work completed in a confined system as a consequence of heat shared with the environment. When calculating stability, alteration in the FEL values is crucial. The conformational shifts of the mutant proteins are investigated by computing the Gibbs energies for the initial two principal components (PC1 and PC2), as presented in Figure 7.

Since the lowest GFE score indicates a highly stable conformation, all of the native structures primarily achieve low energy values. Compared to the WT, MTs demonstrated a considerable variation in energy levels. The energy values for the original N domain structure varied from 0 to 17.4 kJ/mol. However, MT (Y919D) reached an energy value of 20.1, greater than the D5551Y's energy range of 0−16.2. In the case of the WT, it is possible to view the red spots that indicate the locations with the minimum energy. The rest of the MTs, however, appear to have a variety of energy minima states. According to the overall analysis, the WT seems to have more stable clusters compared to the N and C sectors' alteration, which are highly reported in CCA.

4. DISCUSSION

Structural stability is crucial for proteins to play their native roles in a complicated cellular environment. Variations in nucleotide strands are known as mutations, which can result in permanent illnesses or extreme physiological states by changing the native function of a protein or stopping the translation of a protein (mutating into a stop codon). Several studies have highlighted the role of MUC5B mutations in developing complex conditions like idiopathic pulmonary fibrosis (IPF) and rheumatoid arthritis (RA).[48] However, it is a frequently mutated gene as per the TCGA and International Cancer Genome Consortium (ICGC) data sets;[49] therefore, we explored its association with other cancers. One of the recent studies claims that significantly increased expression of MUCINs, including MUC5B, is identified in clear cell renal cell carcinoma (CCRCC) patients having worse survival.[28] Another study investigates the association between the countenance level of mucin gene clusters (MUC2, MUC5A, and MUC5B) and colorectal cancer (CRC), claiming the expression of these mucin gene clusters to be lower in CRC patients.[27] Additionally, it is observed that MUC5B is solely linked to tumor grades, being pretty advanced in poorly distinguished tumors.[27] One more study on papillary thyroid microcarcinomas found nonsynonymous mutations in different genes, including MUC5B, playing a role in cell adhesion and existing in the aggressive category only.[30]

Based on these roles of the MUC5B gene and our interest in cancer studies, we considered the TCGA data for mutational landscape analysis, revealing this gene to be one of the top ten genes in our cancer of interest (CCA). Every mutation type could somehow affect the normal molecular dogma in some way or the other. However, the role of missense mutations is quite obvious, and it directly results in the altered protein structure and, thus, the function. We can see that the ratio of missense mutations is quite higher compared to the other forms of variants and has a higher ratio of SNPs compared to the other variant types. Among other genes, we observed that MUC5B has the most significant *p*-value, which is why we selected this gene for further exploration. Other genes did not even get close to the *p*-value of MUC5B. The rest of the genes have *p*-values of 0.0984 (IDH1), 0.109 (LRP1B), 0.213 (BAP1), 0.249 (EPHA2), 0.372 (DNAH5), 0.418 (MUC16), 0.518 (ARID1A), 0.584 (CHD7), and 0.719 (PBRM1) (Figure S3). Upon plotting the mutational position on the targeted protein, we observed only two types of mutations, missense mutations and frameshift insertions, consistent with the mutational landscape (Figure 1).

We can see that except for Y919D and D5551Y, all other mutations are off the defined domains. Since a gene has particular domains essential for the function, we targeted only the mutations inside these regions to understand their impact. One of the major issues is that this protein has 5762 amino acids in total. It is challenging to simulate as a whole because it is computationally intensive and time-consuming, especially with limited computational resources. Also, the MUC5B protein's crystal structure has not been reported yet, and perhaps it is nearly impossible to homology model such a complicated and big structure. We first cut the protein into segments of interest and then incorporated the shortlisted mutations (Y919D and D5551Y) to understand their dynamics.

