Research Article

Association of Functional Variants and Protein-to-Protein Physical Interactions of Human MutY homolog linked with Familial Adenomatous Polyposis and colorectal cancer Syndrome

Zainularifeen Abduljaleel 1, 2, 3, 4

1 Department of Medical Genetics, Faculty of Medicine, Umm Al-Qura University, P.O.Box: 715, Makkah 21955, Saudi Arabia.

2 Molecular Diagnostics Unit, Department of Laboratory and Blood Bank, Al Noor Specialized Hospital P.O.Box: 6251, Makkah, Saudi Arabia.

3 Science and Technology Unit, Umm Al-Qura University, P.O. Box: 715, Makkah 21955, Saudi Arabia.

4 Molecular Diagnostics Unit, The Regional Lab and Blood Bank, Ministry of Health (MOH), Makkah, Saudi Arabia.

Corresponding Author

Dr.Zainularifeen Abduljaleel
Department of Medical Genetics, Faculty of Medicine, Umm Al-Qura University, P.O.Box: 715, Makkah 21955, Saudi Arabia.
Mobile: +966 570552361
E-mail: zainulbio@gmail.com
Abstract

The human gene MUTYH codes for a DNA glycosylase involved in the repair of oxidative DNA damage. Faulty MUTYH protein activity causes the accumulation of G→T transversions due to unrepaired 8-oxoG:A mismatches. MUTYH germ-line mutations in humans are linked with a recessive form of Familial Adenomatous Polyposis (FAP) and colorectal cancer predisposition. We studied the repair capacity of variants identified in MUTYH-associated polyposis (MAP) patients. MAP is inherited in an autosomal recessive type due to mutations in MUTYH (Y165C, G382D, P54S, A22V, Q63R, G45D, S136P and N43S), indicating that both copies of the gene become inactivated. However, the parents of an individual with an autosomal recessive condition may serve as carriers, each harboring one copy of the mutated gene without showing signs or symptoms of MAP. Six protein partners have been associated with MUTYH, four via direct physical interactions, namely, hMSH6, hPCNA, hRPA1, and hAPEX1. We examined, for the first time, specific interactions of these protein partners with MAP-associated MUTYH mutants using molecular dynamics simulations. The approach provided tools for exploration of the conformational energy landscape accessible to protein partners. The investigation also determined the impact before and after energy minimization of protein-protein interactions and binding affinities of MUTYH wild type and mutant forms, as well as the interactions with other proteins. Taken together, this study provided new insights into the role of MUTYH and its interacting proteins in MAP.

Keywords

MUTYH, Mutyh (human, murine genes coding for MUTYH glycolyase); MUTYH-associated polyposis; DNA mismatch repair; functional variants; protein-protein interactions; molecular dynamics simulation

Introduction

Familial Adenomatous Polyposis (FAP) alterations in the human MUTYH gene have been associated with autosomal recessive colorectal cancer development (Al-Tassan 2002; Out 2010; Farrington 2005). MUTYH is located on Chr1p32-34 and has 16 exons spanning 7,100 base pairs (Cheadle 2007). The phenotype of MUTYH-associated polyposis (MAP) is similar to attenuated FAP, having from 10 to a few hundred colorectal adenomas. It is estimated that ~1% of colorectal cancers are linked to MAP, but this number is likely to increase as more patients are carefully screened and test positive for MUTYH mutations (Out 2010; David 2007). The clinical course of colon cancer is usually silent and initially without symptoms until the disease has reached an advanced stage as shown in the figure 1.

Although no germ-line mutations were detected in the Adenomatous Polyposis Coli (APC) gene, tumor tissue from MAP patients had somatic APC mutations, mostly involving G:C to T:A transversions. KRAS mutations also have been detected in MAP patients (Lipton 2003).

The DNA glycosylase MUTYH plays a crucial role in base excision repair (BER) by preventing the accumulation of mutations resulting from oxidative DNA damage (Slupska 1999; Slupska 1996). MUTYH is an 8-oxoG:A mismatch repair enzyme in humans and is an ortholog of E.
coli MutY enzyme, but has less activity for other mismatches such as G:A or C:A. It is involved in base excisions of the A/G, A/ (Gene Ontology) GO and A/C type of non-pairings (McGoldrick 1995). MUTYH identifies and removes the oxidized purine 2-hydroxyadenine (2-OH-A) from different base pairs, particularly from 2-OH-A:G, indicating that the protein might counter the biological consequences of other oxidized premutagenic DNA bases (Ohtsubo 2000; Hashiguchi 2002). The DNA glycosylase starts BER by removing 8-oxoG from 8-oxoG:C base pairs. MUTYH also removes adenine inserted opposite 8-oxoG during replication (Parker 2003). Subsequent to adenine elimination, apurinic-apyrimidinic endonuclease (APEX) cuts the apurinic site (AP), and the gap filling steps are then accomplished via the long patch BER pathway (Yang 2001; Parlanti 2004). Among DNA glycosylases, exclusion of the base opposite the lesion seems to be a distinctive characteristic of MUTYH and entails both the replication machinery and post-replicative mismatch repair (MMR).

In humans, germ-line biallelic mutations of the MUTYH gene are linked with a recessive form of FAP called MAP. MAP patients exhibit symptoms of colorectal cancer early in adult life, with widespread adenomatous polyps of the colon. Due to the failure of repair of 8-oxoG:A mismatches, tumors of MAP patients accumulate somatic G:C→T:A transversions in cancer-related genes (Lipton 2003). In rodent cells, the inactivation of Mutyh is linked with high mutation rates, frequently because of the increase of G:C→T:A transversions. Many mutations in the MUTYH gene, including missense and insertion/deletion alterations, have been observed in MAP patients. The Y165C and G382D substitutions are the most common variants in the European population, with a frequency of 1%, and have been observed in 80% of patients with mutations in both alleles of the MUTYH gene (Sieber 2003; Gismondi 2004). Therefore, screening for MUTYH mutations might be important in identifying a subset of patients with early onset MAP. Therefore, further large-scale analysis of multiple genes expression datasets might lead to the identification of more representative gene expression signatures associated with CRC predisposition. Herein, we integrated one or more independent CRC gene expression datasets retrospectively, which led to the identification of a five genes that are associated signature associated with colorectal cancer systemic deterioration.

This paper deals with the identification of a MUTYH gene functional network and its relationships with others genes based on biomolecular interactions. The susceptible loci were interrogated for cross-interactions and ranked, using the Networked Gene Prioritize (NGP) method, to detect FAP-associated genes. Furthermore, we studied the MUTYH protein structural configuration and its mutation transition, via molecular dynamic (MD) simulation, and examined protein-protein interaction (PPI) formation of specific complexes, including structural predictions of such complexes. To the best of our knowledge, this is the first report addressing MD simulation of the MUTYH protein domains (I and II). This study focused on chain A and investigated MUTYH DNA binding and protein interactions. Furthermore, comparative MD analyses of the MUTYH protein with other proteins were performed in order to clarify physical interactions between the wild type and mutants.

Results

Deleterious point mutations of MUTYH in human FAP form 8-oxoG:C pairs, which initiate the binding of 8-oxoguanine DNA glycosylase (OGG1) to regenerate G:C pairs via the BER pathway,
with the involvement of PCNA, APE1, MSH6 and RPA1. Both OGG1 and MUTYH proteins have central roles in preventing the build-up of 8-oxodG damage and tumor formation (Fig. 2A-D). The MUTYH gene is associated with various diseases, including polyposis (Fig. 3A). It has been reported that MUTYH interacts with PCNA, APEX1, and RPA1, including the MSH6 subunit of MutSα MMR detection complex proteins (Fig. 3B) (Parker 2001). We examined, for the first time, protein-protein interactions, conformational changes, and binding affinities of wild type and novel mutant MUTYH variants with known partners involved in BER.

Validation of Mutyh associated polyposis associated gene panel in the TCGA CRC dataset

We successively focused on the potential role of the regulated genes that are associated in the CRC cancer panel. Therefore, targeted in a five genes i.e. MUTYH, PCNA, RPA1, APEX1, MSH6 which are associated with polyposis was validated using the TCGA CRC dataset to determine between their relationship to overall survival and disease-free survival (DFS) (Fig. 4A) data from the eBioPortal for Cancer Genomics. The OncoPrint for this gene panel in the TCGA CRC dataset with the proportion of patients overexpressing each gene is presented in (Fig. 4B-E). Interestingly, the combination of this five-gene revealed a higher prognostic value, in which patients overexpressing of the five genes showed month survival on both primary and metastatic (log-rank test P-value: 0.15e-9), living and decreases survival (log-rank test P-value: 0.00) and the affected individual both male and female survival ratio shown (log-rank test P-value: 0.843) than those with lower and over expression of these genes. Data from the univariate analysis were subsequently put into the Cox proportional hazards multiple regression models to identify the independent factors for prognosis.

