Article

Discovery Medicine 2026; 38(214): 3542–3556
https://doi.org/10.24976/Discov.Med.202638214.270

Integrative Characterization of Germline FOXA1 Missense Variants in Asian Participants of the UK Biobank With Gastrointestinal Cancers

Shirin Hui Tan[1,2,3,*], Yan Ting Hee[4], Ah-Jung Jeon[4], Wei Hong Lai[1,2,5], Pei Jye Voon[6], Edmund Ui Hang Sim[3]

[1]Clinical Research Centre, Sarawak General Hospital, Ministry of Health Malaysia, 93586 Kuching, Sarawak, Malaysia
[2]Institute for Clinical Research, National Institutes of Health, Ministry of Health Malaysia, 40710 Shah Alam, Selangor Darul Ehsan, Malaysia
[3]Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[4]Research and Development Lab, MiRXES Lab Pte. Ltd., 138667 Singapore, Singapore
[5]Faculty of Medical and Life Sciences, Sunway University, 47500 Subang Jaya, Selangor, Malaysia
[6]Department of Radiotherapy, Oncology and Palliative Care, Sarawak General Hospital, Ministry of Health Malaysia, 93586 Kuching, Sarawak, Malaysia
*Correspondence: shirinth.crc@gmail.com (Shirin Hui Tan)
Submitted: 12 May 2026 Revised: 2 June 2026 Accepted: 24 June 2026 Published: 25 November 2026

Background: Gastrointestinal (GI) cancers are genomically heterogeneous, yet Asian populations remain underrepresented in large-scale sequencing studies, particularly with respect to germline variants. The interpretation of such variants remains challenging because of limited population-specific data and functional annotations.
Methods: We identified germline Forkhead box A1 (FOXA1) missense variants in Asian individuals with incident GI cancers using whole-exome sequencing data from the UK Biobank. Variants were prioritized based on allele frequency and functional annotation and were evaluated using an integrative structural modeling framework incorporating AlphaFold-based models, local interaction analyses, root-mean-square deviation, and *in silico* stability predictions.
Results: Four FOXA1 missense variants (p.Ala83Thr, p.Leu148Val, p.His387Arg, and p.Ser448Asn) were identified exclusively in the GI cancer subset of this cohort. Structural analyses demonstrated preservation of the overall FOXA1 fold, including the forkhead DNA-binding domain, across all variants. However, each substitution introduced localized perturbations affecting residue packing, the electrostatic environment, and interaction networks, with predicted neutral-to-mildly destabilizing ΔΔG values. Variants located in flexible regions primarily affected local interaction networks, whereas variants in structured regions induced subtle alterations without causing global conformational disruption.
Conclusion: These findings support a model in which FOXA1 missense variants contribute to localized conformational and interaction changes rather than global structural destabilization. This study highlights the utility of integrating population-scale genomics with structure-based analyses for the interpretation of germline variants in underrepresented populations.

Keywords: FOXA1; gastrointestinal cancer; whole-exome sequencing; structural modeling

Introduction

Gastrointestinal (GI) cancers constitute a major global health burden, accounting for a significant proportion of cancer incidence and cancer-related mortality worldwide [1]. Despite advancements in screening, diagnostics, and treatment, outcomes for many GI malignancies remain suboptimal, reflecting substantial biological heterogeneity in disease initiation, progression, and therapeutic response [2]. Evidence suggests that this heterogeneity arises from complex interactions among phenotypic characteristics, genetic variation, and molecular regulatory mechanisms, highlighting the need for integrative analytical approaches that extend beyond single-layer investigations [2].

Despite the increasing availability of population-scale sequencing datasets such as the UK Biobank, the interpretation of germline missense variants remains a major bottleneck in genomic research [3]. This challenge is particularly pronounced in underrepresented populations, for which variant frequencies, population-specific annotations, and functional evidence are limited. Consequently, many missense variants remain classified as variants of uncertain significance, underscoring the need for integrative analytical approaches that combine genomic data with complementary structural and functional inference frameworks.

Large population-based resources such as the UK Biobank facilitate the systematic investigation of phenotypic diversity associated with GI cancers, providing in-

Copyright: © 2026 The Author(s). Published by Discovery Medicine. This is an open access article under the CC BY 4.0 license.
Note: *Discovery Medicine* stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

sights into sociodemographic, anthropometric, biochemical, and lifestyle-related factors associated with these diseases [4,5]. Building on this phenomic foundation, whole-exome sequencing (WES) data from the same cohort provide an opportunity to investigate the role of germline coding variants, including rare and population-specific variants, in GI cancer biology [6]. However, the interpretation of missense variants remains challenging, particularly when variants are infrequent, population-specific, or underrepresented in reference datasets that are primarily composed of individuals of European ancestry [3,7]. In these contexts, variant frequency data and conventional pathogenicity prediction tools are often insufficient to determine biological significance [8].

Missense variants represent a significant proportion of rare coding variants identified through WES. In contrast to truncating mutations, which typically result in an overt loss of protein function, missense substitutions may affect protein behavior through more subtle mechanisms, such as alterations in conformation, stability, or molecular interactions [9]. Their effects are inherently difficult to infer from sequence-based information alone, particularly in underrepresented populations for which comparable data are limited [10]. Therefore, complementary analytical frameworks are required to contextualize the potential impact of missense variants identified in population-based cancer datasets.

Forkhead box A1 (FOXA1) is a pioneer transcription factor that plays a critical role in chromatin remodeling, transcriptional regulation, and epithelial lineage specification [11]. FOXA1 regulates gene expression programs essential for tissue differentiation and homeostasis by binding to condensed chromatin and facilitating access for additional transcriptional regulators [12]. Dysregulation of FOXA1 has been associated with multiple solid tumors, including GI malignancies, in which altered FOXA1 activity is associated with aberrant epithelial differentiation and transcriptional reprogramming [13,14,15]. Recent sequencing studies have identified FOXA1 missense variants in individuals with GI malignancies [16,17]. Nevertheless, the mechanistic implications of these variants remain poorly characterized.

Interpreting missense variants in transcription factors such as FOXA1 poses additional challenges because of their modular protein architecture, which generally consists of structured domains interspersed with regions of increased flexibility or intrinsic disorder [18]. Variants located outside canonical functional domains present particular interpretive challenges because their potential effects on protein conformation, stability, and interaction capabilities cannot be reliably inferred from sequence-based predictions alone [18]. This complexity underscores the need for analytical tools capable of evaluating both global protein architecture and region-specific structural features.

Recent advances in protein structure prediction, particularly the availability of high-confidence AlphaFold models, have created new opportunities to investigate the potential structural consequences of missense variants in proteins lacking experimentally determined full-length structures [19]. Structure-based analyses enable the systematic assessment of the effects of amino acid substitutions on overall protein architecture, local structural environments, residue–residue interactions, and predicted thermodynamic stability [20]. These approaches provide a complementary framework for variant interpretation, particularly for regulatory proteins in which subtle structural alterations may have functional consequences.

In this study, we extended our previous phenomic analysis of GI cancers in the UK Biobank by conducting a structure-based investigation of FOXA1 germline missense variants identified through WES in Asian individuals with GI malignancies [21]. Four FOXA1 missense variants (p.Ala83Thr, p.Leu148Val, p.His387Arg, and p.Ser448Asn) were prioritized for downstream structural analysis. Using AlphaFold-predicted FOXA1 models, we evaluated variant-associated structural perturbations through global and domain-specific root-mean-square deviation (RMSD) analyses, local structural inspection, residue–residue contact mapping, electrostatic surface analysis, and *in silico* stability prediction. This study aimed to provide structural and mechanistic insights into FOXA1 missense variants identified in an underrepresented population cohort.

Methods

Study Population

This study used data from the UK Biobank, a large prospective cohort comprising more than 500,000 participants aged 40–69 years at recruitment [4]. Participants of Asian ancestry were identified based on self-reported ethnicity in the UK Biobank dataset. In this study, "Asian" ethnicity was defined according to the UK Biobank classification and included individuals who reported Chinese, Indian, Pakistani, Bangladeshi, or other Asian backgrounds, consistent with a previous UK Biobank-based study [21]. Incident GI cancer cases were identified through national cancer registry linkages using relevant International Classification of Diseases, Tenth Revision (ICD-10) codes. GI cancers in this study included malignancies of the esophagus (C15), stomach (C16), small intestine (C17), colon (C18), rectosigmoid junction (C19), rectum (C20), anus and anal canal (C21), liver and intrahepatic bile ducts (C22), gallbladder (C23), biliary tract, including other and unspecified parts (C24), pancreas (C25), and other digestive organs (C26) [4]. Demographic, lifestyle, biochemical, and anthropometric variables were extracted from baseline assessment records.