We compared WT and MT RMSDs of both domains and observed noticeable differences. However, we maintain that Y919D did not seem highly deleterious compared to D5551Y because although it fluctuates significantly at the start, it remains pretty stable similar to the WT. One thing unusual with our outcomes, especially in the RMSD estimation, is the higher peaks, almost double the value of commonly observed mutational impacts and MD simulation studies. According to our understanding, this is because the protein is not simulated as a whole. When a protein is in its compact form, it is highly stable compared to only a part of the protein with a long loop structure. We also predicted both domain structures using AlphaFold[50] for comparative analysis and additional validation of our findings. In our case, we observed differences between the structures predicted through Robetta and AlphaFold, e.g., an RMSD difference of 2.98 and 4.76 Å (Figure S5). The AlphaFold indicated that the N domain's structure quality is quite presentable, while the C domain has a poorly predicted structure.

Further, we conducted MD simulations to compare our results with the AlphaFold predicted structure, observing that the Robetta predicted structures were much more stable comparatively with some exceptions. Trajectory analysis (Figure S6) showed similar trends in RMSD and R_g; although there exists a difference in values, it validates our initial findings. In addition, several analyses have been conducted to assess the impact of these alterations from various perspectives. All of the results are relatively consistent and corroborate one another, including R_g, PCA, and FEL analyses. Interestingly, although the Y919D's RMSD value shows stability at the end, if observing the final frame of the WT and MT proteins (Figure S7), we can observe a much clear deviation of MT from the WT, to the level that they even do not look alike although both of these structures are obtained after a 100 ns

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

simulation and precisely superimposed. This reveals how a single-point mutation can alter the overall protein structure.

Although our study brings a comprehensive investigation of MUC5B mutational analysis and provides a wider sight of the structural changes, there are still a few issues with this study that will be fixed in our follow-up work. For instance, one should consider these mutations for mechanistic analysis once the crystal structure of MUC5B is resolved. Additionally, the mutational impact of these proteins should be studied on the entire conformation by conducting long-term simulations. It is also recommended to carry out pan-cancer analysis and explore the mutational expression of the MUC5B gene in other cancers and whether these two mutations also express in related cancers.

5. CONCLUSIONS

Based on the obtained results, we maintain that MUC5B is a frequently mutated gene in cholangiocarcinoma. Its mutations in the N and C domains' vicinity significantly alter the protein conformation. Compared to Y919D, the D5551Y is more deleterious and affects the system's stability. Further studies are required to explore these mutations in other cancers and consider MUC5B protein as a whole for additional long-term simulations and in vitro or in vivo assays.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c04871.

Top 40 differentially expressed genes in cholangiocarcinoma (Figure S1); biological process, cellular components, molecular function, and general and cell cycle pathways (Figure S2); survival analysis of the top ten frequently mutated genes (Figure S3); projection of mutations on the protein structures (Figure S4); superimposition of the Robetta and AlphaFold predicted structures (Figure S5); comparison of the RMSD and R_g of the Robetta and AlphaFold structures for the N and C domains (Figure S6); and superimposed WT and MT protein structures for the N and C domains (Figure S7) (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Aman Chandra Kaushik − *Department of Bioinformatics and Biological Statistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*; Email: amanbioinfo@sjtu.edu.cn

Dong-Qing Wei − *State Key Laboratory of Microbial Metabolism, Shanghai-Islamabad-Belgrade Joint Innovation Center on Antibacterial Resistances, Joint International Research Laboratory of Metabolic & Developmental Sciences and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200030, P. R. China; Zhongjing Research and Industrialization Institute of Chinese Medicine, Zhongguancun Scientific Park, Nanyang, Henan 473006, P. R. China; Peng Cheng Laboratory, Shenzhen, Guangdong 518055, P. R. China*; orcid.org/0000-0003-4200-7502; Email: dqwei@sjtu.edu.cn

Authors

Aamir Mehmood − *Department of Bioinformatics and Biological Statistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*; orcid.org/0000-0001-8713-966X

Sadia Nawab − *State Key Laboratory of Microbial Metabolism and School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*

Yifan Jin − *Department of Bioinformatics and Biological Statistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*; orcid.org/0000-0001-8894-7693

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c04871

Author Contributions

A.M. and A.C.K. designed the study and carried out all of the analyses. S.N. and Y.J. wrote the initial draft of the manuscript and advised on structural biology approaches. A.M wrote the final manuscript. DQW supervised the project and advised on method improvement

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

D.-Q.W. is supported by grants from the National Science Foundation of China (Grant Nos. 32070662, 61832019, 32030063), the Science and Technology Commission of Shanghai Municipality (Grant No. 19430750600), as well as the SJTU JiRLMDS Joint Research Fund and Joint Research Funds for Medical and Engineering and Scientific Research at Shanghai Jiao Tong University (YG2021ZD02). The computations were partially performed at the Pengcheng Lab. and the Center for High-Performance Computing, Shanghai Jiao Tong University.