Impact of nsSNP Variants on Protein Structure, Conformational Flexibility, and Stability

The mutations nsSNPs in the MUTYH gene were identified and their impacts on the substituted residues were determined. A total of eight nsSNP mutations were detected and checked for their impact on MUTYH protein stability (Table 1). Positions to specific independent count (PSIC) scores for site substitution were as follows: 1.250 (A22V; benign, i.e., lacking phenotypic effect), 0.152 (Q63R, benign); 2.227 (G45D, probably damaged); 2.604 (P54S, benign); 0.351 (N43S, benign); 0.070 (P54S, benign); 2.884 (Y165C, probably damaged) and 2.004 (G382D, probably damaged), with normal accessibility score 0.85. Sorting Intolerant from Tolerant (SIFT) identified amino acid substitution influencing protein function and phenotypic outcome. The impact of point mutation on protein stability was ascertained using the Panther method for selected MUTYH nsSNPs within the exonic region (Table 2). This method was based on algorithms to locate homologous sequences using SWISS-PORT version 51.3 and TriEMBL 34.3. The substitution at position 326 from Ser-Arg was predicted to be tolerated (low confidence) with a score of 0.18 and median sequence conversion of 3.23. Threshold for intolerance was <0.05, damaging was <0.001 (Fig. 5). Among 152 SNPs identified in MUTYH, 95 were nsSNPs. Furthermore, a total of 90 samples obtained from the genome variant server (GVS) identified MUTYH with 32 SNPs; however, only five were nsSNPs (Fig. 6A-B). The functionally significant nsSNPs were predicted by SIFT (35.8%) and PolyPhen (37.7%). SNPs&GO further confirmed these results, with high correlations of 0.82 and 0.81, respectively.

The protein stability change was expressed as a linear combination of energy functions, where the proportionality coefficients varied with the solvent accessibility of the mutated residue and were identified with the neural network. The sequence optimality score (sum of negative ΔΔGs) was
computed for each substituted position in the protein. The optimality calculated a stability score, analogous to the free energy difference between a wild type and mutant protein. The stability score was used for predicting whether or not a mutation had an impact on the protein structure, and consequently a potential role in disease. The torsion angles were initialized with constant values of Boltzmann energy (Table 3).

The first identified nsSNP mutation (P54S) substituted proline in the wild type (Fig. 7A), and affected backbone conformation required for correct enzymatic activity. The wild-type Pro residue was highly conserved, unlike other residues in this region. The observed mutant residue was not among others reported before in homologous proteins. The mutation results in the formation of an 'empty space' in the core of the protein (or protein-complex) and causes the loss of hydrophobic interactions in the core of the protein (or protein-complex). The protein stability due to this mutation differed in energy value by 1.47 kcal/mol, solvent accessibility (Acc) 0.00%, and torsion angles ($\phi$, $\psi$) -69.8$^0$, 144.1$^0$; consequently the prediction was for destabilization.

The second nsSNP mutation A22V (Fig. 7B) was buried in the core of the protein (or protein-complex), and the mutant residue likely exhibited steric hindrance. The change in energy value 0.03 kcal/mol, solvent accessibility (Acc) 0.00% and torsion angles ($\phi$, $\psi$) -1118.0$^0$, 105.1$^0$, resulted in predicted protein destabilization.

The third nsSNP mutation Q63R (Fig. 7C) differed from the wild type based on the energy value 0.15 kcal/mol, solvent accessibility (Acc) 69.36% and torsion angles ($\phi$, $\psi$) -92.5$^0$, -49.5$^0$; these parameters predicted protein destabilization.

The fourth nsSNP mutation G45D (Fig. 7D) was located close to a highly conserved region and resulted in charge differences between the wild-type and the mutant proteins. The stability was altered due to a change in energy value 0.69 kcal/mol, solvent accessibility (Acc) 66.97% and torsion angles ($\phi$, $\psi$) 115.9$^0$, -145.4$^0$; hence, the prediction of destabilizing the protein.

The fifth nsSNP mutation S136P (Fig. 7E) substituted a residue on the protein surface, possibly affecting interactions with other molecules or with other surface residues. The torsion angles for this residue were unusual, indicating that only glycine was flexible enough to make the correct orientation, and mutation caused the local backbone to form incorrect conformations that disturbed local structure. The energy value 0.39 kcal/mol, solvent accessibility (Acc) 7.70%, and torsion angles ($\phi$, $\psi$) -68.3$^0$, 140.1$^0$ predicted protein destabilization.

The sixth nsSNP mutation N43S (Fig. 7F) was within a domain annotated in Uniprot as "Nudix hydrolase". The wild-type residue was buried in the core of the protein (or protein-complex), and the stability of this mutation differed in energy value by 0.08 kcal/mol, solvent accessibility (Acc) 47.73%, and torsion angles ($\phi$, $\psi$) -43.8$^0$, 153.3$^0$; thus, the prediction of stability change was neutral.

Two additional mutations (Y165S, G382D) were obtained from Uniprot-entry MUTYH (Q9UEF7). MUTYH domain II structure predictions from I-TASSER showed highly conserved regions for PDB file 3N5N. The Y165S nsSNP mutation caused alteration in the corresponding structural domains (Fig. 7G). Mutation of this fully conserved residue was damaging for the protein. The mutation differed in energy value 0.70 kcal/mol, solvent accessibility (Acc) 45.61%, and torsion angles ($\phi$, $\psi$) -60.3$^0$, 141.1$^0$, causing the protein to destabilize. The nsSNP G382D mutation (Fig. 7H)
affected protein stability, with the difference in energy value from the wild type being 0.64 kcal/mol, solvent accessibility (Acc) 57.89%, and torsion angles ($\phi$, $\psi$) -92.5$^\circ$, -149.3$^\circ$, the outcome being destabilizing.

Collectively, the data suggested that the stability, flexibility, and functional properties of the mutants differed and affected the protein stability of the wild type. Detailed analysis of contact energies before and after energy minimization located functionally or structurally important residues, and helped to classify consequences of each mutation, as well as identifying novel residues for site-directed mutagenesis. The difference in free contact energies (kcal/mol) for denaturation of MUTYH protein domain I and II between wild-type and mutant proteins was also determined using Chemical Computing Group (CCG) Molecular Operating Environment (MOE) (http://www.chemcomp.com) before and after energy minimization (Fig. 7I).

Analysis of Protein-Protein Interaction Networks

The disease-associated genes were obtained from the Online Mendelian Inheritance in Man (OMIM) database based on the Mayo Clinic clinical data set. Colorectal cancer gene networks were identified for this complex disease. Artificial susceptibility loci, each comprising 100 genes, were constructed around 10 putative colorectal cancer genes described in OMIM. We assessed the performance of our classifier on the basis of these various data sources in three different gene networks generated on the basis of a Bayesian framework. One network was generated solely on the basis of GO data (GO network), one network was based on both Microarray Co-expression and predicted PPI data (MA and PPI network), and an overall network was constructed containing all three types of data (GO, MA and PPI network). The microarray co-expression and predicted PPI interaction (MA & PPI) network selected four genes (PCNA, MSH6, APEX1, and RPA1), whereas the GO network identified three genes (OGG1, MSH2, and MUTYH), and the GO, MA and PPI network was comprised of PCNA, MSH6, APEX1, and RPA1 (Fig. 8A-D). However, in the GO, MA and PPI network, among the top five disease genes showing strong interactions with MUTYH, only MSH6 and PCNA were identified.

Molecular Aspects of Protein-Protein Interactions

The physicochemical properties of interaction interfaces were assessed by incorporating a naive Bayes classifier, a simple probabilistic classifier that relies on Bayes' theorem application with strong (naive) independence prediction, into a computational scheme to integrate diverse information. This approach took into consideration dynamic hydrogen bonding effects between the bio-molecules, intermolecular H-bond, H-bond energy, and the specific solvation patterns of water molecules. Electronic polarization and charge-transfer as well as chemical bond formation are also taken into consideration. MUTYH protein domain I showed interactions with MSH6, PCNA, APEX1, and RPA1 proteins. Wild type MUTYH (control) and mutant (affected) proteins were analyzed separately based on interactive docking studies (Fig. 9 and 10). Protein interactions between the wild type proteins were normal based on the calculations of intermolecular H-bond, or H-bond and binding energies (Table 4.) Interactions also were determined for MUTYH with MSH6, PCNA, APEX1, and RPA1. In the wild type, MUTYH protein domain I interacted with MSH6 (Fig. 9A), and binding energy calculations for MSH6 (residues Ser196, Gly159, His113, Glu198) with MUTYH (residues Gly148, Ser150, Ser200, Pro152, Pro199) were -66.75 kcal/mol, whereas the intermolecular energy of H-bond formation was -7.45 kcal/mol. With wild type PCNA (Fig. 9B), the binding energy between residues
Gly111, Leu205, Thr206 and MUTYH (residues Pro29, Ile255, Glu110) was -59.02 kcal/mol. The intermolecular H-bond energy (between three residues in PCNA and MUTYH (Ser154, Ser153) was -2.36 kcal/mol. The MUTYH wild type interacted with APEX1 wild type (Fig. 9C), exhibiting binding energy between APEX1 (Gln71, Arg75, Lys103, Tyr118, Leu104, Trp119) with MUTYH (Leu104, Asn102, Glu72) calculated as -33.85 kcal/mol. The intermolecular H-bond energy between the five residues, two in APEX1 (Lys103, Val142) and three in MUTYH (Glu59, Lys103, Tyr117) was -3.14 kcal/mol. The MUTYH wild type interacted with RPA1 wild type (Fig. 9D), showing binding energy between RPA1 (Val76, Ala31, Met1) and MUTYH (Leu38, Leu102, Glu35, Glu129) of -93.60 kcal/mol. The intermolecular H-bond energy between two residues in RPA1 (Pro23, Ser150) and two in MUTYH (Tyr146, Ala31) was calculated to be -4.52 kcal/mol.