WES data were available for more than 450,000 UK Biobank participants [5]. From this subset, WES data from Asian participants diagnosed with incident GI cancers (n = 87) were extracted and included in the variant analysis. Among these participants, 77 carried at least one FOXA1 missense variant, whereas the remaining 10 were FOXA1 wild type. The 77 variant carriers were included in downstream variant characterization and structural modeling analyses. All analyses were restricted to this genetically defined subset to evaluate FOXA1 sequence variation and its structural implications for the protein. All FOXA1 variants analyzed in this study were germline variants identified through blood-derived WES. Fig. 1 summarizes the cohort selection process and computational workflow used in this study. In addition, a total of 9268 Asian participants without a prior or incident cancer diagnosis (ICD-10 C00–C97) were included as the healthy control subset for comparisons of variant carrier proportions. Asian ancestry in both groups was defined using self-reported ethnic background (UK Biobank Field 21000) [4]. Participants were classified as Asian according to UK Biobank ethnic coding, consistent with our previous analyses [22]. No age or sex matching was performed between cases and controls, as this study was not designed as a matched case-control analysis. The control subset was included to assess the presence or absence of FOXA1 missense variants within a cancer-free Asian UK Biobank population. Relatedness filtering was not applied. However, because no FOXA1 missense variants were identified among any of the 9268 Asian healthy control participants, the inclusion of related individuals within this group would not affect the finding of variant absence in the control population. The difference in sample size between cases (n = 87) and controls (n = 9268) reflects the prevalence of incident GI cancer within the Asian UK Biobank cohort.

Baseline clinical and demographic characteristics of FOXA1 missense variant carriers (n = 77) are presented descriptively in Table 1. Continuous variables are reported as the mean and standard deviation, whereas categorical variables are reported as counts and percentages. Where applicable, percentages exceed 100% because some participants carried more than one FOXA1 variant or had more than one GI cancer diagnosis. No comparison of baseline characteristics between FOXA1 variant carriers and the Asian healthy control subset was performed because no FOXA1 missense variants were identified among healthy controls (n = 9268).

To assess whether baseline characteristics differed across FOXA1 variant subgroups within the Asian GI cancer subset, exploratory analyses were conducted among the 77 FOXA1 variant carriers. Participants carrying more than one FOXA1 variant were assigned to a primary variant subgroup according to a prespecified rarity hierarchy (H387R > L148V > S448N > A83T). Continuous variables were compared across the four variant subgroups using the Kruskal–Wallis test. Categorical variables were compared using the chi-square test. These analyses were exploratory and should be interpreted with caution given the highly imbalanced subgroup sizes and the lack of independence among variant groups resulting from the presence of multi-variant carriers.

WES Data Processing and Variant Annotation

WES data were obtained from the UK Biobank exome sequencing release, which provides centrally processed and quality-controlled variant call files generated using a standardized alignment and variant-calling pipeline [22]. For downstream analyses, population-level variant call format (pVCF) files were generated by subsetting WES data to the target genomic regions using GLnexus (version 1.0.4) on the DNAnexus platform [23]. The resulting pVCF files were annotated using SnpEff Annotate (version 2.1.2) [24]. Genotype-level filtering was applied to retain only high-quality calls, defined as genotype quality (GQ) $\geq$20 and read depth (DP) $\geq$10. Statistical analyses were conducted in R (version 4.4.1; R Foundation for Statistical Computing, Vienna, Austria; https://www.r-project.org/) using Fisher's exact test or chi-square tests to compare variant distributions between population subgroups. Multiple-testing correction was performed using the false discovery rate method.

For variants unique to a given subgroup, genotype proportions (homozygous reference, heterozygous, and homozygous alternate) were examined to assess potential enrichment. Functional annotation of retained variants was performed using the Ensembl Variant Effect Predictor [25]. Nonsynonymous coding variants, including missense, start-lost, and splice-region variants classified as having a moderate impact, were prioritized. Rare variants were defined as those with a minor allele frequency <0.01. However, subsequent structural analyses included both common and low-frequency FOXA1 missense variants identified within the cohort. Additional functional prioritization was based on a Combined Annotation Dependent Depletion (CADD) PHRED score $\geq$20 [26]. Variants were further annotated using SIFT and PolyPhen-2 predictions. Population allele frequencies were obtained directly from the Genome Aggregation Database (gnomAD) browser [27], whereas variant classification according to the American College of Medical Genetics and Genomics (ACMG) guidelines was assessed using InterVar [28,29]. These analyses were used to contextualize FOXA1 variant frequencies within the study cohort and to inform the selection of FOXA1 variants for downstream structural characterization.

FOXA1 Variant Selection and Prioritization

FOXA1 was selected for in-depth analysis because of the presence of missense variants identified among Asian participants with GI cancers and its established role as a pioneer transcription factor involved in chromatin remodeling

and epithelial differentiation. GeneCards and UniProtKB annotations further supported its biological significance in GI epithelial homeostasis and oncogenic transcriptional regulation [30,31]. Four germline FOXA1 missense variants (A83T, L148V, H387R, and S448N) were prioritized for downstream structural modeling based on their identification within the study cohort (Table 2) and their biological plausibility. Although some individuals carried more than one FOXA1 variant, phasing information was unavailable. Therefore, subsequent structural modeling was restricted to single-variant analyses.

Protein Structure Generation and Mutant Model Construction

The wild-type FOXA1 protein structure was obtained from the AlphaFold Protein Structure Database (AF-P55317-F1) [19]. Mutant FOXA1 structures were generated by introducing single amino acid substitutions into the AlphaFold-predicted wild-type structure, followed by structural refinement and visualization in PyMOL (version 3.1.6.1; Schrödinger, LLC, New York, NY, USA). Individual single-variant models were constructed for A83T, L148V, H387R, and S448N. All predicted structures were exported in Protein Data Bank format for downstream analyses. All figures were generated using PyMOL (version 3.1.6.1).

Structural Alignment and RMSD Analysis

Structural comparisons between wild-type and mutant FOXA1 models were performed using PyMOL (version 3.1.6.1). Global RMSD values were calculated using full-backbone alignment of wild-type and mutant structures with the PyMOL align function to evaluate overall conformational deviation. To minimize the influence of intrinsically disordered terminal regions, domain-specific RMSD calculations were also performed for the forkhead DNA-binding domain (residues 169–260), based on UniProt accession number P55317 (FOXA1_HUMAN), which represents the structurally conserved and functionally critical DNA-binding core [30]. Structural overlays were visually inspected to assess changes in backbone positioning, side-chain orientation, local secondary structure integrity, and surface topology.

Residue–Residue Contact Mapping

Local residue–residue interaction networks were analyzed to characterize microenvironmental rearrangements induced by the variants. For each model, residues within a 4 Å radius of the mutated site were identified using the spatial selection tools in PyMOL (version 3.1.6.1). Contacts were classified as conserved, gained, or lost through direct comparisons between wild-type and mutant structures using the same distance threshold. Contact mapping was applied to all single-variant FOXA1 models to determine whether the variants induced local compaction, loosening, or redistribution of adjacent residue interactions.

Protein Stability and Flexibility Prediction

The thermodynamic and dynamic consequences of FOXA1 missense variants were assessed using the DynaMut and DynaMut2 web servers [32,33]. Wild-type and mutant FOXA1 structures were submitted individually to estimate variant-induced changes in Gibbs free energy ($\Delta\Delta G$) and vibrational entropy ($\Delta\Delta S$). $\Delta\Delta G$ values were interpreted as relative indicators of predicted stabilizing (positive values) or destabilizing (negative values) effects, whereas $\Delta\Delta S$ values were used to infer changes in regional protein flexibility. Qualitative classifications generated by DynaMut2 (stabilizing, destabilizing, or inconclusive) were recorded and integrated with the structural modeling analyses.

Surface Accessibility, Disorder, and Domain Annotation

Residue solvent accessibility and burial were assessed using PyMOL visualization in conjunction with AlphaFold per-residue confidence scores (pLDDT). Regions with pLDDT values below 50 were considered intrinsically disordered. FOXA1 domain boundaries and known post-translational modification annotations were retrieved from UniProtKB [31]. The spatial locations of the FOXA1 variants relative to structured domains, regulatory regions, and disordered termini were mapped.

Electrostatic Surface Analysis

Electrostatic surface potential was calculated using the Adaptive Poisson–Boltzmann Solver (APBS) implemented in PyMOL. Protein structures were prepared by assigning atomic charges and radii before the calculations. Electrostatic potential maps were visualized on molecular surfaces, with identical color scales and orientations applied to wild-type and mutant structures to ensure consistent comparisons of local electrostatic changes. Electrostatic surface analysis was restricted to the H387R variant. Among the four variants examined, H387R was the only substitution predicted to alter the charge state of the affected residue at physiological pH. Histidine possesses a predominantly neutral side chain at physiological pH (pKa ~6.0–6.8), whereas arginine retains a permanent positive charge (pKa ~12.5) [34]. The remaining variants (A83T, L148V, and S448N) involve substitutions between residues that are neutral at physiological pH and, therefore, were not selected for electrostatic surface analysis.