■ REFERENCES

(1) Sarcognato, S.; Sacchi, D.; Fassan, M.; Fabris, L.; Cadamuro, M.; Zanus, G.; Cataldo, I.; Capelli, P.; Baciorri, F.; Cacciatore, M.; Guido, M. Cholangiocarcinoma. *Pathologica* **2021**, *113*, 158.

(2) Banales, J. M.; Marin, J. J.; Lamarca, A.; Rodrigues, P. M.; Khan, S. A.; Roberts, L. R.; Cardinale, V.; Carpino, G.; Andersen, J. B.; Braconi, C.; et al. Cholangiocarcinoma 2020: the next horizon in mechanisms and management. *Nat. Rev. Gastroenterol. Hepatol.* **2020**, *17*, 557−588.

(3) Hermanek, P.; Hutter, R.; Sobin, L.; Wagner, G.; Wittekind, C. Digestive System Tumours. In *TNM Atlas*; Springer, 1997; pp 71−152.

(4) Rizvi, S.; Khan, S. A.; Hallemeier, C. L.; Kelley, R. K.; Gores, G. J. Cholangiocarcinoma-evolving concepts and therapeutic strategies. *Nat. Rev. Clin. Oncol.* **2018**, *15*, 95−111.

(5) Munoz-Garrido, P.; Rodrigues, P. M. The jigsaw of dual hepatocellular−intrahepatic cholangiocarcinoma tumours. *Nat. Rev. Gastroenterol. Hepatol.* **2019**, *16*, 653−655.

(6) Brunt, E.; Aishima, S.; Clavien, P. A.; Fowler, K.; Goodman, Z.; Gores, G.; Gouw, A.; Kagen, A.; Klimstra, D.; Komuta, M.; et al. cHCC-CCA: Consensus terminology for primary liver carcinomas with both hepatocytic and cholangiocytic differentation. *Hepatology* **2018**, *68*, 113−126.

(7) Andersen, J. B.; Spee, B.; Blechacz, B. R.; Avital, I.; Komuta, M.; Barbour, A.; Conner, E. A.; Gillen, M. C.; Roskams, T.; Roberts, L. R.; et al. Genomic and genetic characterization of cholangiocarcinoma identifies therapeutic targets for tyrosine kinase inhibitors. *Gastroenterology* **2012**, *142*, 1021−1031.e15.