We further studied interactions of mutated MUTYH with the four wild type proteins. The mutated MUTYH protein interaction with wild type MSH6 protein had an altered physical interaction (Fig. 10A), which was reflected from the binding energy calculation between MSH6 (Arg140, Gln127, Gly111) and MUTYH (Thr137, Gly148), being less (-41.43 kcal/mol) than the MUTYH (control) and MSH6 (control) interaction. Similarly, the intermolecular H-bond energy was -7.92 kcal/mol between six residues of MSH6 (Arg86, Asp197, Gly149, Glu128, Leu76, Leu72) and six residues of MUTYH (Arg121, Tyr109, Thr137, Glu198, Ser146, Ser143). The intermolecular H-bond energy between two residues of APEX1 (Lys103, Val142) with two of MUTYH (Glu59, Thr117) was -4.94 kcal/mol. The mutant MUTYH interaction with wild type APEX1 (Fig. 10C) showed binding energy of -57.87 kcal/mol between three APEX1 residues (Gln117, Gln71, Tyr118) and three in MUTYH (Lys103, Leu104, Gly111). The intermolecular H-bond energy between the two residues of APEX1 (Lys103, Val142) with two of MUTYH (Glu59, Thr117) was -4.94 kcal/mol. Finally, the mutant MUTYH interaction with RPA1 wild type (Fig. 10D) showed altered physical interaction based on the binding energy between RPA1 (Leu102, Val76, Met1) and MUTYH (Leu38, Glu129, Glu35), calculated as -92.83 kcal/mol. The intermolecular energy of H-bond formation between three residues of RPA1 (Glu100, Pro23, Met1) and MUTYH (Thr116, Ala31, Ser150) was -5.71 kcal/mol.

**MD Simulation and Solvent Accessibility of Amino Acid Residues in Mutant Proteins**

Deleterious nsSNPs were searched from a database of single nucleotide polymorphisms (dbSNP) (http://www.ncbi.nlm.nih.gov/SNP), and from the single amino acid polymorphism database (SAAPdb) (http://www.bioinf.org.uk/saap/db/) (Hurst 2009), which had data on SNPs from both dbSNP and Human Genome Variation database (HGVBASE). This approach displays data onto the translated regions of the gene to establish whether the mutation occurred in a region of interest and whether it resulted in a potential missense mutation. We selectively probed two domains from the MUTYH gene for MD simulations. First, PDB: 1X51 structure of the NUDIX domain from human A/G-specific adenine DNA glycosylase alpha-3 splice isoform, in the experimental structure methods from NMR work (Fig. 11A-B), was examined. This region has the UniProt identifier Q9UIF7. The domain I mutations affected chain-A at residues P54S, A22V, Q63R, G45D, S136P and N43S. Second, domain II of MUTYH protein was predicted by homology modeling from I-Tasser, a program that employs best fit according to the C-score showing homology with Q9UIF7 (MUTYH) protein based on TM-score 0.47±0.15, RMSD 12.4±4.3Å and C-score -2.03. The predicted protein was highly conserved with PDB: 3NSN. The crystal structure analysis of the catalytic domain and
interconnector of human MUTYH in the experimental structure methods from X-ray 2.30Å (Fig. 11E and G) revealed this domain I model to have two mutations. These mutations changed the sequence in the glycosylase in a manner that was common in people of European descent. One mutation changed the amino acid Y165C, whereas the second mutation switched the amino acid G382D. The mutations from both MUTYH domains (I and II) positions were examined using SWISS-PORT and viewed separately to obtain altered model structures.

The energy minimizations were achieved by NOMAD-Ref Gromacs server, using force field energy for the native type protein in both domain I and II and mutant type protein structures. The total energy of native structures of domains I and II, and mutant structures, were calculated (Table 3), and ranged from 0.25 to 0.50 RMSD native and mutant protein structure, respectively. Using the CHARMM-GUI server that offers graphical user interface (GUI) for MD simulations, the consequence of mutations was studied using classical molecular dynamics approaches. We analyzed both the native and mutated structures through long simulations in defined solvent conditions, as well as investigating the differences in dynamics and stability in MUTYH domains (I & II).

The simulation created an aqueous solvent environment based on fixed parameters around the chain ‘A’ protein. Both domains of MutY homolog protein 1X51 were solvated showing system with octahedral boundary shapes with 5400 water molecules of 10Å cutoffs to fully solvate the molecule with edge distance 10.0 (Fig. 11C; 11D 11F). MD simulation procedures were employed to examine the dynamic performance of the two MUTYH domains I, and II. Trajectories over 10ns were based on analyzing the RMSD for both domains at 300K. RMSD with 0.10Å at the start of the trajectory (t=0) show the movements of both MUTYH domains took place throughout the thermalization and equilibration phases. The deviation value quickly increased at 300K during the initial 2ns and sustained until the end of simulation (t=10 ns). The MUTYH domains (I and II) mutations exhibited quick elevation in deviation value (1.10 Å) until 10ns and then reduced by the end of simulation.

The results indicated that the MUTYH domains were not stable and did not reach in folded conformation till finish of the simulation at 70 °C. However, it was possible that the Mutated MUTYH domains structure may reach to stable and folded form at lengthier simulation times than the applied 10ns. NOMAD-Ref server performed energy minimizations studies for both wild and mutant structures. The total energy for the native protein structure domain I (1X51) following before energy minimization was 4719.71985 kJ/mol (score; 0.69), whereas prior to after energy minimization it was 3676.5378 kJ/mol (score; −3.31). For the mutant protein domain II form(s) the corresponding before energy minimization was 390 841.524 kJ/mol (score; 0.50), whereas prior to after energy minimization it was 2453.2390 kJ/mol (score; −1.25).

Discussion

In the current study, we retrospectively derived and validated a gene expression signature associated with the risk of systemic relapse in patients with colorectal cancer. Analysis of the data from Colorectal Adenocarcinoma dataset (http://www.cbioportal.org). In the datasets, we identified upregulated and downregulated genes associated with in CRC. Interestingly, several of the identified genes (PCNA, RPA1, APEX1, MSH6, and Mutyh) were identified to be differentially expressed mRNA expression profiling of CRC compared to adjacent normal samples. Higher expression of gene
MSH6 and MutyH was found to be associated with metastasis in different types of human cancers [17]. Our analysis narrowed down the CRC recurrence signature to these five genes (PCNA, RPA1, APEX1, MSH6, and Mutyh) whose expression was associated with OS (log-rank test P-value: 0.15e-9) and the indidual average gender overall survey (log-rank test P-value: 0.843).

MUTYH is involved in the initial steps of MMR and BER pathways, providing key contributions for genomic stability and maintenance. To the best of our knowledge, there has been no report that deals with in-silico modeling and functional/structural aspects of MUTYH variants. The MUTYH complete crystal structure has not been solved. However, the observed functional and interactive activities of MUTYH variants in this study can be related to the amino acid sequence information and structure of E. coli MutyH protein, which is 41% identical with human MUTYH (Wooden 2004; Bai H 2007). MUTYH has a catalytic site resembling the NUDIX domain from human A/G-specific adenine DNA glycosylase alpha-3 splice isoform, determined from NMR (Made 2002). All mutants studied here (Fig. 5A-I) were predicted to be within this catalytic domain.

While the C-terminal domain of E. coli MutyH protein is not needed for its adenine removal activity, it strongly interacts with the Gene Ontology (GO) containing strand to flip the target adenine out of the helix. In support of this theory, it has been shown previously that the removal of this domain from MUTYH protein drastically reduces its adenine incision activity from AGO pair, but not from A-G mismatch (Noll DM 1999). Indeed, it was previously shown that the C-terminal domain of truncated MutY is not only defective in removing “A” but also in binding substrate containing AGO (Noll DM 1999). Examples of variants in which the consequence of MUTYH function may be predicted by their corresponding position in the structure are the P18L, V22M and G25D (Shinnura K 2001, Al-Tassan N 2002). These three SNPs are present in the RPA binding site in the N-terminus of MUTYH, and interfere with the localization of MUTYH to the site of DNA replication. Variants E466X (Jones S 2002) and Y90X (Sampson JR 2003), which were found in individuals of Indian and Pakistani descent, respectively, represent truncated MUTYH protein due to a premature stop codon.

MUTYH domain I mutations probably have clinical significance with regard to cancer risk. Mutation P54S (Coil) showed solvent accessibility 0% (i.e. buried), whereas G45D (bend) showed 70% solvent accessibility, indicating that it is exposed on the outer surface; both alterations were designated as ‘probable-pathogenic’ for gastric cancer. Mutations A22V (beta-ladder, solvent accessibility 60%) and S136P (coil, solvent accessibility 8%, buried) both had predicted phenotypic consequences for colorectal adenomatous polyposis. Variants Q63R (Helix) and N43S (Coil) were predicted to be neutral. Mutation G45D showed 50% solvent accessibility but was buried, and clinical data indicated an associated with high colon cancer risk, as predicted in the modeling studies (“functionally pathogenic”).