Model Confidence and Interpretation Framework

Partial FOXA1 structures have been resolved using crystallography and cryo-electron microscopy; however, these structures represent only the forkhead DNA-binding domain in DNA- or nucleosome-bound conformations and, therefore, do not capture the full-length protein or its intrinsically disordered regulatory regions [35]. Because an experimentally determined full-length FOXA1 structure is

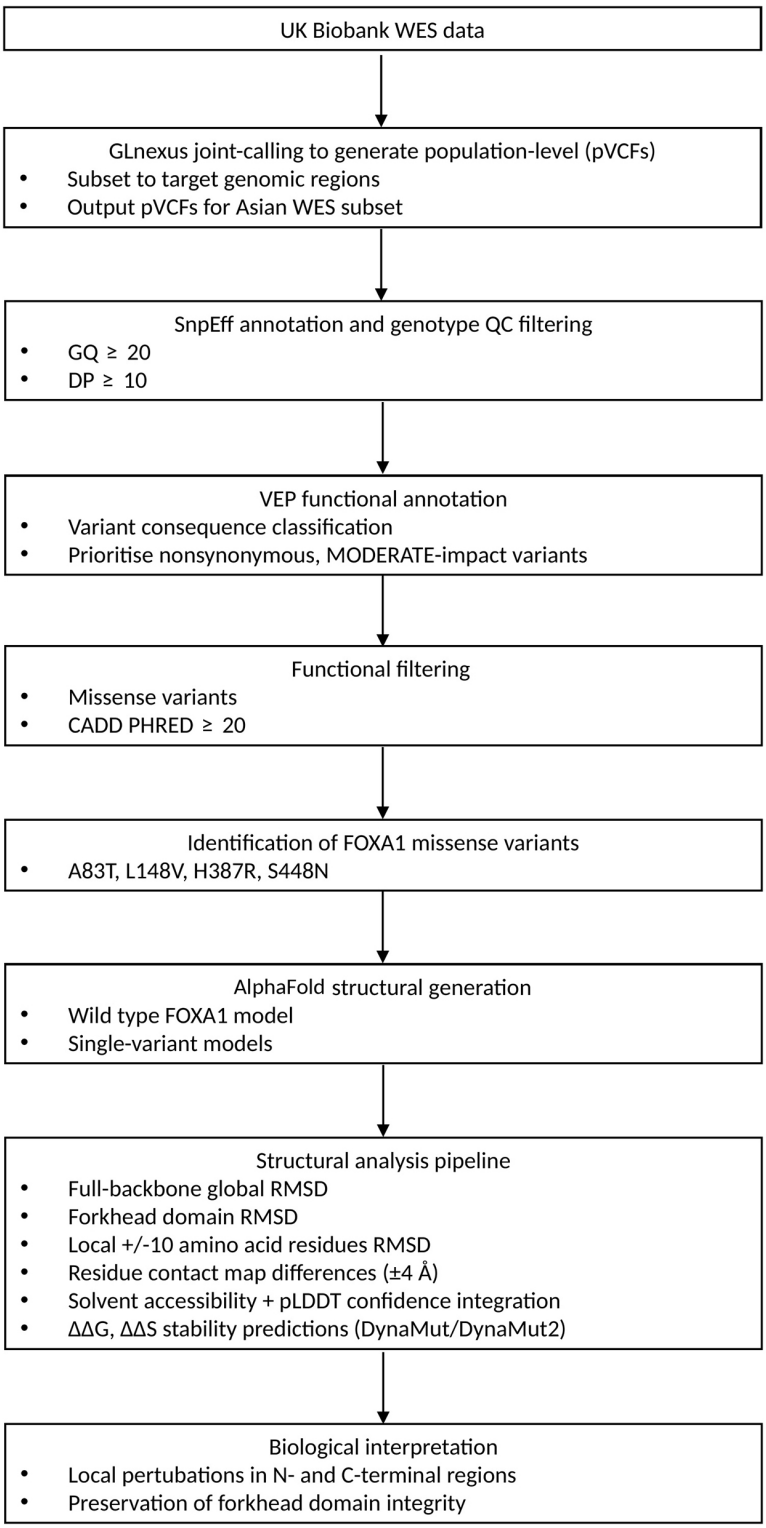


Fig. 1. Cohort selection and computational workflow for FOXA1 variant analysis. Schematic overview of the analytical pipeline used for FOXA1 variant identification, prioritization, and structural characterization. UK Biobank WES data were filtered to identify participants of self-reported Asian ancestry with incident gastrointestinal cancers (ICD-10 C15–C26). Variants underwent quality control, functional annotation, and prioritization before structural modeling of wild-type and mutant FOXA1 proteins. Downstream analyses included global and domain-specific root-mean-square deviation (RMSD) calculations, local structural perturbation assessment, residue–residue interaction analysis, and stability predictions. WES, whole-exome sequencing; GQ, genotype quality; DP, read depth; CADD, Combined Annotation Dependent Depletion; FOXA1, Forkhead box A1; ICD-10, Tenth Revision.

unavailable, all analyses were performed using AlphaFold-predicted models, which provide complete residue coverage across both ordered and disordered regions.

Structural analyses focused on domain-specific RMSD calculations and local perturbations within high-confidence regions to minimize the influence of flexible terminal segments on global structural comparisons. Multiple structural features, including RMSD measurements, residue–residue contact changes, side-chain orientation, surface accessibility, and ΔΔG stability predictions, were evaluated to characterize variant-associated structural effects.

Structural modeling was employed as a complementary analytical approach to provide mechanistic context for FOXA1 missense variants identified through WES. Given the large number of variants identified in population-scale datasets and the limited feasibility of experimentally validating each variant, integrative *in silico* approaches are increasingly used to prioritize variants and provide functional context for downstream experimental investigation.

Results

The study population selection process and analytical workflow are summarized in Fig. 1. Among UK Biobank participants with available WES data, individuals of self-reported Asian ancestry with incident GI cancers (ICD-10 C15–C26) were identified. After the application of the exclusion criteria, 77 participants carrying at least one germline FOXA1 missense variant were included in the downstream structural analyses. The computational pipeline used for variant prioritization and structural characterization is shown in Fig. 1 and outlines the sequence of quality control, functional annotation, structural modeling, and quantitative structural analyses performed in this study.

Cohort Characteristics and FOXA1 Variant Distribution

A total of 77 participants carrying one or more germline FOXA1 variants (A83T, L148V, H387R, and S448N) were identified within the Asian GI cancer subset of the UK Biobank dataset. The most prevalent variant was A83T (n = 74), followed by S448N (n = 13), L148V (n = 2), and H387R (n = 1). The marked predominance of A83T relative to the other variants indicates a highly skewed variant distribution within the cohort, with the remaining variants occurring at substantially lower frequencies. All variants were identified in the heterozygous state except A83T, for which 40 individuals were heterozygous and 34 were homozygous. As several participants carried more than one variant, the sum of variant counts exceeded the total number of carriers (Table 1). However, due to the absence of phasing information, it was not possible to determine whether these variants co-occurred on the same allele.

Table 1. Clinical and demographic characteristics of Asian participants with gastrointestinal cancer carrying FOXA1 missense variants (n = 77).

Variables		N (%)[b]
FOXA1 variants[a]		
	p.Ala83Thr (A83T)	74 (96.10)
	p.Leu148Val (L148V)	2 (2.60)
	p.His387Arg (H387R)	1 (1.30)
	p.Ser448Asn (S448N)	13 (16.88)
Combined FOXA1 variants		
	A83T and S448N	10 (12.99)
	A83T and L148V	2 (2.60)
	A83T and H387R	1 (1.30)
Sociodemographic characteristics		
	Age at cancer diagnosis, years	64.42 (7.39)
	Age at recruitment, years	57.25 (7.61)
Sex		
	Female	28 (36.36)
	Male	49 (63.64)
Ethnicity		
	Bangladeshi	2 (2.60)
	Chinese	17 (22.08)
	Indian	35 (45.45)
	Pakistani	13 (16.88)
	Other Asian background	10 (12.99)
GI cancer subtype[a]		
	Esophageal	4 (5.19)
	Gastric	9 (11.69)
	Small intestine	1 (1.30)
	Colorectal	43 (55.84)
	Liver and intrahepatic bile ducts	9 (11.69)
	Gallbladder	3 (3.90)
	Biliary tract	5 (6.49)
	Pancreatic	7 (9.09)
Smoking status		
	Never	55 (72.37)
	Previous	12 (15.79)
	Current	9 (11.84)
Pack-years of smoking (ever smokers)		20.40 (9.18)
Alcohol consumption		
	Never	25 (32.47)
	Previous	3 (3.90)
	Current	49 (63.64)
Body mass index, kg/m^2		26.98 (4.453)
BMI category		
	Underweight	1 (1.35)
	Normal weight	8 (10.53)
	Overweight	16 (21.05)
	Obese	51 (67.11)

Table 1. Continued.

