Pathogenic nsSNPs that increase the risks of cancers among the Orang Asli and Malays

Nurul Ain Khoruddin1,2, Mohd Nur Fakhruzzaman Noorizhab1,3, Lay Kek Teh1,3, Farida Zuraina Mohd Yusof1,2 & Mohd Zaki Salleh1,3

Single-nucleotide polymorphisms (SNPs) are the most common genetic variations for various complex human diseases, including cancers. Genome-wide association studies (GWAS) have identified numerous SNPs that increase cancer risks, such as breast cancer, colorectal cancer, and leukemia. These SNPs were cataloged for scientific use. However, GWAS are often conducted on certain populations in which the Orang Asli and Malays were not included. Therefore, we have developed a bioinformatic pipeline to mine the whole-genome sequence databases of the Orang Asli and Malays to determine the presence of pathogenic SNPs that might increase the risks of cancers among them. Five different in silico tools, SIFT, PROVEAN, Poly-Phen-2, Condel, and PANTHER, were used to predict and assess the functional impacts of the SNPs. Out of the 80 cancer-related nsSNPs from the GWAS dataset, 52 nsSNPs were found among the Orang Asli and Malays. They were further analyzed using the bioinformatic pipeline to identify the pathogenic variants. Three nsSNPs; rs1126809 (TYR), rs10936600 (LRRC34), and rs757978 (FARP2), were found as the most damaging cancer pathogenic variants. These mutations alter the protein interface and change the allosteric sites of the respective proteins. As TYR, LRRC34, and FARP2 genes play important roles in numerous cellular processes such as cell proliferation, differentiation, growth, and cell survival; therefore, any impairment on the protein function could be involved in the development of cancer. rs1126809, rs10936600, and rs757978 are the important pathogenic variants that increase the risks of cancers among the Orang Asli and Malays. The roles and impacts of these variants in cancers will require further investigations using in vitro cancer models.

1Integrative Pharmacogenomics Institute (iPROMISE), Universiti Teknologi MARA (UiTM), Selangor Branch, Puncak Alam Campus, 42300 Puncak Alam, Selangor, Malaysia. 2Faculty of Applied Sciences, Universiti Teknologi MARA (UiTM), Selangor, Malaysia. 3Faculty of Pharmacy, Universiti Teknologi MARA (UiTM), Selangor Branch, Puncak Alam Campus, 42300 Puncak Alam, Selangor, Malaysia. *email: zakisalleh@uitm.edu.my
protein. The structure and function of the native proteins were found to be altered due to the nsSNPs using a pipeline comprised of several bioinformatics tools\(^3\). A recent study had identified the deleterious nsSNPs on the hOGG1 gene that altered the secondary structure of the expressed protein and destabilized its local conformation, which increases the risks for lung cancer\(^4\). Furthermore, \textit{in-silico} modeling has been widely used to assess the functional impacts of nsSNPs and their possible roles in cancers\(^5,6\). \textit{in-silico} modeling has the advantage of being able to make rapid predictions for the mechanisms of actions of a wide range of compounds in a high-throughput mode. Another advantage is that prediction can be made based on the structure of a compound before it is synthesized\(^7\).

Databases of human variants have been developed with different scopes and contents used to predict diseases\(^8\) in achieving personalized medicine\(^9\). The genome-wide association study (GWAS) database (https://www.ebi.ac.uk/gwas/) is widely used to associate SNPs with diseases. Although there are other existing human variant databases such as ClinVar, COSMIC, SwissVar, and Humbavari, GWAS is the only database that gives a world of information or catalogs on disease mutations in different populations. This database also provides information on the statistically significant variants and the increase/decrease associated risks for each phenotype\(^10\).

The application of genomics, bioinformatics, and the availability of data generated from high-throughput technologies are the fundamental tools for implementing precision medicine not only for cancer diseases but also for other common and rare diseases\(^11\). Various tools have been used to predict the functional effects of nonsynonymous coding variants using basic sequence homology\(^12,13\); empirically derived rules\(^14\); structural and functional features\(^15\); a weighted average of the normalized scores\(^16\); decision trees\(^17,18\); support vector machines\(^19\); and Bayesian classifiers\(^20\). A comprehensive systematic evaluation study on the performances of these widely used prediction methods to identify the pathogenicity of the SNPs is required\(^21\). While new and more algorithms are being developed, the accuracy of prediction using a combination of the different algorithms should be validated. It is recommended that different computational methods are used to determine the impact of different SNPs during the screening step, and further validation should be incorporated in studying the impacts of nsSNPs on specific proteins\(^22\). In addition, complementary methods could be combined in a meta-server to yield more reliable predictions\(^23\). Several recent studies had reported on the use of a combination of various methods to uncover the potential impact of the nsSNPs in understanding the molecular mechanisms of various diseases, which includes cancers\(^24-28\). The combination of these tools allows more accurate prediction using the multiple conservation, structural, or combined methods (conservation and structural). Therefore, combined methods and meta-prediction methods (predictors that integrate multi-predictor results) are important for biomedical applications. This is because they can be applied to a much greater number of single nucleotide variants, considering that many human proteins do not currently have an experimentally defined structure or a close homolog to construct a model. Thus, combined and meta-prediction methods have a wide range of potential applications using the combinations of features yet to be explored\(^29\). As GWAS is usually conducted on a large population size using a high throughput detection method and is costly, some world populations were not studied. Therefore, their disease risks are not available. The Orang Asli are still practicing traditional healing methods, therefore the record on the incidence of cancers among the Orang Asli is lacking. This has posed challenges to the authorities to strategize health programs to ensure the sustainability of the Orang Asli. Due to the lack of phenotypic data on cancers, mining the genomes of the Orang Asli to predict their susceptibility for the different types of cancers would provide important data that allows the scientists to strategize research focus areas and for the authorities to provide relevant funding. In this study, we aimed to develop and validate a bioinformatics pipeline to detect and annotate the cancer-associated nsSNPs, and predict the structural and functional impacts\(^30\) of these nsSNPs that might increase the risks of cancers among the Orang Asli. Using the same pipeline, we also investigate the cancer risks of the Malays, which constitute the biggest population in Malaysia. The database of the Malay genomes was provided by Wong et al.\(^46\) and lacks information on the phenotypic traits, therefore it is interesting to predict the cancer susceptibility risks for this cohort using the established pipeline. The pipeline contains several stages: (i) preprocessing and cleaning of the data; (ii) genotype calling; (iii) filtering and quality control; (iv) annotation of variants; (v) functional prediction of nsSNPs; (vi) statistical analysis of association studies; and (vii) validation of findings. This study is a part of our initiatives to enhance precision health in our country. The bioinformatics pipeline developed in this study will be used in the future to predict genomic variations associated with different diseases.

**Methods**

**Whole genome sequences.** The whole-genome sequences of ninety-eight (98) healthy and unrelated Orang Asli from six different sub-tribes were retrieved from the Whole-Genome-Sequence Database at Integrative Pharmacogenomics Institute (iPROMISE) in the form of a bam file. The Orang Asli were recruited from sub-tribes that are located in the (i) northern region of the Peninsular Malaysia (| Bateq, n = 22; Gua Musang, Kelantan); (2) Lenggong, Perak) and (Kensiu, n = 19; Baling, Kedah); (ii) in the central region (Che Wong, n = 18; Kuala Gandah, Pahang) and (Semai, n = 16; Kuala Lipis, Pahang); in the southern region (Kanaq, n = 7; Kota Tinggi, Johor)]. The mean coverage of whole-genome sequences of Orang Asli across all the 98 samples was 37.39 x (minimum of 18.44 x to a maximum of 46.02 x) and was checked using Qualimap version 2.2.1.

The genomic DNA (gDNA) of each of the 98 Orang Asli individuals was isolated from 300 ml of whole blood using the Wizard Genomic DNA Purification Kit (Promega, Wisconsin, USA). A microvolume
spectrophotometer (NanoDrop 2000, Thermo Scientific) was used to evaluate DNA quantity. Whole-genome sequencing of the 98 Orang Asli were then performed using the Genome Analyzer System (GA IIX) with a target of >30x coverage. The whole-genome sequences of Orang Asli were then assembled by the in-house bioinformatics workflow. Quality on the raw sequence data was checked with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed with Trimmomatic version 2.5 (https://software.broadinstitute.org/gatk/best-practices/) recommended by the Genome Analysis Toolkit (GATK) Best Practices. Briefly, the reads were aligned using BWA version 0.6.1-r104 to the reference human genome GRCh37/hg19 and duplicates were labeled and extracted using Picard version 1.119 (http://broadinstitute.github.io/picard/).

Whole-genome sequences of ninety-six (96) healthy Singaporean Malays were obtained in the form of bam files from Singapore Sequencing Malay Project (http://www.stategen.ns.edu.sg/~SSMP). Malays are Austroasiatic-speaking ethnic group who mainly live in Malaysia, Indonesia, and Singapore in the Southeast Asian region. The mean coverage of the whole-genome sequences of Singapore Malays across all the 96 samples was 47.6x. The depth of coverage for each sample ranged from 35.5x to 81.9x. All the genomic DNA of 96 Malays individuals was collected from the Singapore BioBank. Picogreen was used to measure fluorescence intensity, and the SpectraMax Gemini EM microwell plate reader was used to confirm that the DNA content was greater than 50 ng/l using spectrophotometric settings at 480/520 nm (Ex/Em). Subsequently, DNA samples were sent to the Defense Medical and Environmental Research Institute for preparation. Whole-genome sequencing of 96 Malays were then performed using the Illumina HiSeq 2000 with a target of >30x coverage.

Variant calling pipeline was performed using HaplotypeCaller and BaseRecalibrator (GATK v2.5) for each sequence data (bam file format) of the Orang Asli and Malays. The HaplotypeCaller was used to detect variants and BaseRecalibrator was used for base quality score recalibration (BQSR). Vcf files for each sample were generated for quality-filtering. Variant filtering was performed using SelectVariants (GATK v2.5), to extract SNPs and exclude variants with a read depth of less than 5 or a quality Phred score of less than 30.

The study protocol was approved by Universiti Teknologi MARA Research Ethics Committee [600-RMI (51/6/01) & 600-RMI (5/1/6)] and the Department of Orang Asli Development (Jabatan Kemajuan Orang Asli Malaysia, JAKOA) Research Ethics Committee [JAKOA.PP.30.052 Jld 5(62)].

**Bioinformatics workflow.** High-risk nsSNPs associated with cancer were classified using the GWAS-Catalog as the source of the dataset, and various bioinformatics tools were employed in the workflow (Fig. 1).

**Nonsynonymous SNPs datasets for validation.** The sensitivity, specificity, and accuracy of the functional effect prediction were determined using a combination of five different algorithms (SIFT, PolyPhen-2, Condel, PROVEAN, and PANTHER), with and without conservation (Consurf) and protein stability (I-Mutant). The standard dataset used comprised of nsSNPs associated with breast cancer from ClinVar. The ClinVar dataset includes a total of 100 clinically tested nsSNPs in which 50 nsSNPs were reported as pathogenic while the other 50 nsSNPs were reported as benign (Table S1). The 100 nsSNPs training dataset were randomly chosen out of 1020 clinically tested nsSNPs associated with breast cancer reported in the ClinVar as it is one of the most commonly studied cancer dataset. Although the dataset is primarily associated with breast cancer, the main purpose of using the training dataset is to test the ability of the pipeline to detect all the deleterious nsSNPs. Additionally, the sample size chosen also is sufficient as concluded by Thusberg et al., that the analysis result of using a small dataset (100SNPs) is comparable to a larger size (1000 SNPs) for a training dataset. Datasets of different types of cancer and a larger sample size may also be used to achieve the same objective.

Analytical parameters of studied tools were calculated using Eqs. (1), (2), and (3) according to Fletcher and Glantz.