Domain II also had common mutations based on the clinical report, supporting a role in MUTYH-associated human carcinogenesis (Wooden 2004). Mutation Y165C (Loop) with 29% solvent accessibility was considered as intermediate, whereas G382D (Bend) with 29% solvent accessibility showed probable clinical relevance for increased cancer risk. Implications for protein structure and stability were calculated through atom and torsion angle potentials. The N43S variant was calculated as -0.08 kcal/mol (DDGp) and predicted to be neutral, whereas all other variants were destabilizing.
Modeling of protein-protein interactions also can provide important insights of wild type versus mutant MUTYH function (Marcotte 1999), via predicted binding sites, protein networks, multi-molecular assemblies, mapping of metabolic pathways, and eventual drug design (Eisenberg 2000). Models supported four candidate proteins (MSH6, PCNA, APEX1, and hRPA1) interacting with mutated MUTYH. MUTYH co-localizes with PCNA to replication foci (Boldoghi I 2001; Parker A 2001, Clark A. B 2000; Kleczkowska H. E 2001). In addition, MUTYH interacts with RPA (Parker 2001). The interactions of PCNA with MUTYH and mismatch repair enzymes suggest that PCNA may act as a coordinator of both repair pathways. We hypothesize that proteins involved in DNA replication, mismatch repair, and base excision repair may exist as a multiple-protein complex, and that MUTYH may be orientated in the replication fork to recognize 8-oxoG on the parental strands and to excise misincorporated A on the daughter strand.

It has been suggested that PCNA may act as a molecular adaptor, coordinating and regulating the actions of DNA replication, DNA repair, and cell cycle control (Matsumoto 1999; Warbrick E 1998). Thus, we further analyzed mutated MUTYH variants for functional networks and relationships based on data from the Bio-Molecular Interaction Network Database (BIND) (Alfarano 2005), the Human Protein Reference Database (HPRD) (Peri 2004), Reactome (Joshi 2005), and Kyoto Encyclopedia of Genes & Genomes (KEGG) (Kanehisa 2004). This approach included analysis of susceptibility loci and genes from different loci that might be linked directly (Turner 2003) or indirectly (Ma 2003). These methods are useful to rank candidate genes for each locus, and significantly increase the chance of detecting disease-related genes. Since MUTYH interacts with PCNA (Nakabeppu 2001; Joshi 2005; Brunner 2004) and DNA, it is possible that an indirect association between MUTYH and hMUTSα may occur. Direct physical interaction of these two proteins was determined from affinity-binding experiments with GST-tagged MUTYH and highly purified recombinant hMUTSα and hMUTS beta proteins. MUTYH recognizes the interface between MSH2 and MSH6 heterodimer, rather than the peptide motifs of MSH6. Western blotting suggested that MUTYH interacted with MSH6 but not with MSH2. This result is consistent with studies showing no interactions of MUTYH and MSH2 in HCT15 cell extracts. Although both MUTYH and MSH6 can bind to their DNA substrates, the interaction between both proteins can occur in the absence of DNA. It is interesting to note that the interaction of MUTYH with MSH6 wa substantially weaker in MT1 cells that express wild type MSH2 and missense mutant MSH6, as compared with the parental TK6 cells. The mutations are at the C-terminal region of MSH6 downstream of the ATP binding site. The result suggests that the C-terminal region of MSH6 is important for its interaction with MUTYH. Direct contact of MUTYH with the C-terminal region of MSH6 remains to be determined (Kat A 1993), although MUTSα can still bind mismatched DNA (Hampson 1997). It has been suggested that the mutations in MSH6 in MT1 cells may affect its ATPase activity (Iaccarino I 1998).

**In-silico** MD simulation noted mutations of MUTYH affecting its associations with binding partners, due to binding site variations in hydrophobic versus electrostatic interactions. Generally, the packed protein cores and the ‘hot spot’ regions had residues that contributed significantly to protein stability. The hot spots were also highly conserved by evolution at the protein-binding site (Ma 2003). It has been suggested that a hot spot region may provide a suitable target for drug design (Ma 2005). We sought to explore the structural stability and associated thermodynamics in a realistic environment. Nonpolar effects remained consistently more favorable in wild and mutants,
emphasizing the significance of hydrophobic effects in protein-protein binding. While entropy analytically opposed binding in every case, there was no observed trend in the entropy difference between wild and mutant MUTYH forms. Free energy decomposition showed residues were placed at the interface, making them favorable.

Finally, it was found that the monomer MUTYH model was stable in vitro, and such a model is suitable for protein stability analysis, although the monomer MUTYH binding to DNA binding and to other proteins was weaker. Our results showed that the MUTYH mutations examined caused decomposition of binding free energies. Furthermore, structural analyses indicated a distorted geometry of the binding site, and hence weaker interactions with S136P, G45D, N43S (L37V), A22V, Q63R and P54S.

**Materials and Methods**

**Clinical Dataset**

We retrieved clinical records of MutYH for multiple adenomas, test unit Code 84304 (http://www.mayomedicallaboratories.com) from the Mayo Medical Laboratory, USA. The PCR-based assay was performed to check the presence of NM_001048171.1: c.494A>G p.(Y165C): rs34612342 and NM_001048171.1:c.1145G>A, p.(G382D): rs36053993 mutations in MutYH in multiple adenomas. Other associated mutations in the MutYH gene (NM_001128425.1: c.1213C>T, p.(Pro405Ser): rs121908382 (chain-A, Pos54); NM_001128425.1: c.1118C>T, p.(Ala373Val): rs35532891; (chain-A, Pos22); NM_001128425.1: c.1240C>T, p.(Gln414Arg): rs760420977; (chain-A, Pos63); NM_001128425.1: c.1187G>A, p.(Gly396Asp): rs36053993; (chain-A, Pos45); NM_001128425.1: c.1460_1461dup, p.(S488P): rs587776618; (chain-A, Pos36), and NM_001128425.1: c.1162C>G, p.(Leu388Va): rs587783057; (chain-A, Pos37) were retrieved from Protein Data Bank (PDB) ID: 1X51 (http://www.ebi.ac.uk/pdbe/entry/pdb/1x51/analysis) resulting in heritable susceptibility to colon and stomach cancer syndrome of FAP.

**Patient information and data analysis for Colorectal Cancer**

The current study was conducted on six different colorectal cancer associates: (1) Colorectal Adenocarcinoma dataset (Giannakis M et al., 2016), which included 619 patients' samples that who developed 239 patients are Male and 380 patients are female. (2) Colorectal Adenocarcinoma dataset (Seshagiri S et al., 2012), which included 276 patients that who developed 117 patients are Male and 154 patients are female. (3) Colorectal Adenocarcinoma (Hoadley KA et al., 2018) dataset, which included a total of 594 CRC patients who developed 312 patients are Male and 280 patients are female. Interrogation of the TCGA dataset was conducted as previously described (Hoadley KA et al., 2018). (4) Colorectal Adenocarcinoma (TCGA, Provisional 2014) dataset, which included a total of 640 CRC patients who developed 335 patients are Male and 294 patients are female. (5) Colorectal Adenocarcinoma Triplets (Brannon AR et al., 2014) dataset, which included a total of 138 CRC patients who developed 59 patients are Male and 79 patients are female. (6) Metastatic Colorectal cancer (Yaeger R et al., 2018) dataset, which included a total of 1134 CRC patients who developed 597 patients are Male and 502 patients are female. The relationship of gene expression patterns with patient survival in the TCGA database was queried using the cBioportal database with the formula GENE: EXP > 0, where the gene represents a query of our associated genes are linked with polyposis.
syndrome. In addition, Fisher’s exact test and the Mann-Whitney (Anna Hart 2011) test implied used to investigate the categorical and constant variables. We determined and compared survival curves using the Kaplan-Meier (Kaplan, E. L; Meier, P 1958) method and log-rank tests (Gehan A 1965). Cox proportional hazards model was used to analyze associations between Clinic-Pathological symptoms and patient survival. Overall survival (OS) data was taken from the chiportal (http://www.chiportal.org).

Predictions of Deleterious nsSNPs of Mutyh Gene and their functional impacts

Non-synonymous single nucleotide polymorphisms (nsSNPs) represent common genetic variation that can encode changed amino acid residues in proteins. nsSNPs may impact the structure or function of expressed proteins, and therefore have an effect on disease outcome. In order to evaluate the phenotypic effects of nsSNPs in human DNA repair genes, we considered every polymorphism in terms of various functional properties. The functional properties were computed based on amino acid prediction from programs such as Sorting Intolerant From Tolerant (SIFT) (Ng 2001). We provided a comprehensive, updated list of all validated nsSNPs from dbSNP, the public database of human single nucleotide polymorphisms at the National Center for Biotechnology Information (NCBI); however, we selected for detailed study MutY homolog (E. coli) nsSNPs within the exonic region, and Genome Variation Server (GVS) (http://gvs.gs.washington.edu), situated in the human DNA repair gene list. The list includes repair enzymes and genes associated with the response to DNA damage, as well as those implicated with genetic instability or sensitivity to DNA damaging agents. PolyPhen version 2.0.9 (Ramensky 2002) was used to predict the outcome of an amino acid change on the structure and function of a protein based on specific empirical rules and the encoded sequence. The protein sequence of MUTYH was submitted to PolyPhen2 under ID Q9UIF7, the program calculated PSIC scores (Position to Specific Independent Count) for each variant and estimates the difference between the variant scores, where a difference of >2.884 was considered as detrimental.