Variables		N (%)[b]
Waist circumference, cm		92.13 (13.55)
Hip circumference, cm		101.0 (9.079)
Biochemical parameters		
	Cystatin C, mg/L	0.9890 (0.2565)
	Total cholesterol, mmol/L	5.253 (1.147)
	HDL cholesterol, mmol/L	1.264 (0.3748)
	LDL cholesterol, mmol/L	3.256 (0.8081)
	Apolipoprotein A1, g/L	1.409 (0.2516)
	Apolipoprotein B, g/L	0.9736 (0.2184)

[a] Percentages may exceed 100% because participants may carry more than one FOXA1 variant and/or have multiple GI cancer diagnoses. [b] Continuous variables are presented as mean (standard deviation), and categorical variables are presented as n (%). HDL, high-density lipoprotein; LDL, low-density lipoprotein.

Carriers were predominantly male (63.64%), with a mean age at cancer diagnosis of 64.42 years (SD 7.39) and a mean age at recruitment of 57.25 years (SD 7.61). Most participants were overweight or obese (88.16%), with a mean body mass index (BMI) of 26.98 kg/m^2 (SD 4.453). The majority reported never smoking (72.37%), whereas current alcohol consumption was common (63.64%). FOXA1 variant carriers were distributed across several Asian ethnic subgroups, including Indian (45.45%), Chinese (22.08%), Pakistani (16.88%), Bangladeshi (2.60%), and other Asian backgrounds (12.99%). No statistically significant difference in variant frequency was observed across ethnic groups (χ^2 = 9.04, df = 12, p = 0.70), with A83T consistently identified as the most common variant in all subgroups. These findings suggest that the observed FOXA1 variant distribution was consistent across Asian subgroups within the cohort. The demographic and clinical characteristics of FOXA1 variant carriers among Asian GI cancer cases are presented in Table 1.

In exploratory analyses conducted within the FOXA1 variant carrier subset, baseline biochemical and lifestyle parameters, including BMI, lipid profiles, inflammatory markers, smoking status, and alcohol consumption, did not differ significantly across the FOXA1 variant groups (p > 0.1 for all variables tested using the Kruskal–Wallis test and chi-square test, where applicable). These comparisons should be interpreted with caution given the highly imbalanced subgroup sizes, the lack of independence among variant groups resulting from multi-variant carriers, and the limited statistical power of the study.

Variant-level frequency comparisons between Asian GI cancer cases (n = 87) and Asian healthy controls (n = 9268) were performed using Fisher's exact test. For three of the four variants (A83T, S448N, and L148V), carrier status in the Asian healthy control subset could not be independently verified with the dataset available thus far. These findings should therefore be interpreted with caution. For H387R, a rare variant (gnomAD AF = 1.25 × 10^{-6}), zero carriers among 9268 controls is consistent with population expectations. Given this uncertainty and the limited sample size, these comparisons are interpreted descriptively rather than as evidence of association. The clinical and demographic characteristics of FOXA1 variant carriers are summarized in Table 1, whereas the distribution of FOXA1 missense variants across GI cancer subtypes is presented in Table 2. The genetic and *in silico* annotations of FOXA1 missense variants, including predicted pathogenicity metrics, population allele frequencies, and ACMG classifications, are summarized in Table 3.

GI Cancer Spectrum Among FOXA1 Variant Carriers

Among FOXA1 variant carriers (n = 77), germline missense variants identified through blood-derived WES were observed in individuals diagnosed with GI cancers (ICD-10 C15–C26) spanning the entire GI tract. Within this carrier group, colorectal cancer (CRC; C18–C20) was the most frequent cancer type (n = 43, 55.8%), followed by gastric cancer (C16; n = 9, 11.7%), liver and intrahepatic bile duct cancers (C22; n = 9, 11.7%), and pancreatic cancer (C25; n = 7, 9.1%). Lower frequencies were observed for esophageal (C15), small intestinal (C17), gallbladder (C23), and biliary tract (C24) cancers.

These distributions reflect the underlying composition of GI cancer cases within the Asian UK Biobank cohort and should be interpreted descriptively rather than as evidence of tumor-type enrichment. When stratified by variant, A83T exhibited the broadest distribution across GI cancer subtypes. The highest frequency was observed in CRC (55.4%), followed by liver and intrahepatic bile duct cancer (12.2%), gastric cancer (10.8%), and pancreatic cancer (9.5%). Lower frequencies were observed for biliary tract (6.7%), esophageal (4.1%), gallbladder (4.1%), and small intestinal cancers (1.3%) (Table 2). S448N was also detected across multiple anatomical sites, including the colon, stomach, rectum, and liver, consistent with its occurrence in both upper and lower GI malignancies. L148V and H387R were infrequently observed and were identified in individuals with gastric, colorectal, or hepatobiliary cancers. These site distributions are presented descriptively and should not be interpreted as evidence of variant-specific cancer subtype associations. This consideration is particularly important given the limited numbers of carriers with L148V (n = 2) and H387R (n = 1).

Table 2. Distribution of FOXA1 missense variants across gastrointestinal cancer subtypes.

FOXA1 variants	Total carriers (n)	Gastrointestinal cancer subtypes, n (%) [a]							
		Esophageal	Gastric	Small intestine	Colorectal	Liver and intrahepatic bile ducts	Gallbladder	Biliary tract	Pancreatic
A83T	74	3 (4.1)	8 (10.8)	1 (1.3)	41 (55.4)	9 (12.2)	3 (4.1)	5 (6.7)	7 (9.5)
S448N	13	1 (7.7)	3 (23.1)	0	6 (46.2)	2 (15.4)	0	0	2 (15.4)
L148V	2	0	0	0	1 (50.0)	1 (50.0)	0	0	0
H387R	1	0	1 (100)	0	0	0	0	0	0

[a] Percentages represent the proportion of participants within each FOXA1 variant carrier group.

Table 3. Genetic and *in silico* evidence for FOXA1 missense variants.

Variant	Exon	GI case carriers (n)	Healthy carriers (n)	GI case carrier frequency	Healthy carrier frequency	gnomAD AF	CADD	SIFT	PolyPhen-2	ACMG (InterVar) interpretation
p.Ala83Thr	2	74	ND / ND	0.851	ND / ND	0.596	16.0	Tolerated	Benign	Benign
p.Ser448Asn	2	13	ND / ND	0.149	ND / ND	0.0798	20.5	Tolerated	Benign	Benign
p.Leu148Val	2	2	ND / ND	0.023	ND / ND	0.0173	19.3	Tolerated	Benign	VUS
p.His387Arg	2	1	0	0.011	0	1.25×10^{-6}	25.7	Deleterious	Damaging	VUS

Abbreviations: AF, allele frequency; ACMG, American College of Medical Genetics and Genomics; GI, gastrointestinal; VUS, variant of uncertain significance; ND, not independently verified.

Notes: Population allele frequencies were obtained from gnomAD. Population frequency categories were defined as common (AF >0.05), low-frequency (AF 0.01–0.05), and rare (AF <0.01). Carrier status for A83T, S448N, and L148V in the Asian control subset of the UK Biobank cohort could not be independently verified with the dataset available thus far. The zero-carrier value for H387R is consistent with its rarity in reference populations. GI case carrier frequency represents the proportion of Asian participants with GI cancer (n = 87) carrying each variant, using the total number of Asian GI cancer cases as the denominator.

Multi-Variant FOXA1 Carriers

In this study, a subgroup of participants carrying multiple FOXA1 variants was identified (n = 13), indicating the presence of more than one germline FOXA1 nonsynonymous variant within the same individual. Among these multi-variant FOXA1 carriers, A83T was present in all cases (100%), whereas S448N occurred in 10 participants (76.9%), L148V in 2 participants (15.4%), and H387R in 1 participant (7.7%). This pattern suggests that A83T may represent a common background FOXA1 variant, with S448N, L148V, and H387R occurring as additional co-existing variants.

CRC was the most common GI cancer diagnosis within the multi-variant subgroup, with 5 of 13 participants (38.5%) carrying multiple FOXA1 variants. Pancreatic cancer (C25) was identified in 2 of 13 participants (15.4%) with multiple FOXA1 variants. No multi-variant cases were identified among participants with cancers of the esophagus (C15), small intestine (C17), anus (C21), gallbladder (C23), or biliary tract (C24). Overall, the tumor distribution observed among multi-variant carriers mirrored the predominance of CRC seen among single-variant carriers, although a higher proportion of pancreatic cancers was observed in the multi-variant subgroup.

Structural Analyses of FOXA1 Variants

To interpret the potential structural consequences of FOXA1 missense variants identified through WES, structural modeling was performed to evaluate variant-associated effects on protein architecture and local interaction environments. Fig. 2 illustrates the domain organization of the FOXA1 protein and the positions of the identified missense variants. The forkhead DNA-binding domain (residues 169–260) corresponds to a well-structured region, as indicated by high pLDDT, whereas the N- and C-terminal regions exhibit lower confidence scores, consistent with intrinsic disorder. All four germline FOXA1 variants are located outside the core DNA-binding domain, although L148V lies in close proximity to this region. The spatial distribution of these variants within the full-length FOXA1 structure is illustrated in Fig. 3A.