Sensitivity (Se) is a proportion of the true-positive results (correct identification of pathogenic variants), according to Eq. (1).

\[
Se = \frac{TP}{TP + FN} \times 100\%
\]  

where TP denotes true-positive cases, and FN denotes false-negative cases.

Specificity (Sp) is a proportion of the true negative results (correct identification of benign variants), according to Eq. (2).

\[
Sp = \frac{TN}{TN + FP} \times 100\%
\]  

where TN denotes true negative cases, and FP denotes false-positive cases.

Accuracy (Ac) is the ratio of complete, correct predictions to the total number of predictions, according to the following Eq. (3).

\[
Ac = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%
\]

Datasets. Information on the genetic variants associated with cancers (SNP ID) was retrieved from the GWAS-Catalog database (https://www.ebi.ac.uk/gwas). Residue change, risk allele frequency, phenotype, and protein accession number were retrieved from The NHGRI GWAS Catalog. The dataset was built after 179,365 genetic variants were filtered based on the keywords ‘cancer’, ‘carcinoma’, ‘glioma’, ‘leukemia’, ‘lymphoma’, ‘melanoma’, and ‘sarcoma’ (Table S2).
Figure 1. A workflow diagram for predicting high-risk cancer-related nsSNPs. The training dataset used was ClinVar to validate the capability of the pipeline to identify pathogenic variants based on the prediction of functional effect, conservation, and stability of cancer-related variants reported in Clinvar. The red dotted line represented the results for the training dataset.
Retrieval of SNPs from the whole-genome sequences. The SNPs that are associated with cancer risks were identified using VCFtools102 based on the dataset (Table S2). The variants were then annotated to identify the associated genes, allele frequency (AF), location of the SNPs in the genome sequences, the position of amino acid change in protein sequences, and codon changes using Variant Effect Predictor103. hg19/GRCh37 was used as the reference genome for the analyses.

Identification of the damaging nsSNPs. The functional effects of identified nsSNPs were predicted by using five different bioinformatics tools. These algorithmic programs included Sorting Intolerant From Tolerant (SIFT) [http://sift.jcvi.org/www/SIFT_seq_submit2.html]105, Polymorphism Phenotyping v2 (PolyPhen-2) [http://genetics.bwh.harvard.edu/pph2/]127, Consensus Deleterious (Condol) [http://bbglab.irbbarcelona.org/fannsdb/query/condel]50, Protein Variation Effect Analyzer (PROVEAN) [http://provean.jcvi.org/index.php]105 and Protein Analysis Through Evolutionary Relationships (Panther v14.1) [http://www.pantherdb.org/tools/csnpscore.do]132. SIFT predicts the effects of an amino acid substitution on protein functions. The sequence homology and the physiochemical characteristics were computed using a normalized probability score (SIFT score) for each substitution25. PolyPhen-2 predicts the potential effect of an amino acid substitution on both protein structure and function using a combination of homolog sequence alignment-based methods and protein 3D structure. The prediction is provided as benign, possibly damaging, and probably damaging according to the scores differences of the position-specific independent count (PSIC) between 2 variants (native amino acid and mutant amino acid)27. Condol predicts the effect of coding variants on protein function based on the ensemble score of multiple prediction tools (SIFT, PolyPhen-2, FATHMM, and Mutation Assessor)108. PROVEAN predicts the functional effects of protein sequence variations, including single or multiple amino acid substitutions and in-frame insertions and deletions130. PANTHER estimates the likelihood of a particular nsSNP to cause a functional effect on the protein using position-specific evolutionary preservation27. The description of the tools used is presented in Table 1.

The nsSNPs were considered high-risk if they were predicted to be damaging or deleterious by at least four bioinformatics tools. They were then subjected to further analysis.

Analysis on conservation of protein evolutionary. ConSurf (consurf.tau.ac.il/) is a bioinformatics tool that was utilized to predict the evolutionary conservation of amino acid in CACFD1, RREB1, LRRC34, ETFA, CPVL, INCENP, FARP2, and TYR protein. It is a web server that builds phylogenetic relationships between homologous sequences to estimate the evolutionary conservation of amino acid positions in a protein or DNA molecule. The conservation analysis on the target proteins was performed to show the significance of each residue position for the protein structure or function. The rate of evolution was determined based on the evolutionary relationship between the protein or DNA, its homologs, and the similarity between amino (nucleic) acids as expressed in the substitutions matrix. Furthermore, Consurf offers an accurate estimation of the evolutionary rate using either an empirical Bayesian approach or a maximum probability (ML) method47. Protein sequence in FASTA format was used as the input. UniProtKB accession numbers for the sequences are: CACFD1, Q9UGQ2; RREB1, Q92766; LRRC34, Q8IZ02; ETFA, P13804; CPVL, Q9H3G5; INCENP, Q9NQS7; FARP2, Q94887; and TYR, P14679. Consurf created an output consists of the protein sequence and multiple sequence alignment colored by conservation scores. The conservation score ranged from 1 to 9, where 1 to 4 is considered as variable, 5 to 6 as intermediate, and 7 to 9 as conserved amino acid position. We selected those residues with a high score for the high-risk nsSNP for further analysis.

Analysis of protein stability. I-Mutant Suite is a web server (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi)54 that was used to predict the stability of protein changes caused by a single point mutation. This tool is trained on a ProTherm-derived data set which is the most extensive database on experimental thermodynamic data on free energy changes, which measures protein stability due to mutations107. The RI ranges from 0 to 10, where 10 is the highest reliability107. The free energy change values from the unfolding Gibbs free energy value of the mutant protein minus the unfolding Gibbs free energy value of the native protein. The RI ranges from 0 to 10, where 10 is the highest reliability107. The free energy change values were categorized into three classes: (i) DDG < -0.5 kcal/mol as destabilizing mutations; (ii) DDG > 0.5 kcal/mol as stabilizing mutations; (iii) -0.5 < DDG < 0.5 kcal/mol as neutral mutations108.

Three-dimensional (3D) protein modeling. The 3D structures of native and mutant (due to nsSNPs) proteins were constructed to explore the differences in the structural stability between the native and mutant proteins. The iterative threading assembly refinement (I-TASSER) server is an integrated platform that provides automated protein structure and function prediction based on the sequence-to-structure-to-function framework109. It was employed for the prediction of 3D protein models of native and mutant protein structures with high-risk nsSNPs. It has the most advanced algorithm to build high-quality 3D protein model from amino acid sequences. I-TASSER generates a full-length model of proteins by excising continuous fragments from threading alignments and then reassembling them using replica-exchanged Monte Carlo simulations. SPICKER clusters low-temperature replicas (decoys) generated during the simulation, and the top five cluster centroids are selected for generating full atomic models. The accuracy of the predicted model is reflected in the form of the confidence score (C-score). The C-scores range is between 5 and 2. The greater values of the C-score display...
| Program (website) | Algorithm | Input parameters | Evolutionary analysis | Structural attributes | Computing tools | Effect | Score | Prediction | References |
|-------------------|-----------|-----------------|-----------------------|----------------------|-----------------|-------|------|------------|------------|
| SIFT (http://sift.jcvi.org) | Evolutionary conservation | dbSNP rs ID | Multiple Sequences Alignment | / | Matrix Dir-ichlet | Effect of amino acid substitution on structure/function of protein | 0.00—1 | <0.05 = "Damaging"; 0.05 = "Tolerated" | 26 |
| Polyphen-2 (http://genetics.bwh.harvard.edu/pph2/) | Protein structure/function and evolutionary conservation | dbSNP rs ID | PSIC profiles | Homolog mapping/predictions | Naive Bayesian classifier | Effect of amino acid substitution on structure/function of protein | 0.00—1 | 0.0—0.15 = "Benign"; 0.15—1.0 = "Possibly damaging"; 0.85—1.0 = "Probably damaging" | 2 |
| Condel (http://bg.upf.edu/lanns db/) | Protein structure/function and evolutionary conservation | Genomic coordinate (s), variant(s) | SIFT, PolyPhen-2, MutationAssessor,FATHMM | Homolog mapping/predictions (PolyPhen-2) | Weighted average of the normalized scores from multiple methods | Effect of amino acid substitution on structure/function of protein | 0.00—1 | 0.0 = "Neutral"; 1.0 = "Deleterious" | 36 |
| PROVEAN (http://provean.jcvi.org/index.php) | Evolutionary conservation/alignment and measurement of similarity between variant sequence and protein sequence homolog | Genomic coordinate (s), variant(s) | BLASTP | / | Blocks Substitution Matrix (BLOSUM62) | Functional effect on protein | (−40—12.5) | ≥—2.5 = "Deleterious"; ≤—2.5 = "Neutral" | 104 |
| PANTHER (http://www.pantherdb.org/tools/csnpscoreform.jsp) | Evolutionary conservation/alignment and measurement of similarity between variant sequence and protein sequence homolog | Protein sequences, substitution(s) | Multiple Sequence alignment (PANTHER library) | / | Alignment scores Hidden Markov Models (HMM) | Functional effect on protein | 0.00—4200 | >450 = "Probably damaging"; 450—200 = "Possibly damaging"; ≤200 = "Probably benign" | 52 |
| Consurf (https://consurf.tau.ac.il/) | Evolutionary conservation/alignment and measurement of similarity between variant sequence and protein sequence homolog | Protein sequences (FASTA format), substitution(s) | PSI-BLAST, Multiple sequence alignment (MAFFT (default), PRANK, T-COFFEE, MUSCLE or CLUSTALW) | / | Neighbor-joining Empirical Bayesian or Machine learning, Heuristic algorithm | Evolutionary conservation | 1—9 | 1 = "Most variable positions" (turquoise); 5 = "Intermediate conserved positions" (white); 9 = "Most conserved positions" (maroon) | 47 |
| 1-Mutant (http://gpcr2.biocomp.unibo.it/cgi/predictors/1-<br>Mutant3.0.cgi) | Protein stability changes upon single-site mutations from the protein sequence or protein structure | Protein sequences (FASTA format), substitution(s) | Multiple Sequence alignment Relative Solvent Accessible Area (DSSP Program, DDGMut dataset) | Support Vector Machine (SVM) | Protein Stability changes | DDG > 0, DDG < 0 | ΔΔG ≤ 0.5 kcal/mol = "Destabilizing mutations"; ΔΔG > 0.5 kcal/mol ≤ ΔΔG ≤ 0.5 kcal/mol = "Neutral mutations" | 54 |
| TM-align (https://zhanglab.ccmb.med.umich.edu/TM-align/) | Alignment and measurement of similarity between two protein structures of known/unknown equivalence | Protein Structure (PDB format) | Superposition of two structures (TM-Score) | Heuristic dynamic programming iterations | Protein structure changes | 0—1 | ≤−0.5 = "Randomly chosen unrelated proteins"; >0.5 = "Same fold in SCOP/CATH" | 60 |
| ModPred (http://www.modpred.org/) | Post-translational Modification | Protein sequences (FASTA format) | PSI-BLAST | Homolog mapping/predictions Position-specific scoring matrices (PSSM) | Identification of Posttranslational Modification Sites | 0—1 | >0.5 = "High confidence"; 0.5—0.3 = "Medium"; <0.5 = "Low confidence" | 115 |

Continued
The protein sequences of the selected nsSNPs were submitted as input for each of the modifications models as given by ModPred. The protein sequences of the identified genetic variants and their amino acid substitutions were submitted. MutPred v1.2 generated output scores indicating the probability of deleterious or disease-associated amino acid substitution. The top five features with P value impact on the functional and structural properties would be recorded. The predicted scores were classified based on three hypotheses; (i) g > 0.5 and p < 0.05 as actionable hypotheses; (ii) g > 0.75 and p < 0.05 as confident hypotheses; (iii) g > 0.75 and p < 0.01 as very confident hypotheses.