The Impact of MUTYH Associated Mutations on Protein Structure and Stability

Protein structure effects were carried out from Project HOPE (Have yOur Protein Explanation) developed at the Centre for Molecular and Bio-molecular Informatics (CMBI) Department of Bioinformatics at the Radboud (Venselaar 2010). The molecular origin of disease-related phenotypes caused by mutations in MUTYH was analyzed via data processing, predicting the impact of each specific mutation on the 3D structure and function of the protein. The UniProt database (http://www.uniprot.org/) web server was used for the retrieval of features that can be mapped on the modified sequence (Jain 2009) Homology Derived Secondary Structure of Proteins (HSSP) conservation scores and sequence-based predictions by DAS-servers were used for biological sequence annotation (Prlic 2007). The protein stability prediction were carried out based on the torsion angle (Cologne University Protein Stability Analysis Tool) (CUPSAT) and stepwise multiple regressions were then used to unify the atom and torsion angle potentials to construct the prediction model on both MUTYH domain I and II. Energy functions were predominantly derived from mean force potentials based on the Boltzmann principal. The torsion angle potentials were used in non-redundant structures of MUTYH proteins to derive the torsion angle φ and ψ after running Dictionary of Secondary Structure of Proteins (DSSP) for the entire dataset.
Pathogenic Prediction of Functional Variants Associated with Disease

Several computational methods have been developed for the classification of SNPs according to their predicted pathogenicity; however, in this study we used SNPs&GO and Panther, (Thusberg 2011; Thomas 2003). SNPs&GO program based on Support Vector Machine (SVM) which classify mutation type based on sequence information from Multiple Sequence Alignments (MSAs) with function-based log-odds scores about protein functions describe by Gene Ontology (GO) (Ashburner 2000). Input was based on the selected position of each mutation, for both wild and mutant type residues of MUTYH protein structures. Binary predictions (pathogenic/neutral) were taken into consideration with their confidence values. The effects of a missense variant was categorized into three types; “Probably pathogenic”, “Possibly pathogenic” and “Benign.” We converted these into binary classifications in two ways, first by considering only the “Probably pathogenic” class as pathogenic and the “Possibly pathogenic” and “Benign” classes as neutral, and second, by considering both the “Probably pathogenic” and “Possibly pathogenic” classes as pathogenic, and the “Benign” class as neutral.

Prioritization of Disease Candidate Gene and Networks

The Bayesian approach was used to generate a gene network, based upon data obtained from Gene Ontology (GO), (Kyoto Encyclopedia of Genes and Genomes) (KEGG), Bio-molecular Interaction Network Database (BIND), Human Protein Reference Database (HPRD), and Reactome. These datasets were comprised of nearly 70,000 predicted protein-protein interactions (Lehner 2004; Stelzl 2005). The genes received a primary assigned score of three different susceptibility loci, each containing a disease gene (P, Q or R) and a couple of non-disease genes. At each locus the 3 positional candidate genes increase the scores of genes functionally proximate within the gene network, using a kernel function which models the relationship between gene-gene distance and score effect. When all loci were processed, shuffling the three susceptibility loci 10,000 times across the genome allows resolving of an empirical p-value of each gene, and the eventual ranking of the positional candidate genes per locus. Genes P, Q, and R should then end up as the top ranked genes, according to the significant p-values. The disease genes were ranked within the top 10 per artificial linkage region, when each region contained 100 genes. Based on this method we were able to identify the genes (Turner 2003).

Macromolecular of Protein-Protein Interaction

Using Global Range Molecular Matching (GRAMM-X) Docking Server v.1.2.0 (Tovchigrechko 2006) carried out protein-protein interactions. GRAMM was compiled on SGI R10000 and PC (Linux) and also compiled on Red Hat, with glibc2.0. MUTYH (MutY Homolog E. coli) protein domain I determined by (PDB) ID: 1X5I contained several mutations. We selected 6 mutations associated with heritable predisposition to colon and stomach cancer syndrome of FAP. Multiple transcript variants encoding different isoforms have been found for this gene. (Entrez Gene ID: Q9UIF7; MUTYH). Several proteins interacted with MUTYH. These proteins were used for further docking studies with four other proteins individually with and without mutations. We also compared the docking predictions with other programs, such as Fast Interaction Refinement in Molecular Docking (FIREDOCK) (Andrusier 2007). This method simultaneously targets the problem of
flexibility and scoring of solutions produced by fast rigid-body docking algorithms. The side-chain flexibility was modeled by rotamers and the obtained combinatorial optimization problem was solved by integer linear program (Kingsford 2005). Following rearrangement of the side chains of two proteins (MUTYH domain I with one of four other proteins), the relative position of the docking partners were refined by Monte Carlo (MC) minimization of the binding score function. The binding score ranks candidates and includes atomic contact energy (Zhang 1997), softened van der Waals (vdW) interactions, partial electrostatics and additional estimations of the binding free energy. The interaction studies were carried out using Knowledge-based FADE and Contacts (KFC Server) predicted binding "hot spots" for proteins and Protein-protein interactions by recognizing structural features indicative of important binding contacts with our proteins. These servers analyzed several chemical and physical features surrounding interface residues by using support vector machines (Zhu 2011).

**Molecular Dynamic Simulation of MUTYH Protein**

MD simulations were determined by using the bio-molecular simulation program CHARMM. CHARMM-GUI (http://www.charmm-gui.org) (Jo 2008), offers a web-based graphical user interface to analyze a range of input files and molecular systems to standardize and make use of common as well as advanced simulation techniques in CHARMM. We used selected structures of MUTYH domains I, and II were used in the MD simulation method. In this simulation, the solvated system was neither minimized nor equilibrated; however, 0.15M ions were added in the simulation box by specifying ions (KCl) and concentration (C). The numbers of ions were automatically established by the ion-accessible volume (V), the total charge of the system (Qsys), and by the positive ion (z+) valency, to neutralize the total system charge, \( z^+N^+=z^--Q_{sys} \). The ion-accessible volume, V, was anticipated by subtracting molecular volume from the total system volume, selecting ion-placing method of Monte Carlo (MC). The solvation free energy was expressed as nonpolar and electrostatic contributions, but the nonpolar contribution was again partitioned into repulsive and dispersive contributions using the weak, free energy simulations were performed with explicit solvent water molecules in close proximity to the solute, whereas the influence of the rest of the solvent mass was showed implicitly as an effective Spherical Solvent Boundary Potential (SSBP). KCl was included to neutralize the overall negative charge of the MUTYH protein structure.

The MD simulations were performed with a 2-fs time step, which was smaller than the fastest vibrational period of linear bonds involved to small hydrogen atoms at a constant temperature of 300K and a constant pressure of 1atm under periodic solvent boundary conditions. The periodic box was enriched with water molecules up to 0.72805g/l density. Addition of counter ions counterbalanced the entire system and every ionizable residue were protonated based on their pKa values (pH 7). Using simulated annealing process water molecules were relaxed. Energy minimization was run until the maximum atom speed dropped below 2000m/s, the system was then heated from 0 to 300K. Finally, at 300K and 10,000 ps, equilibrated MD simulation were performed at constant pressure with 12Å vdw cut-off. The electrostatic interactions were analyzed and 2000 snapshots were obtained at each 6.5ps (Essmann U et al 1995). The Particle Mesh Ewald (PME) method (Hockney 1988) was utilized for electrostatics, and a 12Å cutoff was applied for vdw interactions. The TIP3P water model (Horn 2004) was used as the solvent. The initial configuration of ions was verified using short Monte Carlo simulations with a primitive model, for instance vdw interactions. Mutations were studied using The Collaborative Computational Project Number 4 in Protein Crystallography and Molecular Graphic
CCP4MG V.2.5.0 (Potterton 2004), and energy minimization were carried out using NOMAD-Ref server (Lindahl 2006). This server utilizes Gromacs force field for energy minimization according to the steepest descent, conjugate gradient, and Broyden Fletcher Goldfarb–Shanno (L-BFGS) methods (Delarue 2004). We also used the conjugate gradient method for 3D structure optimization and the deviation between the two structures was assessed by their RMSD values.

Conclusions

In conclusion, the resulted presented in this study for the first time show that based on computational approaches, the deleterious SNPs in Mutyh gene, functional network and its relationships with others genes were analyzed via bio-molecular interactions. We also performed comparative MD analyses of the MUTYH protein interactions with other proteins physically interacting with it as well as between the wild type and mutant type structures. The MUTYH protein domains I had six variants P54S, A22V, Q63R, G45D, S136P and N43S variants and likely have great clinical significance associated with cancer risk, whereas protein domains II had two Y165C and G382D. Based on gene ontology the MUTYH protein showed interaction with six proteins, however only four of them (hMSH6, hPCNA, hRPA1, and hAPEX1), exhibited physical interactions. The results suggested that the MUTYH mutations caused decomposition of binding free energies and structural analyses indicated that it also distorted the geometry of the binding site and hence weakened the interactions. Overall results showed that such a computational approach can prove to be very useful for understanding the molecular basis of a disease and significantly contribute to our understanding of protein-protein interactions, mutational implications on the protein structure and its impact on disease progression.

Acknowledgement

The Authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding the work through the research group project No: RGP-VPP-200.