Structural Consequences of the FOXA1 A83T Variant

Structural superposition of the wild-type and A83T FOXA1 models demonstrated preservation of the overall fold, with no displacement of secondary structural elements. Residue 83 is located within the N-terminal region of FOXA1. Substitution of alanine with threonine introduced a polar hydroxyl group and altered the local side-chain orientation (Fig. 3B). Local contact analysis within a 4 Å radius revealed increased residue interactions surround-

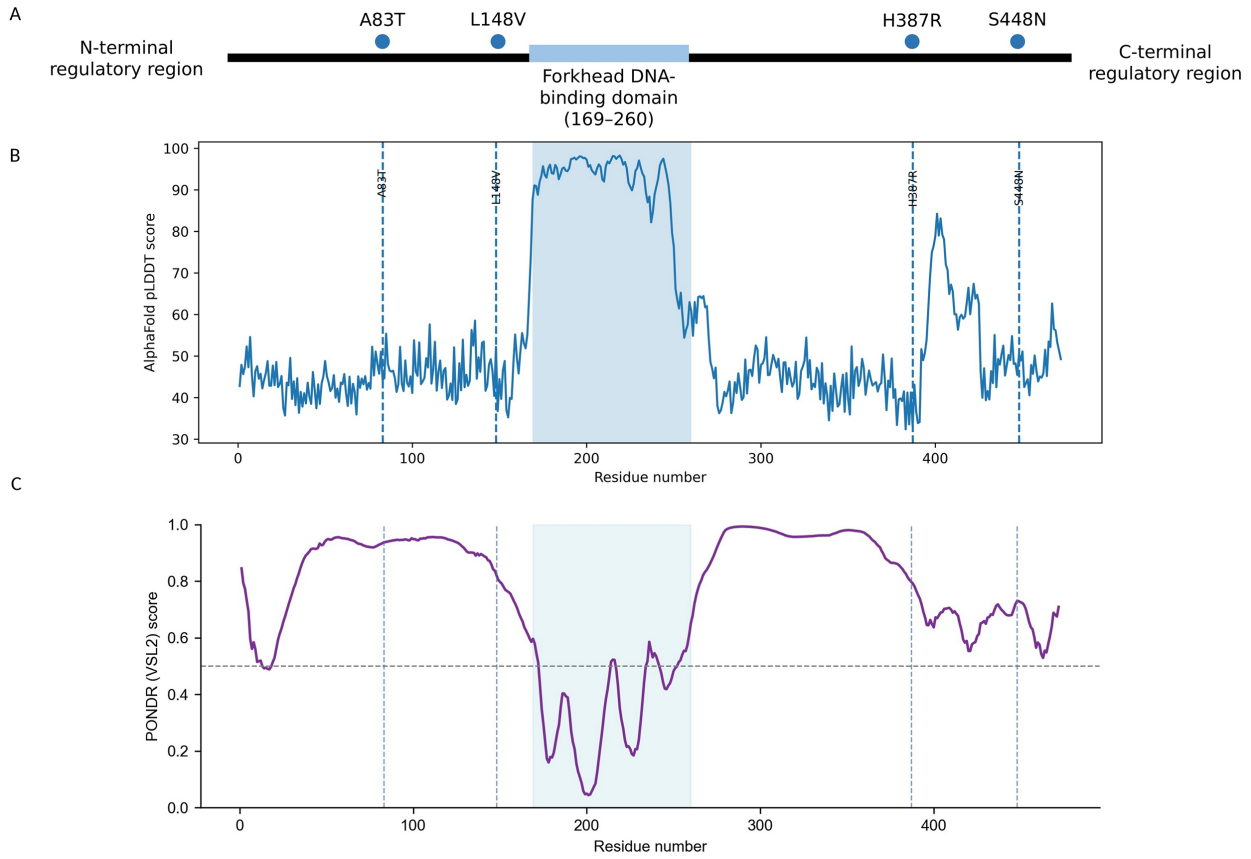


Fig. 2. Integrated structural confidence and intrinsic disorder profiling of FOXA1. (A) Linear representation of FOXA1 (472 amino acids [aa]) showing the locations of four missense variants (A83T, L148V, H387R, and S448N) identified in Asian GI cancer cases. The forkhead DNA-binding domain (residues 169–260) is highlighted, together with the flanking N- and C-terminal regulatory regions. (B) AlphaFold per-residue confidence (pLDDT) scores for the wild-type FOXA1 model (AF-P55317-F1). High confidence scores are observed across the forkhead DNA-binding domain, whereas lower confidence scores are evident in the terminal regions. (C) Intrinsic disorder prediction using PONDR VSL2, showing extensive disorder in the N- and C-terminal regions (scores >0.5) and reduced disorder within the forkhead domain. Variant positions are indicated in all panels for direct comparison.

ing the mutant residue. No steric clashes or long-range structural distortions were observed in the A83T model (Fig. 3B). These findings are consistent with localized modulation of interaction potential within a flexible regulatory region rather than with structural destabilization.

Structural Consequences of the FOXA1 S448N Variant

Structural superposition of the wild-type and S448N FOXA1 models demonstrated preservation of the overall tertiary structure. Residue 448 is located within the C-terminal region of FOXA1. Substitution of serine with asparagine introduced a larger polar side chain and altered the local side-chain orientation.

Local contact analysis revealed an increase in short-range interactions surrounding residue 448 in the mutant model compared with the wild-type structure. These changes were confined to the C-terminal region, with no propagation of structural deviations into the central folded core (Fig. 3E).

Structural Consequences of the FOXA1 L148V Variant

Comparison of the wild-type FOXA1 and L148V models revealed minimal global conformational variation, with preservation of backbone alignment. Residue 148 is positioned within a well-defined region of the protein and is surface-accessible in the wild-type model.

Substitution of leucine with the smaller valine residue reduced side-chain volume and resulted in local repacking of neighboring hydrophobic residues. No steric clashes, disruption of secondary structure, or long-range conformational shifts were observed in the mutant model. The local structural effects associated with the L148V substitution are illustrated in Fig. 3C. These findings suggest conservative hydrophobic repacking without disruption of structural integrity.