Identification of functional and structural properties. MutPred v1.2 and HOPE were used to identify the functional and structural properties of the selected nsSNPs. MutPred is a web application tool that effectively classifies amino acid substitution as being associated with a disease or neutral in human (http://mutpr ed.mutdb.org/). This tool also helps in predicting the deleterious amino acid substitution or molecular cause of disease112. It focuses on a wide range of structural and functional properties, including secondary structure, signal peptide and transmembrane topology, catalytic activity, macromolecular binding, PTMs, metal-binding, and allostery106. Protein sequences (FASTA format) of the identified genetic variants and their amino acid substitutions were submitted. MutPred v1.2 generated output scores indicating the probability of deleterious or disease-associated amino acid substitution. The top five features with P value impact on the functional and structural properties would be recorded. The predicted scores were classified based on three hypotheses; (i) g > 0.5 and p < 0.05 as actionable hypotheses; (ii) g > 0.75 and p < 0.05 as confident hypotheses; (iii) g > 0.75 and p < 0.01 as very confident hypotheses.

HOPE is a web service tool that was used to identify the structural effects of a point mutation on human protein sequence (www.cmbi.ru.nl/hope)113. The protein sequences of the selected nsSNPs were submitted as input. HOPE generated results based on the collected and combined information from several web services and databases. Initially, the algorithm included BLAST against PDB and UniProt to obtain details on the tertiary structure to build a homology model. It was followed by the prediction of the protein features using the Distributed Annotation System114.

ModPred (http://www.modpred.org/)106 is a web server tool that was used for the prediction of post-translational modification (PTM) sites in proteins based on sequence-based features, physicochemical properties, and evolutionary features. A total of 34 logistic regression models were used in ModPred for 23 different PTM sites to simultaneously predict and analyze multiple types of PTM sites to obtain information on the functional and structural impacts of multiple PTM protein regulatory mechanisms. The 34 ensembles of logistic regression models were trained independently for 23 PTMs on a total collection of 126,036 experimentally tested non-redundant protein sites extracted from various public databases such as SwissProt, HPDR, PDB, Phospho.ELM, PhosphoSitePlus & PHOSIDA and literatures105. The PTM sites were predicted to have either low, medium, or high confidence scores. Sites with low confidence have scores of at least 0.5. In contrast, PTM sites with medium and high confidence have different predictor scores that were based on sensitivity and specificity estimates for each of the modifications models as given by ModPred.

Table 1. Description of the functional prediction tools.

| Program (website) | Algorithm | Input parameters | Evolutionary analysis | Structural attributes | Computing tools | Effect | Score | Prediction | References |
|-------------------|-----------|------------------|-----------------------|----------------------|-----------------|--------|-------|------------|-----------|
| MutPred2 (http://mutpr ed.mutdb.org/) | Evolutionary conservation/ alignment and measurement between variant sequence and protein sequence homolog, molecular alterations | Protein sequences (FASTA format), substitution(s) | PSI-BLAST | Homolog mapping/predictions | Neural network ensemble, Machine learning (ML) | Effect of amino acid substitution and their molecular mechanisms | 0–1 | ≥ 0.5 = "Pathogenic" ≤ 0.5 = "Benign" | 106 |

higher confidence for the predicted model109. The best model for each query protein was selected according to C-score values. Default parameters were used for each of the protein structures. The amino acid sequences of the proteins to be modeled were prepared in the FASTA format as input for the server to predict the native and mutant models. The predicted structures were loaded into PyMOL to visualize their molecular structures. PyMol was used to visualize the molecular structures in high-quality 3D images.

The qualities of all predicted protein structures were then validated by ERRAT tools (https://servicesn.mbi.ucla.edu/ERRAT/)110, and Ramachandran Plot. (https://zlab.unmassmed.edu/bu/rama/)111. ERRAT program analyzed the statistics of noncovalent interactions between three types of atoms, which are carbon (C), nitrogen (N), and oxygen (O). Consequently, six types of interactions are possible (CC, CN, CO, NN, NO, and OO). Ramachandran Plot illustrates the statistical distribution of the combinations of the backbone dihedral angles ϕ and ψ, in protein structures. The number of residues in the allowed or disallowed regions of the Ramachandran plot determines the quality of the model. Template modeling aligns (TM-align) was used for comparison between the predicted native and mutant protein models. Its algorithm identifies the best structural alignment between the protein pairs based on the combination of template modeling-score (TM-score), root mean square deviation (RMSD), and the superposition of the structures69. TM-score scores range from 0 to 1, where 1 represents the ideal match between two protein structures. In contrast, the higher value of RMSD represents a more significant difference between native and mutant structures.

Prediction of protein–protein interactions. STRING is a database and web resource dedicated to protein–protein interactions network, including direct (physical) and indirect (functional) interactions115. The database contains data from genomic context, experimental repositories, co-expression, and collections of public text116. The available information in the database will allow us to identify and further understand the experimental and/or theoretical interaction for TYR, FARP2, and LRRC34 for this study.
| Statistical parameters | Model A | Model B | Model C | Model D | Model A3 | Model B3 | Model C3 | Model D3 |
|------------------------|---------|---------|---------|---------|----------|---------|---------|---------|
| TP (N)                 | 50      | 46      | 45      | 42      | 48       | 44      | 39      | 38      |
| FN (N)                 | 0       | 4       | 5       | 8       | 2        | 6       | 11      | 12      |
| TN(N)                  | 25      | 32      | 40      | 47      | 40       | 44      | 46      | 48      |
| FP(N)                  | 25      | 18      | 10      | 3       | 10       | 6       | 4       | 2       |
| Sensitivity (%)        | 100     | 92      | 90      | 84      | 96       | 88      | 78      | 76      |
| Specificity (%)        | 50      | 64      | 80      | 94      | 80       | 88      | 92      | 96      |
| Accuracy (%)           | 75      | 78      | 85      | 89      | 88       | 88      | 85      | 86      |

**Annotation**

- Model A is the combination of five different tools which at least one tool predicted nsSNPs as deleterious/neutral.
- Model B is the combination of five different tools which at least two tools predicted nsSNPs as deleterious/neutral.
- Model C is the combination of five different tools which at least three tools predicted nsSNPs as deleterious/neutral.
- Model D is the combination of five different tools which at least four tools predicted nsSNPs as deleterious/neutral.
- Model A3 is the combination of Model A with the prediction of conservation and protein stability.
- Model B3 is the combination of Model B with the prediction of conservation and protein stability.
- Model C3 is the combination of Model C with the prediction of conservation and protein stability.
- Model D3 is the combination of Model D with the prediction of conservation and protein stability.

Table 2. Performance of 8 different prediction models (Model A, B, C, D, A3, B3, C3 and D3) using functional effect prediction tools (SIFT, PolyPhen-2, Condel, PROVEAN, and PANTHER) and conservation (Consurf) and protein stability (I-Mutant). The prediction tools’ performance was assessed on a standard dataset with all statistical parameters; TP, FN, TN, FP, sensitivity, specificity, and accuracy. TP = True Positive; FN = False Negative, TN = True Negative and, FP = False Positive.

---

**Molecular docking.** The effect of the deleterious point mutations over the binding affinity of FARP2, LRRC34, and TYR, were determined by molecular docking using UCSF Chimera 1.15 tools with Autodock Vina instruments. Protein and the peptide molecule were given as input for the docking experiments. The protein three-dimensional (3D) crystal structure, MYNN (PDB ID:2vpx), SRC (PDB ID:2h8h), and DCT (PDB ID: 4hx1) from RCSB Protein Data Bank (PDB) were used as receptors for LRRC34, FARP2, and TYR respectively. The peptide sequences from native and mutant FARP2, LRRC34, and TYR protein structures were used as the ligands for the docking procedure. The peptide sequences of at least nine amino acid residues of each of the native and mutant FARP2, LRRC34, and TYR proteins were converted into SMILES for the docking analysis. The three-dimensional structure for each ligand was then generated by the Build Structure tool within UCSF Chimera 1.15 software using SMILES as an input. Target proteins and ligands were optimized using the Dock Prep tool from UCSF Chimera 1.15 software with default parameters before docking analysis. These steps include removing solvents, adding hydrogens, and determining the charge. We maximized the grid box size along with the axes X, Y and, Z accordingly to define the binding sites for conducting the docking. The grid box size was set at 40.4399, 37.7452, 39.3645 along the x, y, and z points, respectively for MYNN (PDB ID:2vpx), 69,2063, 68.8481, 75.7427 SRC (PDB ID:2h8h) and 73.99757, 63.0875, 65.1247 DCT (PDB ID: 4hx1). The Autodock from UCSF Chimera 1.15 tools predicted and evaluated ten (10) protein binding sites for each interaction of receptors and ligands. The same binding sites of native and mutant proteins were compared. The PDB format of these input receptors and ligands were converted into a pdbqt format. The docking result and the binding interaction between ligand and receptor proteins were visualized by UCSF Chimera 1.15 tool.

**Results**

**Standard dataset.** The dataset contains a total of 100 nsSNPs in which 50 nsSNPs were reported as pathogenic, and 50 nsSNPs were reported as benign (Table S1). The parameters investigated were compared and are presented in Table 2. The sensitivity, specificity, and accuracy of the prediction for the clinical significance of the nsSNPs were calculated for four (4) models (Model A, B, C, and D). Model A represents at least one tool that predicted nsSNPs as deleterious or benign, and it showed the highest sensitivity (100%), followed by Model B (92%), Model C (90%), and Model D (84%). For specificity and accuracy, Model D showed the highest percentages (specificity 94%, and accuracy 89%) followed by Model C (specificity 80%, and accuracy 85%), Model B (specificity 64%, and accuracy 78%), and Model A (specificity 50%, and accuracy 75%). Further analyses were conducted using the combination of five functional effect tools which investigate the conservation and stability (Model A3, B3, C3, and D3). These models resulted in lower sensitivity of deleterious and benign nsSNPs compared to other models (Model A, B, C, D, A3, B3, and C3). However, Model A3, and B3 showed higher accuracy (88%) compared to Model D (89%) and Model C (85%).
SNPs dataset. The database included a total of 3,608 SNPs (excluding redundant nsSNPs entries), 80 are nsSNPs, 21 are sSNPs, 73 in the 3'UTR, 23 in the 5'UTR, 1,922 in the intronic region, 1,078 in the intergenic region, and the remaining are variants in the coding sequence regions, transcription factor binding sites, stop-gained region, splice region, regulatory region, splice-acceptor, noncoding transcript and in-frame insertion. The details are provided in the Table S2. For further investigation, only nsSNPs were selected.

Cancer-related nsSNPs for whole-genome sequences of Orang Asli and Malays. All of the identified SNPs were searched against the SNPs dataset retrieved from GWAS. Out of 80 nsSNPs associated with cancers from the dataset, a total of 52 nsSNPs were found among the Orang Asli and Malays (43 in Orang Asli and 43 in Malays), as presented in Table 3. Thus, we selected all the 52 identified nsSNPs associated with cancer risks among the Orang Asli and Malays for further investigation.