Author contributions

Z.A., J.D., and W.K., designed research; Z.A., J.D., and M.A. performed research; Z.A., W.K., and M.A. contributed new reagents/ analytic tools; M.A., Z.A., and W.K. analyzed data; and M.A., Z.A., and J.D. wrote the paper. Principle Investigator & Supervisor: J.D

Disclosure of potential conflicts of interest

The author declares that they have no conflict of interest.

Research involving human participants and/or animals

Ethical approval: “All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.”
Informed consent: “Informed consent was obtained from all individual participants included in the study.”

References

Al-Tassan N, et al. (2002) Inherited variants of MYH associated with somatic G:C-->T:A mutations in colorectal tumors. Nat Genet 30(2):227-232.

Ames BN & Gold LS (1991) Endogenous mutagens and the causes of aging and cancer. Mutat Res 250(1-2):3-16.

Ashburner M, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-29.

Andrusier N, Nussinov R, & Wolfe HJ (2007) FireDock: fast interaction refinement in molecular docking. Proteins 69(1):139-159.

Anna Hart (2011) Mann-Whitney test is not just a test of medians: differences in spread can be important. BMJ; 333(7543): 391–393.

Alfaro C, et al. (2005) The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 33(Database issue):D418-424.

Brannon AR (2014). Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. Genome Biol 15(8):454.

Bai H, Grist S, Gardner J, et al. Functional characterization of human MutY homolog (hMYH) missense mutation (R231L) that is linked with hMYH-associated polyposis. Cancer Lett. 2007;250:74–81.

Brunner HG & van Driel MA (2004) From syndrome families to functional genomics. Nat Rev Genet 5(7):545-551.

Boldogh I, et al. (2001) hMYH cell cycle-dependent expression, subcellular localization and association with replication foci: evidence suggesting replication-coupled repair of adenine:8-oxoguanine mispairs. Nucleic Acids Res 29(13):2802-2809.

Cheadle JP & Sampson JR (2007) MUTYH-associated polyposis—from defect in base excision repair to clinical genetic testing. DNA Repair (Amst) 6(3):274-279.

Croitoru ME, et al. (2004) Association between biallelic and monoallelic germ line MYH gene mutations and colorectal cancer risk. J Natl Cancer Inst 96(21):1631-1634.

Colussi C, et al. (2002) The mammalian mismatch repair pathway removes DNA 8-oxodGMP incorporated from the oxidized dNTP pool. Curr Biol 12(11):912-918.

David SS, O'Shea VL, & Kundu S (2007) Base-excision repair of oxidative DNA damage. Nature 447(7147):941-950.
De la Rue M & Dumas P (2004) On the use of low-frequency normal modes to enforce collective movements in refining macromolecular structural models. Proc Natl Acad Sci USA 101(18):6957-6962.

Essmann U, Perera L, Berkowitz ML (1995) A smooth particle-mesh Ewald potential. Journal of Computational Physics 103:5577-5892.

Eisenberg D, Marcotte EM, Xenarios I, & Yeates TO (2000) Protein function in the post-genomic era. Nature 405(6788):823-826.

Farrington SM, et al. (2005) Germline susceptibility to colorectal cancer due to base-excision repair gene defects. Am J Hum Genet 77(1):112-119.

Gismondi V, et al. (2004) Prevalence of the Y165C, G382D and 1395delGGA germline mutations of the MYH gene in Italian patients with adenomatous polyposis coli and colorectal adenomas. Int J Cancer 109(5):680-684.

Gu Y, et al. (2002) Human MutY homolog, a DNA glycosylase involved in base excision repair, physically and functionally interacts with mismatch repair proteins human MutS homolog 2/human MutS homolog 6. J Biol Chem 277(13):11135-11142.

Hashiguchi K, Zhang QM, Sugiyama H, Ikeda S, & Yonei S (2002) Characterization of 2-hydroxyadenine DNA glycosylase activity of Escherichia coli MutY protein. Int J Radiat Biol 78(7):585-592.

Hoadley KA (2018). Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell 173(2):291-304.e6.

Hashimoto K, Tominaga Y, Nakabeppu Y, & Moriya M (2004) Futile short-patch DNA base excision repair of adenine:8-oxoguanine mispair. Nucleic Acids Res 32(19):5928-5934.

Hockney RW, & Eastwood JW. (1988) Computer Simulation using Particles. McGraw-Hill.

Hurst JM, et al. (2009) The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 30(4):616-624.

Horn HW, et al. (2004) Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J Chem Phys 120(20):9665-9678.

Janin J, et al. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52(1):2-9.

Joshi-Tope G, et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33(Database issue):D428-432.

Jain E, et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10:136.
Nakabeppu Y (2001) Regulation of intracellular localization of human MTH1, OGG1, and MYH proteins for repair of oxidative DNA damage. Prog Nucleic Acid Res Mol Biol 68:75-94.

Noll, D. M., A. Gogos, J. A. Granek, and N. D. Clarke. 1999. The C-terminal domain of the adenine-DNA glycosylase MutY confers specificity for 8-oxoguanine. adenine mispairs and may have evolved from MutT, an 8-oxo-dGTPase. Biochemistry 38:6374-6379.

Out AA, et al. (2010) Leiden Open Variation Database of the MUTYH gene. Hum Mutat 31(11):1205-1215.

Ohtsubo T, et al. (2000) Identification of human MutY homolog (hMYH) as a repair enzyme for 2-hydroxyadenine in DNA and detection of multiple forms of hMYH located in nuclei and mitochondria. Nucleic Acids Res 28(6):1355-1364.

Potterton L, et al. (2004) Developments in the CCP4 molecular-graphics project. Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1):2288-2294.

Parlanti E, Pasucci B, Terrados G, Blanco L, & Doglioni E (2004) Aphidicolin-resistant and -sensitive base excision repair in wild-type and DNA polymerase beta-defective mouse cells. DNA Repair (Amst) 3(7):703-710.

Parker AR & Eshleman JR (2003) Human MutY: gene structure, protein functions and interactions, and role in carcinogenesis. Cell Mol Life Sci 60(10):2064-2083.

Parker A, et al. (2001) Human homolog of the MutY repair protein (hMYH) physically interacts with proteins involved in long patch DNA base excision repair. J Biol Chem 276(8):5547-5555.

Peri S, et al. (2004) Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res 32(Database issue):D497-501.

Prlic A, et al. (2007) Integrating sequence and structural biology with DAS. BMC Bioinformatics 8:333.

Ramensky V, Bork P, & Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894-3900.

Ramaswamy S, Ross KN, Lander ES, Golub TR. A 2003 molecular signature of metastasis in primary solid tumors. Nat Genet. 33(1):49–54. pmid:12469122

Roy A, Kucukural A, & Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5(4):725-738.

Stelzl U, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122(6):957-968.

Schulmann K, Pox C, Tannapfel A, & Schmiegel W (2007) The patient with multiple intestinal polyps. Best Pract Res Clin Gastroenterol 21(3):409-426.

Giannakis M (2016). Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. Cell Rep 15(4):857-865.

Gehan A (1965) A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples. Biometrika. 52(1/2):203-23.
Seshagiri S (2012). Recurrent R-spondin fusions in colon cancer. *Nature* 488(7413):660-4.

Slupska MM, Luther WM, Chiang JH, Yang H, & Miller JH (1999) Functional expression of hMYH, a human homolog of the *Escherichia coli* MutY protein. *J Bacteriol* 181(19):6210-6213.

Shinmura K, Yamaguchi S, Saitoh T, Kohno T, Yokota J: Somatic mutations and single nucleotide polymorphisms of base excision repair genes involved in the repair of 8-hydroxyguanine in damaged DNA. *Cancer Letters* 2001, 166:65-69

Sampson JR, Dolwani S, Jones S, Eccles D, Ellis A, Evans DG, Frayling I, Jordan S, Maher ER, Mak T, Maynard J, Pigatto F, Shaw J, Cheadle JP: Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet* 2003, 362:39-41

Sieber OM, et al. (2003) Multiple colorectal adenomas, classic adenomatous polyposis, and germ-line mutations in MYH. *N Engl J Med* 348(9):791-799.

Slupska MM, et al. (1996) Cloning and sequencing a human homolog (hMYH) of the *Escherichia coli* mutY gene whose function is required for the repair of oxidative DNA damage. *J Bacteriol* 178(13):3885-3892.

Shinmura K, et al. (2000) Adenine excisional repair function of MYH protein on the adenine:8-hydroxyguanine base pair in double-stranded DNA. *Nucleic Acids Res* 28(24):4912-4918.

Tovchigrechko A & Vakser IA (2006) GRAMM-X public web server for protein-protein docking. *Nucleic Acids Res* 34(Web Server issue):W310-314.

Thusberg J, Olutubosun A, & Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* 32(4):358-368.

Thomas PD, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. * Genome Res* 13(9):2129-2141.

Turner FS, Clutterbuck DR, & Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 4(11):R75.

Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, & Vriend G (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* 11:548.

Wooden SH, Basset HM, Wood TG, et al. Identification of critical residues required for the mutation avoidance function of human MutY (hMYH) and implications in colorectal cancer. *Cancer Lett*. 2004;205:89–95.

Yang H, et al. (2001) Enhanced activity of adenine-DNA glycosylase (Myh) by apurinic/apyrimidinic endonuclease (Ape1) in mammalian base excision repair of an A/GO mismatch. *Nucleic Acids Res* 29(3):743-752.