(8) Bertuccio, P.; Malvezzi, M.; Carioli, G.; Hashim, D.; Boffetta, P.; El-Serag, H. B.; La Vecchia, C.; Negri, E. Global trends in mortality from intrahepatic and extrahepatic cholangiocarcinoma. *J. Hepatol.* **2019**, *71*, 104−114.
(9) Lindnér, P.; Rizell, M.; Hafström, L. The impact of changed strategies for patients with cholangiocarcinoma in this millenium. *HPB Surg.* **2015**, *2015*, 2015.
(10) Kamsa-Ard, S.; Luvira, V.; Suwanrungruang, K.; Kamsa-Ard, S.; Luvira, V.; Santong, C.; Srisuk, T.; Pugkhem, A.; Bhudhisawasdi, V.; Pairojkul, C. Cholangiocarcinoma trends, incidence, and relative survival in Khon Kaen, Thailand from 1989 through 2013: a population-based cancer registry study. *J. Epidemiol.* **2018**, 197−204.
(11) Strijker, M.; Belkouz, A.; van der Geest, L. G.; van Gulik, T. M.; van Hooft, J. E.; de Meijer, V. E.; Haj Mohammad, N.; de Reuver, P. R.; Verheij, J.; de Vos-Geelen, J.; et al. Treatment and survival of resected and unresected distal cholangiocarcinoma: a nationwide study. *Acta Oncol.* **2019**, *58*, 1048−1055.
(12) Alabraba, E.; Joshi, H.; Bird, N.; Griffin, R.; Sturgess, R.; Stern, N.; Sieberhagen, C.; Cross, T.; Camenzuli, A.; Davis, R.; et al. Increased multimodality treatment options has improved survival for Hepatocellular carcinoma but poor survival for biliary tract cancers remains unchanged. *Eur. J. Surg. Oncol.* **2019**, *45*, 1660−1667.
(13) Koerkamp, B. G.; Wiggers, J. K.; Allen, P. J.; Besselink, M. G.; Blumgart, L. H.; Busch, O. R.; Coelen, R. J.; D'Angelica, M. I.; DeMatteo, R. P.; Gouma, D. J.; et al. Recurrence rate and pattern of perihilar cholangiocarcinoma after curative intent resection. *J. Am. Coll. Surg.* **2015**, *221*, 1041−1049.
(14) Cambridge, W. A.; Fairfield, C.; Powell, J. J.; Harrison, E. M.; Søreide, K.; Wigmore, S. J.; Guest, R. V. Meta-analysis and meta-regression of survival after liver transplantation for unresectable perihilar cholangiocarcinoma. *Ann. Surg.* **2021**, *273*, 240−250.
(15) Spolverato, G.; Kim, Y.; Alexandrescu, S.; Marques, H. P.; Lamelas, J.; Aldrighetti, L.; Clark Gamblin, T.; Maithel, S. K.; Pulitano, C.; Bauer, T. W.; et al. Management and outcomes of patients with recurrent intrahepatic cholangiocarcinoma following previous curative-intent surgical resection. *Ann. Surg. Oncol.* **2016**, *23*, 235−243.
(16) Verma, M.; Davidson, E. A. Mucin genes: structure, expression and regulation. *Glycoconjugate J.* **1994**, *11*, 172−179.
(17) Gendler, S. J.; Spicer, A. P. Epithelial mucin genes. *Annu. Rev. Physiol.* **1995**, *57*, 607−634.
(18) Gendler, S. J.; Lancaster, C. A.; Taylor-Papadimitriou, J.; Duhig, T.; Peat, N.; Burchell, J.; Pemberton, L.; Lalani, E.-N.; Wilson, D. Molecular cloning and expression of human tumor-associated polymorphic epithelial mucin. *J. Biol. Chem.* **1990**, *265*, 15286−15293.
(19) Gum, J. R.; Byrd, J.; Hicks, J. W.; Toribara, N.; Lamport, D.; Kim, Y. Molecular cloning of human intestinal mucin cDNAs: sequence analysis and evidence for genetic polymorphism. *J. Biol. Chem.* **1989**, *264*, 6480−6487.
(20) Gum, J. R.; Hicks, J. W.; Swallow, D. M.; Lagace, R. L.; Byrd, J. C.; Lamport, D. T.; Siddiki, B.; Kim, Y. S. Molecular cloning of cDNAs derived from a novel human intestinal mucin gene. *Biochem. Biophys. Res. Commun.* **1990**, *171*, 407−415.
(21) Porchet, N.; Van Cong, N.; Dufosse, J.; Audie, J.; Guyonnet-Duperat, V.; Gross, M.; Denis, C.; Degand, P.; Bernheim, A.; Aubert, J. Molecular cloning and chromosomal localization of a novel human tracheo-bronchial mucin cDNA containing tandemly repeated sequences of 48 base pairs. *Biochem. Biophys. Res. Commun.* **1991**, *175*, 414−422.
(22) Dufosse, J.; Porchet, N.; Audie, J.; Guyonnet Duperat, V.; Laine, A.; Van-Seuningen, I.; Marrakchi, S.; Degand, P.; Aubert, J. Degenerate 87-base-pair tandem repeats create hydrophilic/hydrophobic alternating domains in human mucin peptides mapped to 11p15. *Biochem. J.* **1993**, *293*, 329−337.