Predicted deleterious nsSNPs among the Orang Asli and Malays. The SNP effect on protein function remains unexplained for a large number of nsSNPs in humans. Five different in-silico nsSNPs prediction algorithms were successfully used to predict the impact of all the nsSNPs on the function, structure, and sequence conservation of the proteins in the Orang Asli and Malays studied in this study. The five tools used were SIFT, PolyPhen-2, CONDEL, PROVEAN, and PANTHER. Different algorithms are used by these in silico methods, which often resulted in outputs with different significant values. SIFT prediction scores range from 0 to 1, values less and equal to 0.05 were considered deleterious; all other values are considered neutral. PolyPhen-2 prediction scores range from 0 (benign) to 1 (probably damaging), values near to 1 are more confidently predicted to be probably damaging. PROVEAN predicted variants as deleterious when the score is below the threshold value of −2.5 and neutral when it is above this value. Besides, CONDEL predicted the results as deleterious if the score is more than 0.5 and neutral if the score is less and equal to 0.5. PANTHER predicted the length of time (in millions of years) of a position in protein sequence, threshold more than 450 million years is considered as probably damaging, between 450 million years and 200 of millions of years as possibly damaging and less than 200 million years as probably benign. This tool used position-specific evolutionary preservation (PSEP) to determine the length of time a position has been preserved in its ancestors. It would be more likely to have a deleterious impact if the position is in longer preservation. The nsSNPs with greater confidence are expected to be truly deleterious.

In this study, we shortlisted 52 nsSNPs with at least four significant scores out of five algorithmic tools used: score < 0.05 in SIFT, >0.9 in PolyPhen-2, < −2.5 in PROVEAN, 1.0 in CONDEL, and >450 million years in PANTHER. Therefore, only the most deleterious nsSNPs would be studied. Based on the scores, 6 out of 43 nsSNPs in the Orang Asli and 6 out of 43 nsSNPs in the Malays were shortlisted. Interestingly, four nsSNPs were found in both populations (Table 3). As a result, the analysis identified eight deleterious amino acid substitutions responsible for the high-risk nsSNP associated with cancers (Table 3). The nsSNPs which are classified as high-risk are rs3124765, rs9379084, rs10936600, rs1801591, rs117744081, rs2277283, rs757978, and rs1126809. They are located on different genes, which are CACFD1, RREB1, LRRC34, ETFA, CPVL, INCENP, FARP2, and TYR, respectively. According to the GWAS database, the eight (8) nsSNPs were associated with the risk of specific cancers, as shown in Table 3. Thus, these eight (8) nsSNPs were further investigated.

Conservation profile of high-risk nsSNPs. ConSurf was further used to investigate the potential impact of the most deleterious nsSNP. It was used to measure the degree of evolutionary conservation of the protein for each amino acid residue. It identifies amino acid positions known to have functional and structural importance through the combination of evolutionary conservation data and solvent accessibility predictions. In this study, all residues of each protein obtained from Consurf were assigned with conservation levels graded with scores ranging from 1 to 9. However, we concentrated only on residues that mapped to the locations of eight (8) high-risk nsSNPs, which we had identified. The server predicted D1171N, I58M, L286, T171I, Y168H, M506T, R402Q, and T260N as highly conserved (Table 4) and their functional and structural importance. The findings further indicated that these eight (8) high-risk nsSNPs were certainly deleterious to the protein functions and structures.

Predicted stability modification. We predicted the stability modifications due to nsSNPs in CPVL, FARP2, CACFD1, RREB1, LRRC34, ETFA, TYR, and INCENP proteins with the help of I-Mutant. The eight (8) nsSNPs that were found associated with cancers were submitted to the I-Mutant 3.0 server to predict the changes in the stability in terms of their free energy change value ($\Delta\Delta G$) and reliability index (RI). Based on the $\Delta\Delta G$ values, all of these nsSNPs have decreased the stability of the respective proteins (Table 5). However, we had excluded two of them, rs1801591 (RI = 0) and rs117744081 (RI = 4), from analysis as they had RI below five (<5). The higher RI value shows higher accuracy in the prediction for stability. Thus, the other six nsSNPs (rs3124765, rs9379084, rs10936600, rs2277283, rs757978, and rs1126809) were further analyzed.

Homology modeling of protein. The three-dimensional (3D) structures of 6 native and mutant proteins were predicted by I-TASSER. In generating the mutant models, all six sequences were submitted to the I-TASSER, where each nsSNP was substituted into the native sequence. UniProtKB accession numbers for the native sequences used are LRRC34, Q8IZ02; FARP2, O94887; and TYR, P14679. The available top 10 templates protein models in PDB which are structurally closest to the query protein sequence were used to model the native and mutant proteins of LRRC34, FARP2 and TYR using I-TASSER. Among the six predicted models for each query protein (LRRC34, TYR, FARP2), the best model was selected based on the highest confidence score (C-score), as shown in S3 Table. C-score is the score of confidence for the prediction of pairwise comparison.
| SNP ID   | Cancer risk                                           | Location          | Gene Symbol | Amino acid change | SIFT   | PolyPhen-2 | ConDel | PROVEAN       | PANTHER |
|---------|-------------------------------------------------------|-------------------|-------------|-------------------|--------|------------|--------|---------------|---------|
| rs12621643 | Acute lymphoblastic leukemia (childhood)             | 2:223,917,983     | KCNE4       | D145E             | Tol    | benign      | Neu    | Neu           | -       |
| rs13014235 | Basal cell carcinoma                                 | 2:202,215,492     | ALS2CR12    | V43L              | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs1050529  | Basal cell carcinoma                                 | 6:31,324,615      | HLA-B       | A65T              | Del_low_con | benign    | Neu    | Neu           | Prob_ben|
| rs1126805** | Basal cell carcinoma or squamous cell carcinoma      | 11:89,017,961     | TYR         | R402Q             | Del    | Prob_dam    | Del    | Prob_dam      |         |
| rs11543198* | Bladder cancer                                       | 15:74,912,328     | CLK3        | R78H              | Tol_low_con | -         | -      | Neu           |         |
| rs35273427 | Breast cancer                                         | 1:120,436,751     | ADAM30      | T737A             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs6964587  | Breast cancer                                         | 7:91,630,620      | AKAP9       | M463I             | Del    | benign      | Neu    | Neu           | -       |
| rs1053338  | Breast cancer                                         | 3:63,967,900      | ATXN7       | K264R             | Tol    | benign      | Neu    | Neu           | Prob_dam|
| rs1324765  | Breast cancer                                         | 9:136,328,657     | CACFD1      | I58M              | Del    | Prob_dam    | Del    | Neu           | -       |
| rs11552449 | Breast cancer                                         | 1:114,448,389     | DCLRE1B     | H61Y              | Del    | benign      | -      | Neu           | Prob_ben|
| rs3815308  | Breast cancer                                         | 19:2,226,676      | DOT1L       | G1386S            | Tol_low_con | benign    | Neu    | Neu           | Prob_ben|
| rs11205303 | Breast cancer                                         | 1:149,906,413     | MTMR11      | M159Y             | Tol    | benign      | -      | Neu           | -       |
| rs9379084  | Breast cancer                                         | 6:27,231,843      | RREB1       | D1171N            | Del    | Prob_dam    | Del    | Del           | Prob_dam|
| rs8050871  | Breast cancer                                         | 16:71,509,796     | ZNF19       | Q218H             | Del    | pos_dam     | Del    | Neu           | Prob_ben|
| rs5797978** | Chronic lymphocytic leukemia                         | 2:242,371,101     | FAR2P       | T260N             | Del    | Prob_dam    | Del    | Del           | Prob_dam|
| rs11539086** | Colorectal cancer                                     | 3:58,552,329      | FAM107A     | E141Q             | Tol    | Prob_dam    | Del    | Neu           | Prob_dam|
| rs4836891  | Colorectal cancer                                     | 9:125,273,574     | OR1J2       | R165Q             | Tol_low_con | benign    | Neu    | Neu           | Prob_ben|
| rs7248888  | Colorectal cancer                                     | 19:46,974,003     | PNMAL1      | C97Y              | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs16845107 | Colorectal cancer                                     | 3:113,127,991     | WDR52       | K284N             | Tol    | benign      | -      | Neu           | -       |
| rs3184504  | Colorectal or endometrial cancer                      | 12:111,884,608    | SH2B3       | W262R             | Del    | Prob_dam    | Del    | Neu           | Prob_ben|
| rs1129506  | Endometrial cancer                                    | 17:29,646,032     | EVI2A       | S23R              | Del_low_con | benign    | Neu    | Neu           | Pos_dam|
| rs2278686 | Endometriosis or endometrial cancer (pleiogrophy)     | 17:46,262,171     | SKAP1       | G161S             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs1229984 | Esophageal cancer                                     | 4:100,239,319     | ADH1B       | H48R              | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs6871    | Esophageal cancer                                     | 12:112,241,766    | ALDH2       | E504K             | Del    | pos_dam     | Del    | Del           | -       |
| rs2274223 | Esophageal cancer                                     | 10:96,066,341     | PLC1E       | H1927R            | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs3765524 | Esophageal cancer and gastric cancer                  | 10:96,058,298     | PLC1E       | T1777I            | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs20541   | Hodgkin's lymphoma                                    | 5:131,995,964     | IL13        | Q144R             | Tol    | benign      | Neu    | Neu           | -       |
| rs3734542* | Lung cancer in ever smokers                           | 6:26,468,326      | BTNF2A1     | R378Q             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs10936600| Multiple myeloma                                      | 3:169,514,585     | LRRC34      | L286I             | Del    | Prob_dam    | Del    | Neu           | Prob_dam|
| rs7193541 | Multiple myeloma                                      | 16:74,666,743     | RFWD3       | J564V             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs34562254| Multiple myeloma                                      | 17:16,842,991     | TNNFSF13B   | P251L             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs1052501 | Multiple myeloma                                      | 3:41,925,388      | UJK4        | A542P             | Del    | benign      | Neu    | Neu           | Prob_ben|
| rs2272007 | Multiple myeloma (hyperdiploidy)                     | 3:41,996,136      | UJK4        | K39R              | Del    | benign      | Neu    | Neu           | Prob_dam|
| rs6793295  | Multiple myeloma and monoclonal gammopathy           | 3:169,518,455     | LRRC34      | S249G             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs1801591 | Non-glioblastoma glioma                              | 15:76,578,762     | ETFA        | T171I             | Del    | Prob_dam    | Del    | Del           | -       |
| rs117744081* | Non-melanoma skin cancer                            | 7:29,132,279      | CPVL        | Y168H             | Del    | Prob_dam    | Del    | Del           | Prob_ben|
| rs11170164** | Non-melanoma skin cancer                            | 12:52,913,668     | KRT5        | G138E             | Del    | pos_dam     | Del    | Del           | Pos_dam|
| rs1229984 | Oral cavity and pharyngeal cancer                    | 4:100,239,319     | ADH1B       | H48R              | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs1492961 | Oral cavity and pharyngeal cancer                    | 4:84,374,480      | HELQ        | V306I             | Tol    | benign      | Neu    | Neu           | -       |
| rs763780  | Pancreatic cancer                                    | 6:52,101,739      | IL17F       | H161R             | Tol    | benign      | Neu    | Neu           | Del          |
| rs2257205 | Pancreatic cancer                                    | 17:56,488,297     | RNF43       | R117H             | Tol    | pos_dam     | Del    | Neu           | Pos_dam|
| rs3795244 | Pancreatic cancer                                    | 17:30,692,396     | ZNF207      | A240S             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs130067* | Prostate cancer                                      | 17:20,817,099     | CDKN1B      | V109D             | Tol    | benign      | Neu    | Neu           | Prob_ben|
| rs2066827 | Prostate cancer                                      | 11:61,908,440     | INCENP      | M506T             | Del    | Prob_dam    | Del    | Del           | Prob_dam|
| rs2292884 | Prostate cancer                                      | 2:238,443,226     | MLPH        | H347R             | Tol    | benign      | Neu    | Neu           | Prob_ben|

Continued
with values ranging from −5 to 2. A greater level of C-score indicates a model with great confidence and vice-versa. Then, PyMol was used to visualize the structures. Structural analysis of demonstrated that the mutants of LRRC34, FARP2, and TYR had structures with deviated orientation compared to the native LRRC34, FARP2, and TYR, respectively (Fig. 2). Compared to the native structure of LRRC34, FARP2, and TYR proteins, their mutant structures have more helixes as presented in Table 6. The numbers of beta-sheets were also different between the native and mutant proteins. The native protein structure of LRRC34 and TYR have more beta-sheets when compared to their mutants. In contrast, the native protein structure of FARP2 has three fewer beta-sheets than its mutant. There are three and two more buried residues in the native LRRC34 (432) and FARP2 (1007) proteins compared to their mutants, respectively. However, buried residues in the native TYR (509) are less than its mutant protein.