Yaeger R. (2018). Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer. *Cancer Cell* 33(1):125-136.e3.

Zhang C, Vasmatis G, Cornette JL, & DeLisi C (1997) Determination of atomic desolvation energies from the structures of crystallized proteins. *J Mol Biol* 267(3):707-726.
FIGURE LEGENDS

Fig. 1.
Colonoscopy screening pictures of polyps. Adenoma-carcinoma sequence of most colorectal cancers arise from adenomatous polyps, some of which progress from small (<5 mm) to large polyps (>1.0 cm), and then to dysplasia and cancer.

Fig. 2.
(A) Base excision repair (BER) pathway induced by oxidative and alkylating agents initiates repair of AP sites in DNA by catalyzing hydrolytic incision of the phosphodiesterase backbone immediately adjacent to the damage. Mutyh structure was in yellow were a DNA duplex containing an 8-oxoG:A mismatch and, in contrast to the E. coli ortholog, the human enzyme has less activity for other mismatches such as G:A or C:A. DNA glycosylase starts BER by removing 8-oxoG from 8-oxoG:C base pairs. In the second pathway, the Mutyh DNA glycosylase takes out adenine inserted opposite 8-oxoG during replication. (B) The DNA mismatch repair protein hMSH6 structure is the DNA strand (upper) contains a mismatched base was in red. The strand discrimination signal, which probably corresponds to strand discontinuities at the replication fork, was located near the mismatch. In vitro, hPCNA was in red and RFC was in violet proteins were bound to the strand discontinuity that must be present in the DNA substrate reaction to MutS searches binds to DNA in the nucleotide-free or ADP-bound state. The binding of MutS to DNA induces bending of both homo- and hetero duplex DNA was in blue. When MutS encounters the mismatch, a bend was turned into a sharp kink. This triggers a conformational change in the protein and the exchange of ADP to ATP. (C) Replication protein 70kDa DNA-binding subunit hRPA1, Undamaged strand with hRPA1 reaction to Complex with XPG. (D) hPCNA was in pink color, fragment of damaged strand removed was part of this pathways reaction to both DNA Polymerase Epsilon filling the gap and DNA Polymerase Epsilon filling the gap in these pathways.

Fig. 3.
(A) Disease relationships for MUTYH. (B) MUTYH interacting proteins.

Fig. 4.
(A) Overview of clinical attributes and a scatter plot of mutation count versus fraction of genome altered for each case in the TCGA endometrial cancer study. (B) The overall survival of Primary and
Metastasis. (C) The overall survival of Living and Decreases. (D) ERBB2 mRNA expression is increased in samples with DNA amplification, and ERBB2 protein abundance was higher in samples with increased mRNA. A plot showed the relationship between ERBB2 mRNA abundance and CNA in the ERBB2 genes i.e. (MutyH, hPCNA, hRPA1, hAPEX1 and hMSH6) in tumors from the selected colorectal cancer. The “x” indicates individual tumors, and the circles indicate tumors with missense mutations. (E) The overall survival for the total number of average infected both male and females.

Fig. 5.

The alignments utilized for prediction were closely associated high median conservation, and then many positions appear conserved and important for function. Prediction accuracy on deleterious substitutions was high but many functionally neutral substitutions were erroneously predicted to become deleterious. SIFT calculates the median conservation value which measures the diversity of the sequences in the alignment. There were 21 SNPs analysis for the functional studies but only 6 SNPs have damaging (0.001) was in red, other were tolerated was in blue that were based in M (Median Conservation score), and green (Normal). The score was the regularized probability that the amino acid alteration was tolerated. The SIFT predicts substitutions with scores less than 0.05 as deleterious. Some of found that substitutions with scores less than 0.1. The quantitative score allows prioritizing their amino acid variations by ranking them from the lowest scores to the highest.

Fig. 6.

Map of MutyH gene showing location of both Exons and Intron, and of 32 SNPs from 90 samples, but with the strength of association of the nsSNPs (recessive model for the risk allele), a common samples with combine variation indicated by color. (A) A haplotypes and its tagging SNPs spanning Intron 1-6 were most strongly associated with colon cancer. The frequency and no-monomorphic filters were applied to set of genotypes. For tag SNPs colors was in black, if there were multiple population groups, these filters were applied to the set of genotypes for each population group. Allele Frequency Cutoff (%): 0, monomorphic sites excluded R² Threshold for Clusters: 0.8, Minimal Genotype Coverage (%) of SNPs to Be Tag SNPs: 85, Minimal Genotype Coverage (%) of SNPs to Be Clustered: 70, Data Merging: common samples with combined variations in Population was PDR90. (B) HapMap linkage disequilibrium for MutyH gene was shown. Analysis of linkage disequilibrium (LD) between polymorphic sites in a locus identifies "clusters" of highly correlated sites based on the r² LD statistic. Sites were binned into sets of highly informative markers to minimize redundant our data. This result was for the development of a minimal set of SNPs that can be used for large-scale genotyping of similar sample populations. In both (A and B) figures showing color representing was in pink (splice variant), red (non-synonymous), greed (coding synonymous), blue (coding) and brown (mRNA UTR).

Fig. 7.
Panel of MUTYH mutations in-group of wild-type residues was shown in green and the mutant a residue was shown in purple (A-D and F-G). Other remains wild type a residue was shown in pink and mutant residues were shown in blue (c and h). Relevant side chains that interact with the residue of interest were depicted and labeled. β-strands, helices, and coils indicate the secondary structure elements that form the scaffold for the interacting residues. The impact of each mutation on the protein structure was assessed as follows. (A) The wild-type residue was more hydrophobic than the mutant residue. The mutation was located within a domain. The wild-type residue was a proline (Pro 54A) was in violet and mutant residue (Ser 54B) was in green. Prolines were known to be very rigid and therefore induce a special backbone conformation which might be required at this position. (B) The mutant residue was a valine (Val 22B) was in green and wild type residue (Ala 22A) was in violet. The wild-type residue was not conserved at this position, our mutant residue was among the observed residue types at this position in other homologous sequences. (C) Mutation of a Glutamine into an Arginine at position 63. The original wild-type residue (Gln 63B), was in violet and newly introduced mutant residue was in green often differ in these properties. The mutant residue was bigger than the wild-type residue (Arg 63A). The wild-type residue was neutral, the mutant residue was positively charged. The mutation was located within a domain, annotated in Uniprot as: "Nudix hydrolase". (D) The wild-type residue was a glycine (Gly 45A) was in violet, the most flexible residue of all. This flexibility might be necessary for the protein's function. Mutation of this Aspartic acid (Asp 45B), was in green can abolish this function probably damaging to the protein. (E) The mutant residue (Pro 136A) was in blue was bigger than the wild-type residue (Ser 136B) was in pink. The residue was located on the surface of the protein, mutation of this residue can disturb interactions with other molecules or other parts of the protein. (F) The size difference between wild-type (Asn 43A) was in violet and new mutant residue (Ser 43B) was in green makes that the new residue was not in the correct position to make the same hydrogen bond as the original wild-type residue did. The difference in hydrophobicity will affect hydrogen bond formation. (G) A mutation of wild type residue (Tyr 165A) was in violet and mutant residue (Cys 165B) was in green. Mutation of a 100% conserved residue was usually damaging for the protein. (H) The wild-type residue was a glycine (Gly 382A) was in pink and mutant residue (Asp 382B) was in blue, the most flexible residue of all. This flexibility might be necessary for the protein's function. Mutation of this aspartic acid can abolish this function. (I) Our protein residues contact energies before and after energy minimization can be used to locate functionally or structurally important residues, to determine and classify consequences of mutations, to identify residues for site-directed mutagenesis.

Fig. 8.

The positional candidate gene prioritization technique using gene networks, depicted in this figure were typically four different gene to gene interaction data sources that were combined in a Bayesian way. (A) In which four disease genes (hPcna, hMsh6, hApx1 and hRja1) have been implicated but there were only four having physical interaction. Susceptibility loci, each containing 100 genes, were defined around 10 known colorectal cancer genes. The 10 highest-ranked genes for each locus were shown in the network, with colors indicating the locus in which they reside. (B-C) the GO, MA and PPI network led to four colorectal cancer genes were high significant [hPcna, hMsh6, hApx1 and hRja1 (were in circles)] being ranked in the top 10 chromosome. All genes get a preliminary score of 0 assigned & four different susceptibility loci, each containing a disease gene were P (green), Q (yellow), and R (orange) should then finale up as the top ranked genes. (D) The most significant p-values was analyzed. Each locus the four positional candidate genes increase the scores of genes
functionally proximate within the gene network, by a kernel function which models the relationship between gene to gene distance & score effect.