(23) Toribara, N.; Roberton, A.; Ho, S.; Kuo, W.; Gum, E.; Hicks, J.; Gum, J., Jr; Byrd, J.; Siddiki, B.; Kim, Y. Human gastric mucin. Identification of a unique species by expression cloning. *J. Biol. Chem.* **1993**, *268*, 5879−5885.
(24) Bobek, L.; Tsai, H.; Biesbrock, A. R.; Levine, M. J. Molecular cloning, sequence, and specificity of expression of the gene encoding the low molecular weight human salivary mucin (MUC7). *J. Biol. Chem.* **1993**, *268*, 20563−20569.
(25) Ho, S. B.; Roberton, A. M.; Shekels, L. L.; Lyftogt, C. T.; Niehans, G. A.; Toribara, N. W. Expression cloning of gastric mucin complementary DNA and localization of mucin gene expression. *Gastroenterology* **1995**, *109*, 735−747.
(26) Kim, Y. S.; Gum, J. R., Jr. Diversity of mucin genes, structure, function, and expression. *Gastroenterology* **1995**, *109*, 999−1001.
(27) Iranmanesh, H.; Majd, A.; Mojarad, E. N.; Zali, M. R.; Hashemi, M. Investigating the Relationship Between the Expression Level of Mucin Gene Cluster (MUC2, MUC5A, and MUC5B) and Clinicopathological Characterization of Colorectal Cancer. *Galen Med. J.* **2021**, *10*, No. e2030.
(28) Meng, H.; Jiang, X.; Huang, H.; Shen, N.; Guo, C.; Yu, C.; Yin, G.; Wang, Y. A MUCINs expression signature impacts overall survival in patients with clear cell renal cell carcinoma. *Cancer Med.* **2021**, *10*, 5823−5838.
(29) Hou, B.; Xie, H.; Piao, S.; Guo, Z.; Shan, B.; Mei, J. Identification of Muc5B mutation as a positive predictive biomarker for mTORC1/2 inhibition by ATG-008 in lung cancer. *Cancer Res.* **2022**, *82*, 4032.
(30) Song, J.; Wu, S.; Xia, X.; Wang, Y.; Fan, Y.; Yang, Z. Cell adhesion-related gene somatic mutations are enriched in aggressive papillary thyroid microcarcinomas. *J. Transl. Med.* **2018**, *16*, No. 269.
(31) Zheng, H.; Duan, W.; Zhao, Q.; Li, C.; Wang, G.; Zhang, Y.; Bai, Y.; Zhou, Y. Effect of MUC5B mutation on prognosis by enhancing the infiltration and antitumor immunity of cytotoxic T lymphocytes in the endometrial cancer. *Am. Soc. Clin. Oncol.* **2020**, *38*, No. e18105.
(32) Liu, J.; Lichtenberg, T.; Hoadley, K. A.; Poisson, L. M.; Lazar, A. J.; Cherniack, A. D.; Kovatich, A. J.; Benz, C. C.; Levine, D. A.; Lee, A. V.; et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **2018**, *173*, 400−416.e11.
(33) Colaprico, A.; Silva, T. C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T. S.; Malta, T. M.; Pagnotta, S. M.; Castiglioni, I.; et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44*, No. e71.
(34) Kaushik, A. C.; Mehmood, A.; Kumar, A.; Babu, A.; Wei, D.-Q.; Zhao, Z. Mining Cancer Cell Line-Based Drugs to Benefit KRAS (G12D) Pancreatic Adenocarcinoma Patients. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; IEEE, pp 2423−2428.
(35) Kaushik, A. C.; Mehmood, A.; Wang, X.; Wei, D.-Q.; Dai, X. Globally ncRNAs expression profiling of TNBC and screening of functional lncRNA. *Front. Bioeng. Biotechnol.* **2021**, *8*, No. 523127.
(36) Kaushik, A. C.; Mehmood, A.; Wei, D.-Q.; Dai, X. Robust Biomarker Screening Using Spares Learning Approach for Liver Cancer Prognosis. *Front. Bioeng. Biotechnol.* **2020**, *8*, No. 241.
(37) Mehmood, A.; Khan, M. T.; Kaushik, A. C.; Khan, A. S.; Irfan, M.; Wei, D.-Q. Structural dynamics behind clinical mutants of PncA-Asp12Ala, Pro54Leu, and His57Pro of Mycobacterium tuberculosis associated with pyrazinamide resistance. *Front. Bioeng. Biotechnol.* **2019**, *7*, No. 404.
(38) Wang, Q.; Mehmood, A.; Wang, H.; Xu, Q.; Xiong, Y.; Wei, D.-Q. Computational screening and analysis of lung cancer related non-synonymous single nucleotide polymorphisms on the human kirsten rat sarcoma gene. *Molecules* **2019**, *24*, 1951.
(39) Mayakonda, A.; Lin, D.-C.; Assenov, Y.; Plass, C.; Koeffler, H. P. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Res.* **2018**, *28*, 1747−1756.
(40) Allaire, J. *RStudio: Integrated Development Environment for R*; Citeseer: Boston, MA, 2012; Vol. *770*, pp 165−171.
(41) Kim, D. E.; Chivian, D.; Baker, D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* **2004**, *32*, W526−W531.