TM-scores and RMSD values of each mutant model were calculated using TM-align. TM-score measures the similarity of topological models for native and mutant proteins, whereas RMSD evaluated the average distance from native α-carbon backbones to mutant models. The mutant model with the highest TM-score value is T171I (0.975), followed by R402Q (0.938), L286I (0.934), T260N (0.929), and Y168H (0.909). The highest TM-score value indicates that the mutant models generated are still in the same folding dimension of the native models but not perfectly the same. Besides, these mutant models were found to be different from the native based on RMSD values shown in Table 5. The nsSNP models of I58M, D1171N, and M506T have very low TM-score values of 0.346, 0.319, and 0.262, respectively, which correspond to randomly chosen unrelated proteins. Hence, we

| SNP ID | Cancer risk          | Location | Gene Symbol | Amino acid change | SIFT | PolyPhen-2 | ConDel | PROVEAN | PANTHER |
|--------|----------------------|----------|-------------|-------------------|-----|------------|--------|---------|---------|
| rs11071896 | Testicular germ cell tumor | 15:66,821,250 | ZWILCH | S344G | Tol | benign | Neu | Neu | Prob_ben |
| rs6793295 | Thyroid cancer | 3:169,518,455 | LRRC34 | S249G | Tol | benign | Neu | Neu | Prob_ben |

Table 3. List of 52 nsSNPs identified among the Orang Asli and the Malays and functional effect predicted by five in silico programs. Del = Deleterious, Tol = Tolerated, Pro_dam = Probably damaging, Pos_dam = Possibly damaging, Prob_ben = Probably benign, Neutral = Neu,—= Not predicted. *nsSNPs which are found in Orang Asli only. **nsSNPs which are commonly found in Malays only. The highlighted rows were the selected nsSNPs for further investigation.

| SNP ID | UniprotKb Accession Number | Amino Acid Change | Conservation Score | Prediction |
|--------|----------------------------|-------------------|--------------------|------------|
| rs9379084 | Q92766 | D1171N | 9 | Highly conserved |
| rs3124765 | Q9UGQ2 | IS8M | 8 | Highly conserved |
| rs10936600 | Q8IZ02 | L286I | 9 | Highly conserved |
| rs1801591 | P13804 | T171I | 9 | Highly conserved |
| rs117744081 | Q9H3G5 | Y168H | 8 | Highly conserved |
| rs2277283 | Q9NQS7 | M506T | 9 | Highly conserved |
| rs1126809 | P14679 | R402Q | 8 | Highly conserved |
| rs757978 | O94887 | T260N | 8 | Highly conserved |

Table 4. Conservation profile of amino acids in proteins with high-risk nsSNPs by ConSurf. 1 ≤ conservation score ≤ 4 = variable, 5 ≤ conservation score ≤ 6 = intermediate, and 7 ≤ conservation score ≤ 9 = highly conserved.

| nsSNP ID | Amino Acid Change | Gene Symbol | Stability | RI | ΔΔG (kcal/mol) | TM-Score | RMSD (Å) |
|----------|-------------------|-------------|-----------|----|----------------|----------|----------|
| rs3124765 | IS8M | CACFD1 | Decrease | 8 | −1.19 | 0.346 | 4.84 |
| rs9379084 | D1171N | RREB1 | Decrease | 7 | −1.74 | 0.319 | 4.41 |
| rs10936600 | L286I | LRRC34 | Decrease | 5 | −1.00 | 0.934 | 2.06 |
| rs1801591 | T171I | ETFA | Decrease | 0 | −0.48 | 0.975 | 1.13 |
| rs117744081 | Y168H | CPVL | Decrease | 4 | −1.50 | 0.909 | 2.66 |
| rs2277283 | M506T | INCENP | Decrease | 6 | −0.88 | 0.262 | 2.56 |
| rs757978 | T260N | FARP2 | Decrease | 5 | −1.01 | 0.929 | 3.21 |
| rs1126809 | R402Q | TYR | Decrease | 9 | −1.39 | 0.938 | 2.56 |

Table 5. I-Mutant 3.0 and TM-align predictions for nsSNPs associated with cancers among the Orang Asli and Malays. RI = Reliability Index. RMSD = Root Mean Square Deviation. ΔΔG ≤ −0.5 kcal/mol = destabilizing mutations, ΔΔG > −0.5 kcal/mol = stabilizing mutations, −0.5 kcal/mol ≤ ΔΔG ≤ 0.5 kcal/mol = neutral mutations. 0.0 < TM-score < 0.30 = random structural similarity, 0.3 ≤ TM-score ≤ 1.00 = in about the same fold 0.5 ± 1. Highlighted rows are the excluded nsSNPs.
Figure 2. Graphical representations of amino acid changes due to the most deleterious nsSNPs and close-up view for substitution of amino acids (green = native residue; red = mutant residue). (a) Superimposed structures of native LRRC34 protein and its mutant with substitution from Leucine to Isoleucine at position 286. (b) Superimposed structures of wild type FARP2 protein and its mutant having substitution from Threonine to Asparagine at position 260. (c) Superimposed structures of the native TYR protein and its mutant having substitution from Arginine to Glutamine at position 402.
Finally selected only three mutants L286I, T260N, and R402Q for further analysis, based on the results provided by I-Mutant and TM-align (Table 5).

The modeled structures were validated using ERRAT program and Ramachandran Plot Server to check the reliability of predicted protein structures. The ERRAT results showed that the qualities for the native and mutant LRRC34, FAR2P, and TYR protein were good with scores of 93.86, 57.17, 75.15, 87.28, 70.27, and 73.51, respectively (Table S3). Ramachandran plots for the native and mutant LRRC34, FAR2P, and TYR protein models showed 87.74%, 71.75%, 85.00%, 87.50%, 69.16, and 85.00% of the residues were located in the allowed regions, and only a few amino acids were deviated (Table S3).

Those three selected mutant protein models were then superimposed on the native protein models to show the location of observed mutations (Fig. 2). The details of the selected native and mutant protein models included the protein templates used to predict the structures and C-score are provided in the Table S3.

### Functional and structural modifications of genetic variants.

Three (3) nsSNPs were shortlisted and submitted to the MutPred2 server. MutPred2 predicts the modification of structural and functional protein structures, including the altered order or disordered interface, transmembrane protein, metal binding, DNA binding, loss of allosteric site, and gain of allosteric site. Based on Table 7, the R402Q mutation showed the highest probability score (0.78), followed by T260 mutation (0.73) and L286 mutation (0.55). An amino acid substitution is predicted as pathogenic if a probability score is 0.50 and above.

The HOPE was further used to explore the structural effects of these three amino acid substitutions. It was shown that the substitution of L286, T260, and R402 were highly conserved. Based on Fig. 3, the L286I mutation is predicted as pathogenic if a probability score is 0.50 and above.

### Table 6. The top 10 templates used for homology modeling, and the alpha helix, beta sheet and exposed/buried residues used by I-TASSER.

| Gene Symbol | Mutation | Template | C-score |
|-------------|----------|----------|---------|
| LRRC34 | L286I | 1a4yA, 2bnh, 4perA, 4k17A, 6hsBa, 4kxIK, 2p1pB, 3ognB, 5hywA, 4q62A | 0.55 |
| FAR2P | T260N | 4gzuA, 4ly6A, 3vkhA, 6ehBA, 5saCA, 3j9BA, 6ar6A, 6busA, 5h64A, 5d06A | 0.70 |
| TYR | R402Q | 4gzuA, 4ly6A, 3vkhA, 6ehBA, 5saCA, 3j9BA, 6ar6A, 6busA, 5h64A, 5d06A | 0.74 |

### Table 7. Probability scores and top prediction features of deleterious mutations by MutPred2 and ModPred.

| Gene Symbol | Mutation | Top Prediction Features | Score | PTMs | Score |
|-------------|----------|-------------------------|-------|------|-------|
| LRRC34 | L286I | Altered Ordered interface (P-value = 0.01) | 0.55 | Proteolytic cleavage | 0.07 |
| FAR2P | T260N | Altered DNA binding (P-value = 2.8e−03) | 0.70 | Proteolytic cleavage | 0.49 |
| TYR | R402Q | Altered Disordered interface (P-value = 0.03) | 0.74 | Proteolytic cleavage | 0.58 |
arthritis, cancer, cardiovascular disease, and inflammation. This represents a remarkably significant prediction by ModPred (Table 7).

**Protein–protein interactions analysis.** The STRING server was used to investigate the interaction of HLA-G with various proteins. The interaction analysis revealed that LRRC34 is related to Leucine-rich repeat-containing 32 (LRRC32), Leucine-rich repeat containing 31 (LRRC31), Leucine-rich repeats and IQ motif containing 4 (LRRIQ4), Actin related protein T3 (ACTRT3), Myoneurin (MYNN), Protein FAM196B (FAM196B), Transmembrane protein 174 (TMEM174), Ly6/PLAUR domain-containing protein 6 (LYPD6), Aspartyl aminopeptidase (DNPEP) and DAZ-associated protein 1 (DAZAPI) as shown in Fig. 4.

While FARP2 is related to cell division control protein 42 homolog (CDC42), Proto-oncogene tyrosine-protein kinase Src (SRC), Tyrosine-protein kinase Fyn (FYN), Neuropilin-1 (NRP1), Plexin-A1 (PLXNA1), Plexin-A2 (PLXNA2), Plexin-A3 (PLXNA3), Plexin-A4 (PLXNA4), Semaphorin-3A (SEMA3A), and Tyrosine-protein kinase Fes/Fps (FES) as shown in Fig. 4.

The interaction analysis also revealed that TYR is related to Short transient receptor potential channel 1 (TRPC1), Tyrosine 3-monoxygenase(TH), Phenylalanine hydroxylase (PAH), Aromatic-L-amino-acid decarboxylase (DDC), Thyroid peroxidase (TPO), L-dopachrome tautomerase (DCT), Melanocyte protein PMEL (PMEL), Melanoma antigen recognized by T-cells 1 (MLANA), P protein (OCA), and Microphthalmia-associated transcription factor (MITF) as shown in Fig. 4.

**Molecular docking analysis.** Autodock Vina, UCSF Chimera 1.15 tools predicted and evaluated a total of 10 protein binding sites along with hydrogen bond interaction and their binding affinities from the docking analysis. The resulting interactions between the native and mutant LRRC34, FARP2, and TYR were compared with those calculated docking results in the same protein binding sites using the exact dimensions of the grid boxes. Thus, a binding site was predicted for each receptor-ligand docking. Molecular docking of SRC, DCT, and MYNN with native and mutant FARP2, TYR, and LRRC34 modeled structures showed differences in the binding affinities (Table 8). The binding affinity of SRC with native FARP2 was $-8.2$ kcal/mol, while for mutant was $-7.8$ kcal/mol. The binding affinity of DCT with native TYR was $-8.1$ kcal/mol, while for mutant was $-8.0$ kcal/mol. The binding affinity of MYNN with native LRRC34 was $-5.4$ kcal/mol, while for mutant L286I was $5.2$ kcal/mol. In
addition, SRC, DCT, and MYNN were bound to the same binding pockets for the native and mutant FARP2, TYR, and LRRC34 proteins, respectively. From the analysis of the binding pose, these three proteins (SRC, DCT, and MYNN) showed significant deviations between the native and mutant protein complexes (Fig. 5). Moreover, interaction analysis of SRC, DCT, and MYNN with the native and mutant FARP2 TYR and LRRC34 proteins showed a reduction in the number of hydrogen bonds with residues in mutant proteins (Table 8). Five residues such as Lys68, Tyr65, Leu5, Ser164, and Gln167 have interactions with SRC in native FARP2 but were absent in mutant proteins. Three residues, Lys152, Ser134, and Lys152, interact with DCT in native TYR but were absent in mutant proteins. Two residues, Asn39 and Ala42, have interactions with MYNN in native LRRC34, but Asn39 was absent in mutant protein.