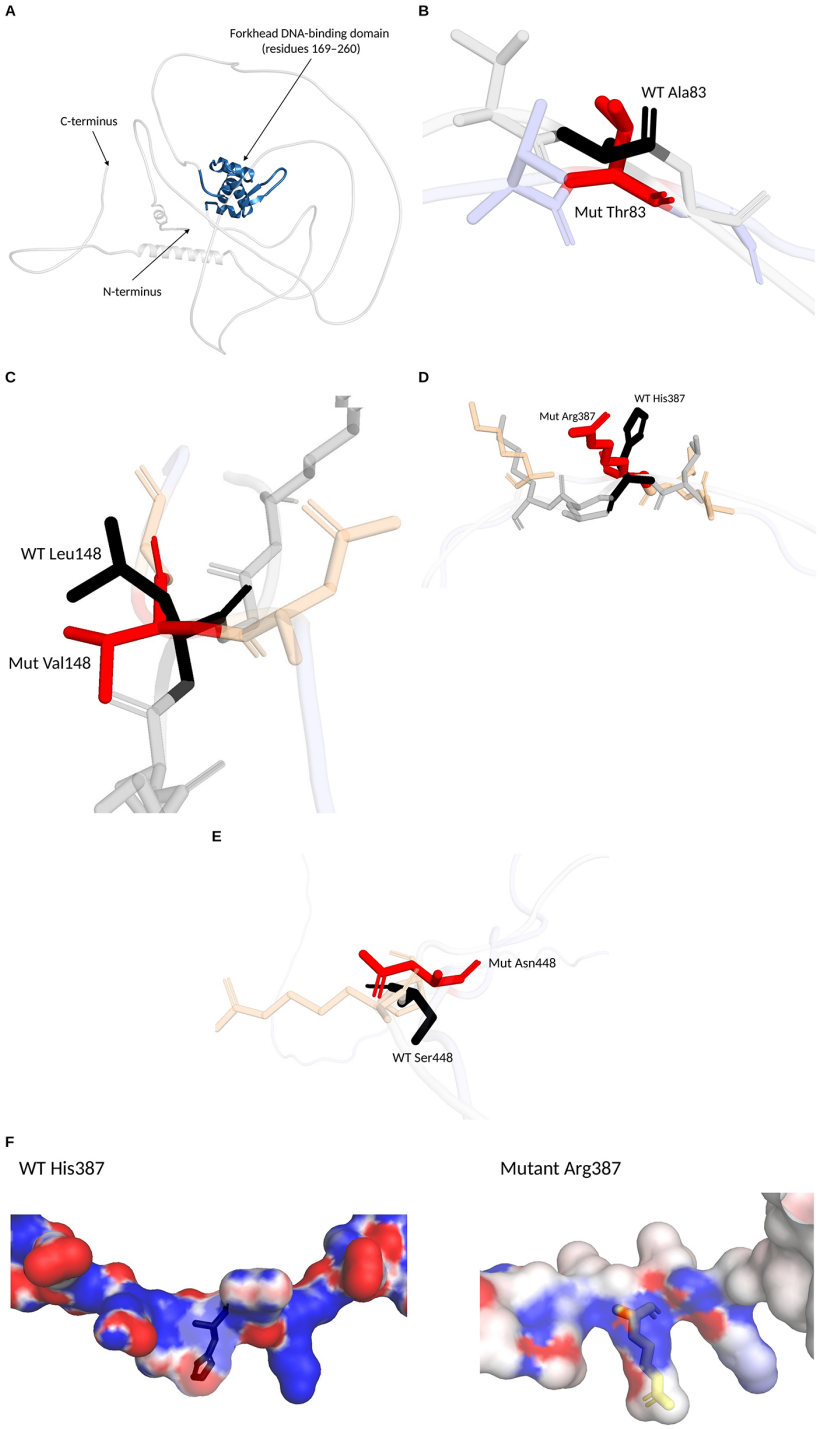


Fig. 3. Structural modeling of FOXA1 germline missense variants. (A) Overview of the AlphaFold-predicted full-length FOXA1 structure. The forkhead DNA-binding domain (residues 169–260) is highlighted in blue, whereas the remaining regions are shown in light gray. The N- and C-termini are labeled to indicate the overall protein orientation. (B–E) Local structural comparison of wild-type and mutant residues for four FOXA1 missense variants: A83T (B), L148V (C), H387R (D), and S448N (E). In each panel, the wild-type residue is shown in black and the mutant residue in red. Surrounding residues within a 4 Å radius of the variant site are displayed as sticks (light gray) to illustrate the local structural environment. These views highlight local packing and interaction changes without disrupting the overall protein fold. All structures were generated using AlphaFold models and visualized using PyMOL. (F) Electrostatic surface representation of the H387R variant. Wild-type (His387) and mutant (Arg387) structures are shown side by side. Electrostatic potentials were calculated using APBS, with red indicating negative charge and blue indicating positive charge. The variant site is highlighted to illustrate differences in local electrostatic distribution between the wild-type and mutant structures.

Table 4. Structural modeling and stability analyses of FOXA1 missense variants.

Variant	Global RMSD (Å)	Forkhead domain RMSD (Å)	DynaMut2 ΔΔG (kcal/mol)	DynaMut ΔΔG (kcal/mol)	ΔΔS (kcal/-mol/K)	Predicted structural interpretation
p.Ala83Thr	39.83	1.03	−0.41	−0.20	−0.484	Mild destabilization with reduced local flexibility
p.Leu148Val	37.98	1.09	−0.06	0.002	−0.007	Near-neutral structural effect with minimal flexibility change
p.His387Arg	38.45	1.10	0.25	−0.147	0.029	Predominantly near-neutral energetic effect with minimal destabilization
p.Ser448Asn	51.40	1.10	−0.24	−0.687	0.587	Mild destabilization with increased molecular flexibility

Abbreviations: RMSD, root-mean-square deviation; ΔΔG, predicted change in Gibbs free energy; ΔΔS, predicted change in vibrational entropy.

Notes: Global RMSD values reflect structural differences across the full-length FOXA1 model, whereas forkhead domain RMSD values were calculated specifically for the DNA-binding domain (residues 169–260). Negative ΔΔG values indicate predicted destabilizing effects, whereas positive ΔΔG values indicate stabilizing effects. ΔΔS values were used to infer relative changes in local molecular flexibility.

Structural Consequences of the FOXA1 H387R Variant

Structural alignment of the wild-type FOXA1 and H387R models demonstrated preservation of the overall protein structure, including the forkhead DNA-binding domain. Residue 387 is located outside the forkhead DNA-binding domain within a loop-proximal region of the C-terminal portion of FOXA1 (Fig. 3D). Among the four variants examined, H387R was the only substitution selected for electrostatic surface analysis because it was the only variant predicted to alter the charge state of the affected residue at physiological pH [34]. Replacement of histidine with arginine introduces a permanent positive charge, resulting in local electrostatic remodeling at the protein surface without affecting the core DNA-binding architecture (Fig. 3F). Structural modeling demonstrated preservation of the forkhead fold, with a low domain-specific RMSD value (1.10 Å; Table 4), indicating that this substitution does not induce structural perturbation at the domain level.

Nevertheless, the electrostatic properties of the protein surface in this region were altered. These findings suggest that H387R primarily induces electrostatic remodeling within a structurally constrained region of the protein. Although this study did not directly evaluate FOXA1–DNA binding interactions, the structural location and physicochemical properties of the substitution suggest that the local electrostatic environment is modified without compromising domain architecture.

Global, Domain-Specific, and Local RMSD Analyses

Global RMSD values calculated across full-length FOXA1 models ranged from 37.98 to 51.40 Å (Table 4). These elevated values likely reflect conformational variability within the intrinsically disordered N- and C-terminal regions and differences in the alignment of flexible segments rather than major alterations of the structured protein core. In contrast, RMSD values calculated specifically for the forkhead DNA-binding domain (residues 169–260) remained consistently low across all variants (1.03–1.10 Å), supporting preservation of the structured core domain.

Predicted Stability and Flexibility Changes

Stability predictions generated using DynaMut2 indicated predominantly mild destabilizing or near-neutral ΔΔG effects across FOXA1 variants, with values ranging from −0.41 to −0.06 kcal/mol (Table 4). Variant-specific differences in flexibility were observed (Table 4), with S448N showing increased flexibility (ΔΔS = 0.587 kcal/mol/K), whereas A83T demonstrated reduced flexibility (ΔΔS = −0.484 kcal/mol/K). L148V exhibited minimal energetic change, consistent with a largely neutral structural effect. In contrast, H387R showed discordant stability predictions between the two prediction tools, although the magnitude of the predicted changes was small, suggesting an overall minimal or context-dependent effect on protein stability.

Discussion

In this analysis of the Asian GI cancer subset of the UK Biobank, four germline *FOXA1* missense variants (A83T, L148V, H387R, and S448N) were identified in individuals diagnosed with GI cancers. The predominance of A83T among carriers is consistent with its high background frequency in the general population (gnomAD AF = 0.596) and does not, in itself, indicate cancer-specific enrichment. S448N, a low-frequency variant (gnomAD AF = 0.0798), was also observed predominantly in individuals with CRC, although causal relationships cannot be inferred from these data (Table 2). Through AlphaFold-based structural modeling and detailed local interaction analyses, we demonstrated that all four variants induced localized microenvironmental changes without disrupting the global FOXA1 fold. Col-

lectively, these findings support a model in which FOXA1 variants induce region-specific structural alterations rather than global destabilization.

Across all models, FOXA1 maintained its overall tertiary structure, including the integrity of the forkhead DNA-binding domain and α-helical core. No significant unfolding, domain displacement, or gross conformational distortion was observed. Domain-specific RMSD values remained consistently low despite elevated global RMSD measurements, suggesting that the greater conformational variability primarily reflects flexibility within the intrinsically disordered terminal regions rather than destabilization of the structured DNA-binding core. These findings indicate that the identified substitutions do not abolish FOXA1 structural integrity but instead induce subtle local perturbations. This structural preservation is consistent with the functional requirements of transcription factors, which often maintain stable structural cores while permitting flexibility within regulatory regions [36]. Several FOXA1 variants identified in this study, including A83T and S448N, are common in population databases, suggesting that these variants are unlikely to represent highly penetrant pathogenic mutations when considered individually. Nevertheless, common and low-frequency missense variants may still influence local conformational dynamics, protein interactions, or regulatory activity within specific genomic or population contexts. In contrast, H387R represents a genuinely rare variant that demonstrated localized electrostatic remodeling without major perturbation of the FOXA1 structural core.

The four FOXA1 variants differ substantially in population allele frequency and ACMG classification (Table 3). A83T (gnomAD AF = 0.596) and S448N (gnomAD AF = 0.0798) are both common variants and were assigned benign ACMG classifications by InterVar. These variants are unlikely to represent individually penetrant pathogenic drivers. L148V (gnomAD AF = 0.0173) is a low-frequency variant classified as a VUS. H387R (gnomAD AF = 1.25×10^{-6}) is a rare variant that is also classified as a VUS. The higher CADD score for H387R (25.7), together with its deleterious SIFT and damaging PolyPhen-2 predictions, suggests a greater predicted functional impact. Nonetheless, the numbers of carriers with L148V (n = 2) and H387R (n = 1) were very small. Therefore, the structural findings for all four variants should be interpreted as hypothesis-generating, irrespective of population frequency category.

The A83T substitution occurs within the N-terminal intrinsically disordered regulatory region of FOXA1. Replacement of alanine with threonine introduces a polar hydroxyl-containing side chain, resulting in altered local side-chain orientation and increased local interaction potential. Contact mapping demonstrated an increase in residue interactions within a 4 Å radius surrounding the mutant residue relative to the wild-type structure. These findings suggest that A83T induces localized reorganization of the microenvironment, characterized by increased local polarity and hydrogen-bonding potential without evidence of regional unfolding or structural destabilization.