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735

(42) Sheik, S.; Sundararajan, P.; Hussain, A.; Sekar, K. Ramachandran plot on the web. *Bioinformatics* **2002**, *18*, 1548−1549.
(43) Mehmood, A.; Kaushik, A. C.; Wang, Q.; Li, C.-D.; Wei, D.-Q. Bringing structural implications and deep learning-based drug identification for KRAS mutants. *J. Chem. Inf. Model.* **2021**, *61*, 571−586.
(44) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(45) Cheatham, T. E. I.; Miller, J.; Fox, T.; Darden, T.; Kollman, P. Molecular dynamics simulations on solvated biomolecular systems: the particle mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *J. Am. Chem. Soc.* **1995**, *117*, 4193−4194.
(46) Liu, Y.; Pezeshkian, W.; Barnoud, J.; De Vries, A. H.; Marrink, S. J. Coupling coarse-grained to fine-grained models via Hamiltonian Replica Exchange. *J. Chem. Theory Comput.* **2020**, *16*, 5313−5322.
(47) Mehmood, A.; Nawab, S.; Wang, Y.; Chandra Kaushik, A.; Wei, D. Q. Discovering potent inhibitors against the Mpro of the SARS-CoV-2. A medicinal chemistry approach. *Comput. Biol. Med.* **2022**, *143*, No. 105235.
(48) Juge, P. A.; Lee, J. S.; Ebstein, E.; Furukawa, H.; Dobrinskikh, E.; Gazal, S.; Kannengiesser, C.; Ottaviani, S.; Oka, S.; Tohma, S.; Tsuchiya, N.; Rojas-Serrano, J.; Gonzalez-Perez, M. I.; Mejia, M.; Buendia-Roldan, I.; Falfan-Valencia, R.; Ambrocio-Ortiz, E.; Manali, E.; Papiris, S. A.; Karageorgas, T.; Boumpas, D.; Antoniou, K.; van Moorsel, C. H. M.; van der Vis, J.; de Man, Y. A.; Grutters, J. C.; Wang, Y.; Borie, R.; Wemeau-Stervinou, L.; Wallaert, B.; Flipo, R. M.; Nunes, H.; Valeyre, D.; Saidenberg-Kermanac'h, N.; Boissier, M. C.; Marchand-Adam, S.; Frazier, A.; Richette, P.; Allanore, Y.; Sibilia, J.; Dromer, C.; Richez, C.; Schaeverbeke, T.; Liote, H.; Thabut, G.; Nathan, N.; Amselem, S.; Soubrier, M.; Cottin, V.; Clement, A.; Deane, K.; Walts, A. D.; Fingerlin, T.; Fischer, A.; Ryu, J. H.; Matteson, E. L.; Niewold, T. B.; Assayag, D.; Gross, A.; Wolters, P.; Schwarz, M. I.; Holers, M.; Solomon, J. J.; Doyle, T.; Rosas, I. O.; Blauwendraat, C.; Nalls, M. A.; Debray, M. P.; Boileau, C.; Crestani, B.; Schwartz, D. A.; Dieude, P. MUC5B Promoter Variant and Rheumatoid Arthritis with Interstitial Lung Disease. *N. Engl. J. Med.* **2018**, *379*, 2209−2219.
(49) Peng, L.; Li, Y.; Gu, H.; Xiang, L.; Xiong, Y.; Wang, R.; Zhou, H.; Wang, J. Mucin 4 mutation is associated with tumor mutation burden and promotes antitumor immunity in colon cancer patients. *Aging* **2021**, *13*, 9043.
(50) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583−589.

https://doi.org/10.1021/acsomega.2c04871
ACS Omega 2023, 8, 3726−3735