Discussion

The exponential increase in the number of nsSNPs detected makes the investigation of the biological significance of each nsSNP by wet laboratory experiments impossible. Alternatively, in silico programs may be used to predict the effects due to mutations and explain the underlying biological mechanisms. nsSNPs in the coding regions can lead to amino acid change and alterations in protein function and account for susceptibility to disease. Identification of deleterious nsSNPs from tolerant nsSNPs is important in analyzing individual susceptibility to disease and understanding disease pathogenesis.

In this study, we have developed a pipeline (Fig. 1) to identify the pathogenic nsSNPs associated with cancers. Although there are various computational tools available to predict the deleterious or damaging effects of nsSNPs on protein structure and function, we had used five different tools (SIFT, PolyPhen-2, Condel, PROVEAN, and PANTHER) to determine the nsSNPs functional effects, while Consurf was used to estimate the evolutionary conservation of the amino/nucleic acid positions in a protein/DNA and protein. I-Mutant 3.0 was used to predict the impact of nsSNPs on the functions or structures of the pathogenic proteins. Among them, SIFT algorithm is the most commonly used tool for SNP characterization to determine deleterious nsSNPs. This method computes a conservation score that provides an insight into the impact of nsSNPs on the functional property of proteins. PolyPhen-2 is considered one of the most reliable tools to predict the functional impact of nsSNPs based on protein sequence, phylogenetic information, and structural information. Condel on the other hand integrates and reflects the combination of scores from different methods (SIFT, PolyPhen2, Mutation Assessor, FATHMM)
classify the nsSNPs. It provides insight into the impact of the mutation on the biological activities of the proteins affected. PROVEAN algorithm is capable of predicting the functional impacts of the amino acid substitution on a protein sequence with commensurable performance and accuracy. It utilizes alignment-based scores to measure the change in sequence variation correlated with the biological function of a protein. Additionally, SIFT, PolyPhen-2, Condel, and PROVEAN, are easy and quick to employ, which allows direct batch queries. Other tools include PANTHER, a powerful and unique method with a curated database of protein families, trees, subfamilies and functions, and evolutionary relationships. It uses phylogenetic trees, multiple sequence alignments, and statistical technique to evaluate the deleterious effects of nsSNPs, making it a viable platform for SNP characterization. Consurf is another widely used tool that can pinpoint critically important sites (nsSNPs) within the functional regions. It is a statistically robust approach that estimates the evolutionary rates due to amino acids substitutions and maps them onto the homologous sequence and/or structures. I Mutant 3.0 tool measures the change in protein-free energy caused by a specific mutation. It helps to detect the changes in protein 3D conformation stability. The tools used in this study cover a wide range of prediction techniques (Table 1), combining the findings from each tool in the pipeline will help to identify the most deleterious nsSNPs more accurately. Specific targeted genotyping assays could be developed to detect these nsSNPs identified to be impactful and further investigated in a local cohort of cancer patients. The prediction can also help scientists to focus their study on understanding the impact of these nsSNPs by prioritizing the most deleterious nsSNPs.

Figure 5. Images of the superimposed native and mutant structural models docked against target proteins with high probabilities values that affect protein functions. (a) Superimposed image of SRC (orange) docked against native (blue) and mutant (green) FARP2 protein and interaction of SRC with (b) native and (c) mutant FARP2 protein structures. (d) Superimposed image of DCT (blue) docked against native (yellow) and mutant (pink) TYR protein and interaction of DCT with (e) native and (f) mutant TYR protein structures. (g) Superimposed image of MYNN (purple) docked against native (blue) and mutant (red) LRRC34 protein and interaction of MYNN with (h) native and (i) mutant LRRC34 protein structures. Hydrogen bonds are presented in a straight blue line.
The bioinformatics workflow developed was validated using the breast cancer dataset from ClinVar, which acts as a standard dataset. The standard dataset has been annotated and we believe it is the most appropriate dataset for functional effect prediction. The standard dataset contained a total of 100 nsSNPs that were clinically associated with breast cancer (Table S1). The sensitivity, specificity, and accuracy of four models (Model A, B, C, and D) in predicting the clinical significance were determined. Model D represents at least four tools that predicted nsSNPs as deleterious or benign, and it showed the highest percentages of specificity (94%), and accuracy (89%), followed by Model C (specificity 80%, and accuracy 85%), Model B (specificity 64%, and accuracy 78%) and Model A (specificity 50%, and accuracy 75%). While Model A has the highest sensitivity (100%) followed by Model B (92%), Model C (90%), and Model D (84%). The highest sensitivity scores mean that fewer potentially deleterious nsSNPs were missed. Thus, we concluded that Model D using at least four out of five tools had the best performance in predicting the most deleterious nsSNPs.

Further analyses using the combination of five functional effect tools with conservation and stability tools showed that Model D3 had the highest specificity (96%), but the lowest sensitivity (76%) in identifying deleterious and benign nsSNPs. Despite not having the highest accuracy, Model D3 was able to classify both pathogenic and benign SNVs accurately (86%). The validated workflow is adequate with good sensitivity, specificity, and accuracy to classify the deleterious and neutral nsSNPs in ClinVar using a combination of SIFT, PolyPhen-2, Condel, PROVEAN, PANTHER, Consurf, and I-Mutant.

The GWAS database was used to identify nsSNPs associated with cancer risks as it is the most extensive SNPs database. We only focused on nsSNPs as they are capable of altering protein function, structure, conformation, and interaction which cause the increased risk of cancer. Out of the 80 nsSNPs associated with cancer risks from the GWAS dataset, a total of 52 nsSNPs were identified among the Orang Asli and Malays (43 in Orang Asli and 43 in Malays). They were subjected for further analysis.

Hence, we conducted the concordance analysis with SIFT, PolyPhen-2, Condel, PROVEAN, PANTHER, Consurf, I-Mutant, ModPred, and MutPred tools to predict the most deleterious nsSNPs among the Orang Asli and Malays (Table 3). From the functional effect prediction analysis, a total of 8 out of 52 nsSNPs which were associated with cancers from both populations were identified as the most deleterious nsSNPs by SIFT, PolyPhen-2, Condel, PROVEAN, and PANTHER (Table 3). The most deleterious nsSNPs were identified based on the criteria that at least four scores out of five algorithmic tools used were significant, which are score < 0.05 in SIFT, > 0.9 in PolyPhen-2, < −2.5 in PROVEAN, 1.0 in Condel, and > 450 million years in PANTHER. The identified nsSNPs were rs3124765 (CACFD1), rs9379084 (RREB1), rs10936600 (ETFA) rs1801591 (LRRC34), rs17744081 (CPVL), rs2277283 (INCENP), rs757978, (FARP2) and rs1126809. (TYR). In terms of the useability of these five tools for prediction, different algorithms for evolutionary conservation, protein function or structure, alignment, and measurement of similarity between variant sequence and protein sequence homologs were analyzed. Hasan et al. had reported that the combination of the best individual tools, FATHMM, iFish, and Mutation Assessor, in one classifier called Meta (Combined Scores through J48 "CSTJ48") enhances the predictive power of these tools. However, no specific classifier outperforms overall datasets in pathogenic predictability. Additionally, these tools have proven performance in identifying deleterious nsSNPs, and these make them useful for our study. Thus, these eight (8) nsSNPs identified were further investigated.

The Consurf server had predicted the eight (8) variations, D1171N, I58M, L286, T171I, Y168H, M506T, R402Q, and T260N, were highly conserved (Table 4), and this emphasizes their functional and structural importance. Evolutionary information is essential to understand the mutations potentially affect human health. The evolution of amino acids influence their properties such as size, shape, hydrophobicity, and charge of amino acids at the molecular level. For example, 53 missense mutations that caused cystic fibrosis were found within highly conserved positions. These regions were significant for conserving the structural and functional integrity of the CFTR protein. Besides, functional sites of proteins like DNA interaction sites, protein–protein interaction sites, and enzymatic sites are essential for biological functions. This may suggest that the nsSNPs found in these conserved regions have higher deleterious effects than other non-conservative nsSNPs and may significantly affected the biological functions. The findings further indicated that these eight (8) high-risk nsSNPs were indeed deleterious to the protein functions and structures.

I-Mutant predicts the protein stability of mutants based on the free energy change value (ΔΔG) and reliability index (RI). I-Mutant predicted 6 out of 8 variants (rs3124765, rs9379084, rs10936600, rs2277283, rs757978, and rs1126809) to have decreased stability. Protein stability is important for the protein structural and functional behavior. Protein stability affects the conformational structure of the protein, such as protein misfolding, aggregation, and degradation, and thus determines its function. From the results, we believe that the six variants might had affected the proteins function by affecting their stability.

For structural analysis, the six native and mutant protein structures (CACFD1, RREB1, LRRC34, INCENP, FARP2, and TYR) were successfully generated using I-TASSER as there are no available close homologous templates. I-TASSER generates full-length models by the iterative structural fragment reassembly method, which consistently drives the threading alignment relative to the native state. They were then verified by ERRAT and Ramachandran Plot Server, which proved the stability, reliability, and consistency of the tertiary structures of the proteins. The three-dimensional structures for the native and mutant proteins predicted by I-TASSER clearly revealed the structural changes resulting from amino acids substitutions (Fig. 2). Furthermore, the changes predicted on the sequence-based homology modeling between the native and mutant on the LRRC34, FARP2, and TYR proteins, support the prediction of the pathogenicity of the deleterious substitutions.

TM-align were utilized to calculate the comparison between the predicted native and mutant protein structures based on TM-score and RMSD value. In most cases, common protein structure modeling tools may construct realistic full-length models with an RMSD value less than 6.5 Å if alignment has a TM-score of more than 0.5. Following the criteria of RMSD < 6.5 Å and TM-score > 0.5, three mutants, 158M (CACFD1), D1171N (RREB1), and M506T (INCENP) with TM-scores below 0.5, were excluded. TM-scores below 0.5 correspond to
numerous studies have shown that this alteration can significantly change the protein function by modifying its structures. Although the effect of modification at proteolytic cleavage sites on these proteins has still not been published, it can alter the ability of proteins to interact with other proteins. For example, the changes at the PTM site might disrupt the protein interaction, leading to cancers. A previous study had reported that mutations of melanogenic enzyme tyrosinase (TYR) result in hypopigmentation of the hair, skin and eyes. Besides, DCT is one of the factors that could probably increase the damage caused by PTM impairment.