The S448N variant is located within the intrinsically flexible C-terminal region of FOXA1. Replacement of serine with asparagine introduces a larger polar side chain with increased hydrogen-bonding potential. Structural analyses demonstrated increased local residue contacts and enhanced short-range interactions surrounding the mutant residue, suggesting localized compaction of the interaction network without disruption of the overall fold. Notably, serine 448 has been reported as a potential phosphorylation site in FOXA1 [37], raising the possibility that this substitution may influence post-translational regulation despite preservation of overall structural integrity.

The L148V substitution occurs within a structured region of FOXA1 and constitutes a conservative hydrophobic-to-hydrophobic amino acid replacement. Although leucine and valine possess similar physicochemical properties, the shorter side chain of valine altered the local hydrophobic packing geometry. Structural modeling revealed subtle repacking of neighboring residues without disruption of secondary structure or global conformation. These findings are consistent with a conservative local structural adjustment rather than structural destabilization.

H387R is located outside the forkhead DNA-binding domain within a loop-proximal region of the C-terminal portion of FOXA1. Replacement of histidine with arginine introduces a permanent positive charge, resulting in local electrostatic remodeling at the protein surface without affecting the core DNA-binding architecture. Domain-specific RMSD values remained low (1.10 Å; Table 4), indicating preservation of the structured core domain. Although the functional consequences of this electrostatic alteration were not directly assessed, these findings suggest localized modification of the FOXA1 surface environment while preserving overall structural integrity.

Taken together, these findings indicate that the FOXA1 variants identified in this cohort predominantly induce localized structural perturbations rather than global conformational disruption. Such subtle structural effects are consistent with structure–function relationships observed in transcription factors, in which local changes may influence regulatory interactions without abolishing the overall architecture [38]. These findings are also consistent with previous evidence demonstrating that FOXA1 variants can alter pioneering activity and transcriptional programs [39].

FOXA1 acts as a pioneer transcription factor that modulates chromatin accessibility and regulates epithelial lineage differentiation in GI tissues [36,39,40]. Although transcriptional activity was not directly evaluated in this study, the observed structural perturbations predominantly affected flexible regulatory regions and surface-exposed interfaces involved in protein–protein interactions rather than

the structured DNA-binding core. Therefore, the structural findings support a model in which FOXA1 variants may influence regulatory interface dynamics without abolishing structural integrity or DNA-binding capacity. Such subtle structural modulation is increasingly recognized as a mechanism contributing to transcriptional heterogeneity in cancer [39,40,41]. These effects may involve altered interaction networks, including oncogenic protein–protein interactions [42].

In this study, FOXA1 germline missense variants were analyzed in a subset of Asian participants diagnosed with GI cancers in the UK Biobank. Although several variants are present in the broader population databases at varying frequencies, carrier status for A83T, S448N, and L148V in the Asian healthy control subset reflects the variant-calling approach used in this study and could not be independently reconciled with their population allele frequencies prior to publication. These findings postulate a mechanistic framework for understanding how FOXA1 germline missense variants may influence transcription factor behavior in GI epithelial tumors. Larger ancestry-diverse sequencing studies incorporating functional validation will be required to determine the biological and clinical relevance of these variants across populations.

This study was based on *in silico* structural modeling, energetic assessment, and static contact network analyses. Computational structural approaches are increasingly used for variant interpretation in large-scale genomic studies, particularly when experimental validation of individual variants is not feasible [38,43,44,45]. Accordingly, these findings should be interpreted as hypothesis-generating rather than as definitive evidence of pathogenicity. This caution is particularly relevant for A83T, which is a common polymorphism (gnomAD AF = 0.596) and for which the observed structural perturbations are unlikely to represent pathogenic mechanisms. The analyses focused on static equilibrium structures and localized structural perturbations and did not include molecular dynamics simulations or experimental biochemical validation. In addition, the limited number of Asian participants with GI cancer and available WES data restricted the statistical power for population-level inference. The current study was not designed or powered to establish statistical associations or estimate cancer risk. However, this limitation does not affect the validity of the structural insights, which were derived from established biophysical modeling approaches. Phasing information was also unavailable, precluding determination of whether FOXA1 variants occurred on the same allele. Additionally, the absence of confirmed carrier counts for A83T, S448N, and L148V in the Asian healthy control subset (n = 9268) reflects the site-unification logic of the GLnexus joint-calling pipeline used in this study, in which a genomic position is retained as a called site only if at least one sample within that batch carries a non-reference allele. Positions with no non-reference calls in a given batch are therefore not represented in the output. While this is consistent with a true absence of carriers, the population allele frequencies of these three variants make this outcome difficult to reconcile at face value, and could not be independently verified with the dataset available thus far, as this would require direct access to the UK Biobank Research Analysis Platform. These findings should therefore be interpreted with caution. This does not apply to H387R, a rare variant (gnomAD AF = 1.25×10^{-6}) for which zero carriers is consistent with population expectations. Future studies incorporating experimental approaches such as electrophoretic mobility shift assays and protein stability measurements (e.g., isothermal titration calorimetry) will be important for validating these predictions [46,47].

Conclusion

In conclusion, AlphaFold-based structural modeling demonstrated that FOXA1 germline missense variants identified in Asian individuals with GI cancers preserved the overall FOXA1 fold while inducing localized perturbations that depended on residue position and structural context. These findings support a model in which FOXA1 variants may influence transcriptional regulation through subtle region-specific conformational modulation rather than overt structural destabilization. This study further highlights the utility of integrating population-scale genomics with structural modeling approaches to contextualize germline variants identified in underrepresented populations. Moreover, this study addresses an underinvestigated area by characterizing the structural consequences of germline FOXA1 missense variants in a GI cancer cohort of Asian ancestry.

Availability of Data and Materials

The datasets generated or analyzed during the current study are available to bona fide researchers via the UK Biobank. Applications for access to UK Biobank data can be submitted through the UK Biobank access portal https://www.ukbiobank.ac.uk/use-our-data/apply-for-access/.

Author Contributions

SHT, WHL, EUHS and PJV designed the study. SHT, YTH, AJJ and WHL analysed and interpreted the data. SHT acquired the UK Biobank data and constructed the FOXA1 structural models. SHT drafted the manuscript and had primary responsibility for statistical analysis and final content. All authors have been involved in revising it critically for important intellectual content. All authors gave final approval of the version to be published. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

The UK Biobank received ethical approval from the National Information Governance Board for Health and Social Care and the National Health Service North West Centre for Research Ethics Committee. This study was also approved by the Medical Research and Ethics Committee (MREC), Ministry of Health Malaysia (NMRR ID-23-00931-SPO). The study was conducted in accordance with the principles of the Declaration of Helsinki and relevant ethical guidelines and regulations. Written informed consent was obtained from all UK Biobank participants at the time of their original recruitment. No additional participants were recruited for the present study.

Acknowledgment

This research was conducted using the UK Biobank Resource under application number 96759. The authors thank the Director General of Health Malaysia for granting permission to publish this article.

Funding

This research received no external funding.

Conflict of Interest

Ah-Jung Jeon and Yan Ting Hee are employed by Mirxes Lab Pte. Ltd. Other authors declare no conflict of interest.

Declaration of Generative AI and AI-Assisted Technologies in Manuscript Preparation

ChatGPT (GPT-5.3 Instant, OpenAI) was used to assist with language polishing and improving the clarity of specific sections of this manuscript. After its use, the authors thoroughly reviewed, verified, and revised all content to ensure accuracy and originality.

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a Cancer Journal for Clinicians. 2024; 74: 229–263. https://doi.org/10.3322/caac.21834

[2] MacDonald WJ, Purcell C, Pinho-Schwermann M, Stubbs NM, Srinivasan PR, El-Deiry WS. Heterogeneity in Cancer. Cancers. 2025; 17: 441. https://doi.org/10.3390/cancers17030441

[3] McInnes G, Sharo AG, Koleske ML, Brown JEH, Norstad M, Adhikari AN, et al. Opportunities and challenges for the computational interpretation of rare variation in clinically important genes. American Journal of Human Genetics. 2021; 108: 535–548. https://doi.org/10.1016/j.ajhg.2021.03.003

[4] Conroy MC, Lacey B, Bešević J, Omiyale W, Feng Q, Effingham M, et al. UK Biobank: a globally important resource for cancer research. British Journal of Cancer. 2023; 128: 519–527. https://doi.org/10.1038/s41416-022-02053-5

[5] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018; 562: 203–209. https://doi.org/10.1038/s41586-018-0579-z

[6] Backman JD, Li AH, Marcketta A, Sun D, Mbatchou J, Kessler MD, et al. Exome sequencing and analysis of 454,787 UK Biobank participants. Nature. 2021; 599: 628–634. https://doi.org/10.1038/s41586-021-04103-z

[7] Fatumo S, Chikowore T, Choudhury A, Ayub M, Martin AR, Kuchenbaecker K. A roadmap to increase diversity in genomic studies. Nature Medicine. 2022; 28: 243–250. https://doi.org/10.1038/s41591-021-01672-4