The three shortlisted nsSNPs were submitted to MutPred2, HOPE, and ModPred tools to predict the modification of structural and functional protein structures. MutPred2 predicts the modification of structural and functional protein structures, including the altered ordered or disordered interface, transmembrane protein, metal binding, DNA binding, loss of allosteric site, and gain of allosteric site. HOPE was used to further explore the structural effects of these three amino acid substitutions. It was shown that the substitution of L286I, T260N, and R402Q in LRRC34, FARP2, and TYR proteins, respectively (Table 7). The scores given by MutPred for this PTM was very low for proteolytic cleavage (Table 7). The LRRC34 is a nucleolar protein that plays a role in the ribosome biogenesis of pluripotent stem cells. Mutations in some of the related proteins or modifications at ribosome biogenesis may result in severe implications for the organism, depending on the degree of the modification and the involvement of the tissue. The changes at the PTM site might influence the structure of LRRC34 protein, which may lead to multiple myeloma. For example, impaired or modified ribosome synthesis due to the mutation of the ribosomal proteins was reported in many cancers such as chronic lymphocytic leukemia, colorectal cancers, and glioma. FARP2 has been reported as a potential regulator of chronic lymphocytic leukemia pathogenesis that influences protein activity encoded by MYC gene. MYC gene is known as a proto-oncogene and produces a nuclear phosphoprotein that plays a role in the cell cycle progression, apoptosis, and cell transformation. The mutation may disrupt the MYC protein activity. Although the effect of modification at proteolytic cleavage sites on these proteins has still not been published, numerous studies have shown that this alteration can significantly change the protein function by modifying its position, stability, or inter-protein interactions others. Proteolytic cleavage of modified residues in the protein may be necessary for some of the essential functions of the protein. Besides, those nsSNPs can disrupt proteins that could probably increase the damage caused by PTM impairment.

Protein–protein interaction network analysis showed the interactions of LRRC34, FARP2, and TYR with ten different proteins. This analysis is important in predicting the functionality of interacting genes or proteins and understanding the functional relationships and evolutionary conservation of the interactions among the genes. Besides, our literature search demonstrated that LRRC34, FARP2, and TYR interact with other proteins. LRRC34 interacts with two major nucleolar proteins, Nucleophosmin (NPM1) and Nucleolin (NCL), in ribosome biogenesis of pluripotent stem cells. The mutation in LRRC34 might affect ribosome biogenesis and lead to tumorigenesis. FARP2 interacts with PLXN4, SEMA3A, and NRP1 in Sema3A-Nrp1/PlxnA4 signaling pathway that controls dendritic morphogenesis. The mutation in FARP2 might disrupt the formation of axonal and dendritic morphologies for the neurodevelopment that ultimately lead to risks of cancers. TYR interacts with TH, MITF, and PAH in the melanogenesis pathway. Due to the nonsynonymous mutation in TYR, the melanin synthesis might be disrupted, leading to tumorigenesis. Therefore, any changes in these protein function/structure would have an impact on many disease pathways.

The structural analysis was performed by using molecular docking. The study aims to identify the correct poses of ligands in the binding pocket of a protein and to predict the affinity between the ligand and the protein, which may enhance or inhibit its biological function.

The molecular docking analysis of SRC, DCT, and MYNN with native and mutant FARP2, TYR, and LRRC34 modeled structures showed a difference in binding affinity, reduction in the number of hydrogen bonds with residues in mutant proteins (Table 8), and a significant deviation between native and mutant protein complexes (Fig. 5), respectively. SRC proto-oncogene plays an essential role in development, growth, progression, and metastasis of some human cancers, including those of the colon, breast, pancreas, and brain. FARP2 were identified as guanine nucleotide exchange factors (GEFs) for RhoGTPases that play regulatory roles in neuronal development, and several studies have revealed the genetic alterations in Ras homologous RhoGEFs in several human cancers. Thus, the deviation observed in the bound SRC molecule with mutant FARP2 protein might disrupt the protein interaction, leading to cancers. A previous study had reported that mutations of melanogenic enzyme tyrosinase (TYR) result in hypopigmentation of the hair, skin and eyes. Besides, DCT is one of the proteins that could participate in the melanogenesis pathway.
the related enzymes that catalyzes different post-TYR reactions in melanin biosynthesis. TYR and DCT also have been proposed to interact with and stabilize each other in multi-enzyme complexes. Thus, the deviation observed in the bound DCT molecule can reduce the catalytic efficiency of TYR. LRRC34 is a member of the leucine-rich repeat-containing protein family that has been suggested to be implicated in the maintenance and regulation of pluripotency. MYNN protein is a member of the BTB/POZ and zinc finger-containing family involved in transcriptional regulation. It has also been shown to interact with a few other proteins, including LRRC34, which are part of the transcription factors that participate in DNA repair. A study showed that disruption of LRRC34 protein function could result in reduced expression of some pluripotency genes. Its altered expression impacts the pluripotency-regulating genes and interacts with other proteins known to be involved in ribosome biogenesis. This molecular docking analysis further evaluates our hypothesis as to whether T260N, R402Q, and L238I mutants have deleterious effects on FAR2P, TYR, and LRRC34 proteins, respectively. The most prominent change was noticed in T260N, R402Q, and L238I, where a significant loss of H-bond interactions within the binding pocket residues can be observed compared to that in the native protein. These H-bonds were disrupted when the amino acid in mutants was replaced with other amino acids, which altered the binding affinity. The change in the number of hydrogen bonds indicates the deleterious effect of amino-acid substitution. Therefore, an increase or decrease of hydrogen bonds of the native form could destabilize the protein and affect protein functions. As a result, genetic mutation which alters the protein structure, therefore influences how the protein interacts with its ligands, potentially leading to a disease condition. This method has been previously used to discover functionally significant variants that may play a role in disease mechanisms. Molecular docking analysis conducted in this study revealed that T260N, R402Q, and L238I mutants could significantly affect the functional activity of FAR2P, TYR, and LRRC34 proteins, respectively.

Conclusion

With the advancement of genomics, predicting and preventing diseases that are preventable will definitely bring a new facet to medical practice. We had illustrated that with the availability of a local genome database, we could predict disease risks in our population using a validated bioinformatics pipeline and the established GWAS and ClinVar database. The pipeline will help strategize experimental research to prioritize studies on the SNPs with predicted functional impact as thousands and millions of SNPs with unknown functions are detected using whole-genome sequencing technologies.

In this study, a bioinformatics pipeline was developed and validated to predict the effects of nsSNPs, rs1126809, rs757978, and rs10936600 on the functional and structural changes on TYR, FAR2P, and LRRC34 proteins, respectively. The analysis also provides significant insight into the deleterious effects of these nsSNPs on the protein structures.

These three (3) nsSNPs were predicted to confer high risks of multiple myeloma, chronic lymphocytic leukemia, and basal cell carcinoma or squamous cell carcinoma in the Orang Asli and Malays population. The prediction pipeline developed in this study helps to reduce the number of extensive investigations and wet lab experiments which are required to explain the impacts of these nsSNPs on the structures and functions of these proteins. We intend to analyze further the risks conferred by these SNPs in the cancer patients in the local population.

We believed that a similar approach could be used to develop and validate bioinformatics pipelines in annotating and predicting the functional effects of SNPs related to other diseases. This study also allows us to establish a database of predicted phenotypes based on the new SNPs identified in our population.

Received: 25 April 2021; Accepted: 26 July 2021
Published online: 09 August 2021

References

1. Collins, F. S., Brooks, L. D. & Chakravarti, A. Erratum: A DNA polymorphism discovery resource for research on human genetic variation (Genome Research (1998) 8 (1229–1231)). Genome Res. 9, 210 (1999).
2. Capriotti, E. & Altman, R. B. Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics 12, S3 (2011).
3. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 14, S3 (2013).
4. Petukh, M., Kucukkalk, T. G. & Alexov, E. On human disease-causing amino acid variants: statistical study of sequence and structural patterns. Hum. Mutat. 36, 524–534 (2015).
5. Chasman, D. & Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: Structure-based assessment of amino acid variation. J. Mol. Biol. 307, 683–706 (2001).
6. Kucukkalk, T. G., Petukh, M., Li, L. & Alexov, E. Structural and physico-chemical effects of disease and non-disease nsSNPs on proteins. Curr. Opin. Struct. Biol. 32, 18–24 (2015).
7. Lander, E. S. The New Genomics: Global Views of Biology. Science (80-) 274, 536–539 (1996).
8. AbdulAzeez, S. & Borgio, J. F. In-silico computing of the most deleterious nsSNPs in HBA1 gene. PLoS ONE 11, 1–13 (2016).
9. Akhtar, M. et al. Identification of most damaging nsSNPs in human CCR6 gene: In silico analyses. Int. J. Immunogenet. https://doi.org/10.1111/ijj.12449 (2019).
10. Badgjur, N. V., Tarapara, B. V. & Shah, F. D. Computational analysis of high-risk SNPs in human CHK2 gene responsible for hereditary breast cancer: A functional and structural impact. PLoS ONE 14, e0220711 (2019).
11. Chakraborty, R., Gupta, H., Rahman, R. & Hasija, Y. In silico analysis of nsSNPs in ABCB1 gene affecting breast cancer associated protein P-glycoprotein (P-gp). Comput. Biol. Chem. 77, 430–440 (2018).
12. Datta, A., Mazumder, M. H., Chowdhury, A. S. & Hasan, M. A. Functional and structural consequences of damaging single nucleotide polymorphisms in human prostate cancer predisposition gene RNASEL. Biomed. Res. Int. 2015, 1 (2015).
13. Wang, Q. et al. Computational screening and analysis of lung cancer related non-synonymous single nucleotide polymorphisms on the human kirsten rat sarcoma gene. Molecules 24, 1951 (2019).
14. Abduljalil, Z. Structural and Functional Analysis of human lung cancer risk associated hOGG1 variant Ser326Cys in DNA repair gene by molecular dynamics simulation. *Non-coding RNA Res.* **4**, 109–119 (2020).

15. Rajasekaran, R., Sudandiradoss, C., Doss, C. G. P. & Sethumadhavan, R. Identification and in silico analysis of functional SNPs of the BRCA1 gene. *Genomics* **90**, 447–452 (2007).

16. Chandrasekaran, G. et al. Computational modeling of complete HOXB13 protein for predicting the functional effect of SNPs and the associated role in hereditary prostate cancer. *Sci. Rep.* **7**, 1–18 (2017).

17. Amberg, A. In Silico Methods. (Springer, 2013)https://doi.org/10.1007/978-3-642-25240-2_55

18. International & T., Mutation, A., Savige, J. & Ars, E.,. Welter, D. Identification and in silico analysis of functional SNPs of the BRCA1 gene. *Genomics* **90**, 447–452 (2007).

19. Ritter, D. I.

20. Rajasekaran, R., Sudandiradoss, C., Doss, C. G. P. & Sethumadhavan, R. Identification and in silico analysis of functional SNPs of the BRCA1 gene. *Genomics* **90**, 447–452 (2007).

21. Al-Obaiade, M. A. I., Ibrahim, B. A., Al-Humaisi, S. & Abdel-Salam, A.-S.G. Genomic and Bioinformatics approaches for analysis of genes associated with cancer risks following exposure to tobacco smoking. *Front. Public Health* **6**, 1–7 (2018).

22. Liu, Y., Yi, Y., Wu, W., Wu, K. & Zhang, W. Bioinformatics prediction and analysis of hub genes and pathways of three types of gynecological cancer. *OncoLett.* **18**, 617–628 (2019).

23. Thomas, P. D. et al. PANTHER-A library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).

24. Sim, N. L. et al. SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, 452–457 (2012).

25. Vaser, R., Adusumalli, S., Leng, S. N., Sikk, M. & Ng, P. C. SIFT missense predictions for genomes. *Nat. Protoc.* **11**, 1–9 (2016).

26. Ramensky, V. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* **30**, 3984–3990 (2002).

27. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet. https://doi.org/10.1002/0471142950.hg0720s76 (2013).

28. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. 459–473 (2005) https://doi.org/10.1016/j.jmb.2005.08.020.

29. Yue, P. & Moult, J. Identification and Analysis of Deleterious SNPs. 1263–1274 (2006) https://doi.org/10.1016/j.jmb.2005.12.025.

30. Kerr, I. D. et al. Assessment of in silico protein sequence analysis in the clinical classification of variants in cancer risk genes. *J. Community Genet.* **8**, 87–95 (2017).