[8] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American Journal of Human Genetics. 2016; 99: 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

[9] Li M, Petukh M, Alexov E, Panchenko AR. Predicting the Impact of Missense Mutations on Protein-Protein Binding Affinity. Journal of Chemical Theory and Computation. 2014; 10: 1770–1780. https://doi.org/10.1021/ct401022c

[10] Chen E, Facio FM, Aradhya KW, Rojahn S, Hatchell KE, Aguilar S, et al. Rates and Classification of Variants of Uncertain Significance in Hereditary Disease Genetic Testing. JAMA Network Open. 2023; 6: e2339571. https://doi.org/10.1001/jamanetworkopen.2023.39571

[11] Zaret KS, Mango SE. Pioneer transcription factors, chromatin dynamics, and cell fate control. Current Opinion in Genetics & Development. 2016; 37: 76–81. https://doi.org/10.1016/j.gde.2015.12.003

[12] Zaret KS. Pioneer Transcription Factors Initiating Gene Network Changes. Annual Review of Genetics. 2020; 54: 367–385. https://doi.org/10.1146/annurev-genet-030220-015007

[13] Ma W, Jiang J, Li M, Wang H, Zhang H, He X, et al. The clinical significance of forkhead box protein A1 and its role in colorectal cancer. Molecular Medicine Reports. 2016; 14: 2625–2631. https://doi.org/10.3892/mmr.2016.5583

[14] Seachrist DD, Anstine LJ, Keri RA. FOXA1: A Pioneer of Nuclear Receptor Action in Breast Cancer. Cancers. 2021; 13: 5205. https://doi.org/10.3390/cancers13205205

[15] Liu N, Wang A, Xue M, Zhu X, Liu Y, Chen M. FOXA1 and FOXA2: the regulatory mechanisms and therapeutic implications in cancer. Cell Death Discovery. 2024; 10: 172. https://doi.org/10.1038/s41420-024-01936-1

[16] Lazar SB, Pongor L, Li XL, Grammatikakis I, Muys BR, Dangelmaier EA, et al. Genome-Wide Analysis of the FOXA1 Transcriptional Network Identifies Novel Protein-Coding and Long Noncoding RNA Targets in Colorectal Cancer Cells. Molecular and Cellular Biology. 2020; 40: e00224–20. https://doi.org/10.1128/MCB.00224-20

[17] Chen X, Li D, Wang N, Yang M, Liao A, Wang S, et al. Bioinformatic analysis suggests that UGT2B15 activates the Hippo-YAP signaling pathway leading to the pathogenesis of gastric cancer. Oncology Reports. 2018; 40: 1855–1862. https://doi.org/10.3892/or.2018.6604

[18] Fawzy M, Marsh JA. Assessing variant effect predictors and disease mechanisms in intrinsically disordered proteins. PLoS Computational Biology. 2025; 21: e1013400. https://doi.org/10.1371/journal.pcbi.1013400

[19] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596: 583–589. https://doi.org/10.1038/s41586-021-03819-2

[20] David A, Sternberg MJE. Protein structure-based evaluation of missense variants: Resources, challenges and future directions.

Current Opinion in Structural Biology. 2023; 80: 102600. https://doi.org/10.1016/j.sbi.2023.102600

[21] Tan SH, Guan CA, Bujang MA, Lai WH, Voon PJ, Sim EUH. Identification of phenomic data in the pathogenesis of cancers of the gastrointestinal (GI) tract in the UK biobank. Scientific Reports. 2024; 14: 1997. https://doi.org/10.1038/s41598-024-52421-9

[22] Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nature Genetics. 2021; 53: 942–948. https://doi.org/10.1038/s41588-021-00885-0

[23] Yun T, Li H, Chang PC, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus. Bioinformatics. 2021; 6: 5582–5589. https://doi.org/10.1093/bioinformatics/btaa1081

[24] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012; 6: 80–92. https://doi.org/10.4161/fly.19695

[25] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biology. 2016; 17: 122. https://doi.org/10.1186/s13059-016-0974-4

[26] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics. 2014; 46: 310–315. https://doi.org/10.1038/ng.2892

[27] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature. 2024; 625: 92–100. https://doi.org/10.1038/s41586-023-06045-0

[28] Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. American Journal of Human Genetics. 2017; 100: 267–280. https://doi.org/10.1016/j.ajhg.2017.01.004

[29] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine : Official Journal of the American College of Medical Genetics. 2015; 17: 405–424. https://doi.org/10.1038/gim.2015.30

[30] Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current Protocols in Bioinformatics. 2016; 54: 1.30.1–1.30.33. https://doi.org/10.1002/cpbi.5

[31] Ahmad S, Jose da Costa Gonzales L, Bowler-Barnett EH, Rice DL, Kim M, Wijerathne S, et al. The UniProt website API: facilitating programmatic access to protein knowledge. Nucleic Acids Research. 2025; 53: W547–W553. https://doi.org/10.1093/nar/gkaf394

[32] Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Research. 2018; 46: W350–W355. https://doi.org/10.1093/nar/gky300

[33] Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. Protein Science : a Publication of the Protein Society. 2021; 30: 60–69. https://doi.org/10.1002/pro.3942

[34] Modha N, Alexov E. Functional and Structural Characterization of Pathogenicity of Human Arginine-Histidine Variants. Journal of Computational Biophysics and Chemistry. 2026; 25: 3185–3209. https://doi.org/10.1142/S2737416526400041

[35] Zhou BR, Feng H, Huang F, Zhu I, Portillo-Ledesma S, Shi D, et al. Structural insights into the cooperative nucleosome recognition and chromatin opening by FOXA1 and GATA4. Molecular Cell. 2024; 84: 3061–3079.e10. https://doi.org/10.1016/j.molcel.2024.07.016

[36] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, et al. The Human Transcription Factors. Cell. 2018; 172: 650–665. https://doi.org/10.1016/j.cell.2018.01.029

[37] Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Research. 2015; 43: D512–20. https://doi.org/10.1093/nar/gku1267

[38] Bailey CG, Gupta S, Metierre C, Amarasekera PMS, O'Young P, Kyaw W, et al. Structure-function relationships explain CTCF zinc finger mutation phenotypes in cancer. Cellular and Molecular Life Sciences : CMLS. 2021; 78: 7519–7536. https://doi.org/10.1007/s00018-021-03946-z

[39] Adams EJ, Karthaus WR, Hoover E, Liu D, Gruet A, Zhang Z, et al. FOXA1 mutations alter pioneering activity, differentiation and prostate cancer phenotypes. Nature. 2019; 571: 408–412. https://doi.org/10.1038/s41586-019-1318-9

[40] Iwafuchi-Doi M, Donahue G, Kakumanu A, Watts JA, Mahony S, Pugh BF, et al. The Pioneer Transcription Factor FoxA Maintains an Accessible Nucleosome Configuration at Enhancers for Tissue-Specific Gene Activation. Molecular Cell. 2016; 62: 79–91. https://doi.org/10.1016/j.molcel.2016.03.001

[41] Hwang J, Likasitwatanakul P, Deshmukh SK, Wu S, Kwon JJ, Toye E, et al. Structurally Oriented Classification of FOXA1 Alterations Identifies Prostate Cancers with Opposing Clinical Outcomes and Distinct Molecular and Immunologic Subtypes. Clinical Cancer Research : an Official Journal of the American Association for Cancer Research. 2025; 31: 936–948. https://doi.org/10.1158/1078-0432.CCR-24-3471

[42] Sharifi Tabar M, Francis H, Yeo D, Bailey CG, Rasko JEJ. Mapping oncogenic protein interactions for precision medicine. International Journal of Cancer. 2022; 151: 7–19. https://doi.org/10.1002/ijc.33954

[43] Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (New York, N.Y.). 2023; 381: eadg7492. https://doi.org/10.1126/science.adg7492

[44] Caswell RC, Gunning AC, Owens MM, Ellard S, Wright CF. Assessing the clinical utility of protein structural analysis in genomic variant classification: experiences from a diagnostic laboratory. Genome Medicine. 2022; 14: 77. https://doi.org/10.1186/s13073-022-01082-2

[45] Keskin Karakoyun H, Yüksel ŞK, Amanoglu I, Naserikhojasteh L, Yeşilyurt A, Yakıcıer C, et al. Evaluation of AlphaFold structure-based protein stability prediction on missense variations in cancer. Frontiers in Genetics. 2023; 14: 1052383. https://doi.org/10.3389/fgene.2023.1052383

[46] Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nature Protocols. 2007; 2: 1849–1861. https://doi.org/10.1038/nprot.2007.249

[47] Vallejo DD, Rojas Ramírez C, Parson KF, Han Y, Gadkari VV, Ruotolo BT. Mass Spectrometry Methods for Measuring Protein Stability. Chemical Reviews. 2022; 122: 7690–7719. https://doi.org/10.1021/acs.chemrev.1c00857