31. Dobson, R. J., Munroe, P. B., Caulfield, M. J. & Saqi, M. A. S. Predicting deleterious nsSNPs: An analysis of sequence and structural attributes. *BMC Bioinformatics* **7**, 3–11 (2006).

32. Krishnan, V. G. & Westhead, D. R. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* **19**, 2199–2209 (2003).

33. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.* **30**, 1237–1244 (2009).

34. Calabrese, R., Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729–2734 (2006).

35. Kulkarni, V., Errami, M., Barber, R. & Garner, H. R. Exhaustive prediction of disease susceptibility to coding base changes in the human genome. *BMC Bioinformatics* **9**, 1–10 (2008).

36. Tian, J. et al. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics* **8**, 8–9 (2007).

37. Thusberg, J., Olatawosen, A. & Vihinen, M. Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.* **32**, 358–368 (2011).

38. Kumar, A., Rajendran, V., Sethumadhavan, R. & Purohit, R. Identifying novel oncogenes: A machine learning approach. *Inter-discip. Sci. Comput. Life Sci.* **5**, 241–246 (2013).

39. Kumar, A. et al. Computational SNP Analysis: Current Approaches and Future Prospects. *Cell Biochem. Biophys.* **68**, 233–239 (2014).

40. Zhang, M., Huang, C., Wang, Z., Lv, H. & Li, X. In silico analysis of non-synonymous single nucleotide polymorphisms (nsSNPs) in the human GJA3 gene associated with congenital cataract. *BMC Mol. Cell Biol.* **21**, 1–13 (2020).

41. Kumar, A. & Purohit, R. Use of long-term molecular dynamics simulation in predicting cancer associated SNPs. *PLoS Comput. Biol.* **10**, e1003318 (2014).

42. Kamaraj, R., Rajendran, V., Sethumadhavan, R., Kumar, C. V. & Purohit, R. Mutational analysis of FUS gene and its structural and functional role in amyotrophic lateral sclerosis 6. *J. Biomol. Struct. Dyn.* **33**, 834–844 (2015).

43. Kamaraj, B., Rajendran, V., Sethumadhavan, R. & Purohit, R. In-silico screening of cancer associated mutation on PLK1 protein and its structural consequences. *J. Mol. Biol.* **19**, 5587–5599 (2013).

44. Kamaraj, B. & Purohit, R. In silico screening and molecular dynamics simulation of disease-associated nsSNP in TYR1 gene and its structural consequences in OCA3. *Biomed. Res. Int.* **2013**, 1 (2013).

45. Tang, H. & Thomas, P. D. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *203*, 635–647 (2016).

46. Wong, L. P. et al. Deep whole-genome sequencing of 100 southeast Asian Malays. *Am. J. Hum. Genet.* **92**, 52–66 (2013).

47. Ashkenazy, H. et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* **44**, W344–W350 (2016).

48. Capriotti, E., Fariselli, P. & Casadio, R. A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics* **20**, 63–68 (2004).

49. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score 0.5? *Bioinformatics* **26**, 889–893 (2010).

50. González-Pérez, Á. & López-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score. *Condel. Am. J. Hum. Genet.* **88**, 440–449 (2011).

51. Cho, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7**, e46688 (2012).

52. Tang, H. & Thomas, P. D. PANTHER-PSEP: Predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics* **32**, 2230–2232 (2016).

53. UNIP: Technical notes: Calculating the human development indi. *Tech. notes* **37**, 14 (2016).

54. Calabrese, R., Capriotti, E., Fariselli, P., Pl, M. & Casadio, R. Protein Folding , Misfolding and Diseases : The I-Mutant Suite Supplementary informations. 9–10 (2008).

55. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13 Suppl 4**, (2012).

56. Singh, P. K., Mistry, K. N., Chiramaana, H., Rank, D. N. & Joshi, C. G. Association of damaging nsSNPs of XRCC1 with breast cancer. *Mol. Gene* **14**, 147–151 (2017).

57. Arshad, M., Bhatti, A. & John, P. Identification and in silico analysis of functional SNPs of human TAGAP protein: A comprehensive study. *PLoS ONE* **13**, 1–13 (2018).
58. Singh, S., Gupta, M., Sharma, A., Seem, R. K. & Changotra, H. The Nonsynonymous Polymorphisms Val276Met and Gly393Ser of E2F1 Gene are Strongly Associated with Lung, Head and Neck Cancers. Genet. Test. Mol. Biomarkers 22, 498–502 (2018).
59. Hassan, M. S., Shaalan, A. A., Dessouky, M. I., Abdelnaim, A. E. & ElHefnawi, M. Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity. Genomics 111, 869–882 (2019).
60. Pettersen, E. F. et al. UCSF Chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612 (2004).
61. Tanchuk, V. Y., Tanin, V. O., Vovk, A. I. & Poda, G. A New, Improved Hybrid Scoring Function for Molecular Docking and Scoring Based on AutoDock and AutoDock Vina. Chem. Biol. Drug Des. 87, 618–625 (2016).
62. Rudnicki, W. R., Mroczek, T. & Cudek, P. Amino acid properties conserved in molecular evolution. J. Mol. Biol. 340, 263–280 (2005).
63. Miller, M. P. Understanding human disease mutations through the use of interspecific genetic variation. Hum. Mol. Genet. 10, 2319–2328 (2001).
64. Deller, M. C., Kong, L. & Rupp, B. Protein stability: A crystallographer’s perspective. Acta Crystallogr. Sect. Struct. Biol. Commun. 72, 72–95 (2016).
65. Leidy. A Missense Mutation in CLIC2 Associated with Intellectual Disability is Predicted by In Silico Modeling to Affect Protein Structure and Functions.
66. Zhang, Y. & Skolnick, J. TM-align: A protein structure alignment algorithm based on the TM-score. Nat. Struct. Mol. Biol. 33, 2302–2309 (2005).
67. Doss, C. G. P. & Sethumadhavan, R. Investigation on the role of nsSNPs in HNPCC genes - A bioinformatics approach.
68. Peleg, O., Choi, J. & Shakhnovich, E. I. Evolution of Specificity in Protein-Protein Interactions.
69. Ko, J. M. I. N., Yang, J., Jeong, S. & Kim, H. Mutation spectrum of the TYR and SLC45A2 genes in patients with oculocutaneous albinism. 943–948 (2012) https://doi.org/10.3892/mrr.2012.764.
70. Piazzi, M., Bavelloni, A., Gallo, A., Faenza, I. & Blalock, W. L. Role of ribosomal protein mutations in tumor development (Review). Acta Biochimica Polonica 63, 1313–1324 (2016).
71. Klein, T., Eckhard, U., Dufour, A., Solis, N. & Overall, C. M. Proteolytic cleavage - mechanisms, function, and ‘omic’ approaches to determine protein function. Biochem. J. 405, 2156–2158 (2011).
72. Danelon, V. Modularity and Distinct Plexin-A4 / FARP2 / Rac1 Signaling Controls Dendrite Morphogenesis.
73. Jin, W. Regulation of Src family kinases during colorectal cancer development and its clinical implications. Cancer (Basel) 12, 1339–1420 (2020).
74. Parkin, A., Man, J., Tippon, P. & Pajic, M. Targeting the complexity of Src signalling in the tumour microenvironment of pancreatic cancer: from mechanism to therapy. FEBS J. 286, 3510–3539 (2019).
75. Jen, W. Regulation of Src family kinases during colorectal cancer development and its clinical implications. Cancer (Basel) 12, 1339 (2020).
76. Hoekstra, N., Kortum, L. & Schramm, S. C. The Ensembl Variant Effect Predictor.
77. Li, C. et al. Genome-wide Association Studies in Humans Links Nucleotide Metabolism to Leukocyte Telomere Length. Am. J. Hum. Genet. 106, 389–404 (2020).
78. Glessner, J. T. et al. Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circ. Res. 115, 884–896 (2014).
79. Sun, K., Wu, H., Wang, H. & Kung, M. H. TM-align: a protein structure alignment algorithm based on the TM-score. Acta Crystallogr. Sect. Struct. Biol. Commun. 63, 1339–1351 (2014).
80. Peleg, O., Choi, J. & Shakhnovich, E. I. Evolution of Specificity in Protein-Protein Interactions. Biophys. J. 107, 1686–1694 (2014).
81. Li, C. et al. A Computational Approach to Identify the Biophysical and Structural Aspects of Methylenetetrahydrofolate Dehydrogenase (MTHFR) Mutations (A222V, E429A, and R594Q) Leading to Schizophrenia. Adv. Bioinformatics 2017, 1–13. doi:10.1155/2017/2861061.
82. Chen, D. et al. Regulation of protein-ligand binding affinity by hydrogen bond pairing. Sci. Adv. 2, (2016).
83. Akter, S., Hossain, S., Hosen, M. I. & Shekhar, H. U. Comprehensive characterization of the coding and non-coding single nucleotide polymorphisms in the tumor protein p63 (TP63) gene using in. Sci. Rep. 63, 1–13 (2021).
84. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120 (2014).
85. Han, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Bioinformatics 00, 1–3 (2013).
86. McKenna, A. et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20, 1297–1303 (2010).
87. Fletcher, G. S. Clinical EPIDEMIOLOGY: The essentials. (Lippincott Williams & Wilkins, 2005).
88. Glantz, S. A. Primer of Biostatistics. (McGraw-Hill Inc., 1997).
89. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. J. Mol. Biol. 40, 442–451 (1975).
90. Danecik, P. et al. The variant call format and VCFtools. Bioinformatics 27, 2156–2158 (2011).
91. McLaren, W. et al. The Ensembl Variant Effect Predictor. Genome Biol. 17, 1–14 (2016).
104. Choi, Y. & Chan, A. P. PROVEAN web server: A tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* **31**, 2745–2747 (2015).
105. Pejaver, V. *et al.* The structural and functional signatures of proteins that undergo multiple events of post-translational modification. *Protein Sci.* **23**, 1077–1093 (2014).
106. Pejaver, V. *et al.* MutPred2: inferring the molecular and phenotypic impact of amino acid variants. 1–28 (2017) https://doi.org/10.1101/134981.
107. Capriotti, E., Fariselli, P. & Casadio, R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* **33**, 306–310 (2005).
108. Capriotti, E., Fariselli, P., Rossi, I. & Casadio, R. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9**, 1–20 (2008).
109. Jianyi, Y. *et al.* The I-TASSER Suite: Protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
110. Bowie, J., Luthy, R. & Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* *(80-.)*, **253**, 164–170 (1991).
111. Anderson, R. J., Weng, Z., Campbell, R. K. & Jiang, X. Main-chain conformational tendencies of amino acids. *Proteins Struct. Funct. Genet.* **60**, 679–689 (2005).
112. Li, B. *et al.* Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744–2750 (2009).
113. Venselaar, H., te Beek, T. A. H., Kuipers, R. K. P., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11**, (2010).
114. Venselaar, H., Ah, T., Kuipers, R. K. P., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11**, 548 (2010).
115. von Mering, C. *et al.* STRING 7 - Recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.* **35**, 358–362 (2007).
116. Szklarczyk, D. *et al.* STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
117. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
118. Lang, P. T. *et al.* DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA* **15**, 1219–1230 (2009).

**Author contributions**

N.A.K.: performed the bioinformatics analysis and prepare the draft. M.N.N.: verified the bioinformatics analysis. L.K.T.: verified the data presented and figures prepared, approved the final draft. F.Z.M.Y.: Proofread the draft paper. M.Z.S.: conceived the study, verified data presented, and secured grant for the study, approved the final draft.

**Funding**

The funding was provided by Ministry of Higher Education, Malaysia,[600-RMI/LRGS 5/3 (1/2011)-1].

**Competing interests**

The authors declare no competing interests.

**Additional information**

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-95618-y.

**Correspondence** and requests for materials should be addressed to M.Z.S.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021