**Fig. 9.**

PPI site in the MUTYH structure domain I in condition of wild type (control) interact with all wild type proteins. (A) The MUTYH structure was performed in order to identify potential protein-protein interaction sites. The first solvent-accessible surface of MUTYH was shown in yellow and binds with hMSH6 (grey) and the putative protein were in circle marks protein interaction site. The positions of amino acids with favorable desolvation energy were indicated in red showing interaction positions (Glu198, Gly148, Ser150, His113, Gln127, Pro152, Pro199, Ser196 and Gly159) of the side chains identified by the bioinformatics analysis. The hMSH6 protein (chain A) was drawn as a molecular surface. The side chains of the relevant interactive amino acids were shown as sticks. In binding energy in between MUTYH domain I and hMSH6 was -66.75 kcal/mol, intermolecular H-bond energy was shown as green -7.45 kcal/mol. (B) The second solvent-accessible surface of MUTYH and binds with hPCNA protein was in grey and the putative PPI site were marked by a circle. The position of amino acids with favorable desolvation energy was indicated in red showing interaction positions (Glu111, Pro29, Ile255, Thr206 and Leu205) of the side chains. The hPCNA protein chain A was drawn as a molecular surface. The side chains of the relevant interactive amino acids were shown as sticks. In binding energy in between MUTYH domain I and hPCNA was -59.02 kcal/mol, intermolecular H-bond in between MUTYH and hPCNA (His44, Ser154 and Ser153) energy was shown as green -2.36 kcal/mol. (C) The third solvent-accessible surface of MUTYH was shown in yellow and binds with hAPEX1 protein was in grey and the putative PPI site was marked by a circle. The position of amino acids with favorable desolvation energy was indicated in red and blue showing interaction positions (Lys103, Arg75, Glu72, Asn102, Leu104, Tyr118 and Gln71) of the side chains. The hAPEX1 protein chain A was drawn as a molecular surface. The side chain of the relevant interactive amino acids was shown as sticks. In binding energy in between MUTYH domain I and hAPEX1 was -33.85 kcal/mol, intermolecular H-bond MUTYH and hAPEX1 (Lys103, Glu59, Ser143, Tyr117 and Val142) energy was shown as green -3.14 kcal/mol. (D) The fourth solvent-accessible surface of MUTYH structure was in yellow and binds with hRPA1 protein were in grey and the putative PPI site was marked by a circle. The position of amino acids with favorable desolvation energy was indicated in red showing interaction positions (Pro29, Met1, Gln35, Leu102, Val176) of the side chains. The hRPA1 protein chain A was drawn as a molecular surface. The side chains of the relevant interactive amino acids were shown as sticks. In binding energy in between MUTYH domain I and hRPA1 was -93.60 kcal/mol; intermolecular H-bond in between MUTYH and hRPA1 (Ser150, Met1, Tyr146, Pro23 and Ala31) energy was shown as green -4.52 kcal/mol.

**Fig. 10.**

PPI site with MUTYH structure domain I in condition of effected (mutant type) interact with all wild type other proteins. (A) The MUTYH structure was performed in order to identify potential PPI sites. The first solvent-accessible surface of MUTYH was shown in yellow and bind with hMSH6 was in grey and the putative protein a circle marks protein interaction site. The position of amino acids with favorable desolvation energy was indicated in red showing interaction positions (Gly148, Thr137, Gln127, Arg140, Gly111 and Phe115) of the side chains. The hMSH6 protein chain A was drawn as a molecular surface. The side chain of the relevant interactive amino acids was shown as sticks. In
binding energy in between hMSh6 and MUTYH protein domain I was -41.43 kcal/mol, intermolecular H-bond energy was shown as green -7.92 kcal/mol. (B) the second solvent-accessible surface of MUTYH and binds with hPCNA protein was in grey and the putative PPI site was marked by a circle. The position of amino acids with favorable desolvation energy was indicated in red showing interaction positions (Asn36, Leu50, Thr51, Gln125 and Leu47) of the side chains. The hPCNA protein chain A was drawn as a molecular surface. The side chain of the relevant interactive amino acids was shown as sticks. In binding energy in between MUTYH domain I and hPCNA were -74.40 kcal/mol, intermolecular H-bond in between MUTYH and hPCNA (Gln125, Thr137, Asp29, Asn65, Gln40 and Thr51) energy was shown as green -7.92 kcal/mol. (C) The third solvent-accessible surface of MUTYH was shown in yellow and bind with hAPEX1 was in grey and the putative PPI site was marked by a circle. The position of amino acids with favourable desolvation energy was indicated in red and blue showing interaction positions (Gln117, Gln71, Glu72, Arg75, Tyr118, Leu104 and Lys103) of the side chains. The hAPEX1 (chain A) was drawn as a molecular surface. The side chain of the relevant interactive amino acids was shown as sticks. In binding energy in between MUTYH domain I and hAPEX1 was -57.87 kcal/mol, intermolecular H-bond of MUTYH and hAPEX1 (Glu59, Lys103, Val142 and Thr117) energy was shown as green -4.94 kcal/mol. (D) The fourth solvent-accessible surface of MUTYH was shown in yellow and binds with hRPA1 protein was in grey and the putative PPI site was marked by a circle. The position of amino acids with favorable desolvation energy was indicated in red showing interaction positions (Met1, Gln35, Val176, Leu38, Glu129 and Leu102) of the side chains. The hRPA1 protein (chain A) was drawn as a molecular surface. The side chain of the relevant interactive amino acids was shown as sticks. In binding energy in between hRPA1 and MUTYH domain I was -92.83 kcal/mol, intermolecular H-bond in between MUTYH and hRPA1 (Thr116, Glu100, Ala31, Pro23, Ser150 and Met1) energy was shown as green -5.71 kcal/mol.

Fig. 11.

(A) The isolated NMR structure of NUDX from human A/G specific adenine DNA glycosylase alpha-3 splice isoform. (B) The MUTYH homolog protein domain I PDB: 1X51 with six mutations (S136P, G45D, N43S, A22V, Q63R and P54S) before energy minimization. (C) The MUTYH protein domain I solvated in to simulation box for MD simulation by solvent boundary utilizer and solvated system cutoff 10Å. (D) After MD simulation and energy minimization of MUTYH L-histidine uncharged with a proton on NE2 an isomer of HIS (HSD131, HSD93, HSD97 and HSD87). (E) The MUTYH protein structure domain II predicted from ID: Q9UIF7 along with prediction of two mutations at before energy minimization. (F) The MUTYH protein domain II solvated in to simulation box of truncated octahedron explicit water solvated system with water molecules cutoff same 10Å. (G) After MD simulation and energy minimization of MUTYH L-histidine uncharged with a proton on NE2 an isomer of HIS (HSD 186, HSD 39, HSD 528, HSD 522, HSD 468, HSD 184, HSD 26, HSD 326 and HSD 421).

TABLE LEGEND

Table 1.

The Impact of Single Point Mutations on Protein Stability
Predict disease related polymorphism on *MutY homolog (E. coli)* on MUTYH protein structure.

**Neutral:** Neutral Polymorphism; **Disease:** Disease-related Polymorphism; **F [X]:** Frequency of residue X in the sequence profile. **Nali:** Number of aligned sequences in the mutated site; **NGOs:** Number of Gene Ontology terms; **LGOs:** Total log odd score of all the Gene Ontology terms.

**Table 2.**

*Summary of MutY homolog (E. coli) SNPs within the Exonic Mutyh gene regions*

**Table 3.**

*The Impact in the Protein Structure Relative Stability upon Single-Site Mutation*

Protein stability changes upon mutation. Wild type (WT); and mutant amino acid (MT), Solvent accessibility, in percent (Acc), Predicted stability energy values ΔΔG, in kcal/mol (ΔΔGp).

**Table 4.**

*Empirical Protein-Protein Interaction of Wild and Mutant type Structure*

Binding interaction in between controls (wild type) & effected (Mutant type) proteins. VdW - Van der Waals forces to the binding energy, ACE - Atomic contact energy (ACE) to the binding energy, HB - Hydrogen bonds to the binding energy.
Fig 1.
Fig 2.
**Fig 3.**

- **A:** DNA oxidative damaged base in Homo sapiens
  - G-to-T transversions
  - MutYH recognizes deaminated, alkylated or mismatched bases
  - A from 8-oxoG+T
  - DNA with damaged base removed (LP-BER)
  - 8-oxoG from 8-oxoG+T
  - Damaged DNA with bifunctional glycosylase (With AP lyase activity)

- **B:** Heteroduplex DNA with nick 5' to mismatch
  - ERC
  - Mismatch in heteroduplex DNA with nick 5' to mismatch bound by MutSalpha
  - Mismatch in heteroduplex DNA with nick 3' to mismatch bound by MutSalpha-MutSalpha

- **C:** Fragment of damaged strand removed
  - Undamaged stand with RPA
  - Complex with XPG

- **D:** DNA Polymerase Delta filling the gap
  - DNA Polymerase Epsilon filling the gap
**Fig 4.**

**A.** MutYH gene interacting proteins

**B.** Disease relationships for MutYH gene
Fig 7.
Fig 8.
Fig 9.
Table 1. The Impact of Single Point Mutations on Protein Structure

| dbSNPs ID     | Chain | Mutation | Secondary structure | Solvent Accessibility | Prediction | RI | Probability | Method       |
|---------------|-------|----------|---------------------|-----------------------|------------|----|-------------|--------------|
| rs121908382   | A     | P54S     | Coil                | 0% Buried             | Disease    | 4  | 0.699       | F[P]=100% F[S]=0% Nali=25 |
| rs35352891    | A     | A22V     | Beta Ladder         | 60% Buried            | Disease    | 1  | 0.548       | F[A]=75% F[V]=0% Nali=25 |
| rs121908383   | A     | Q63R     | Helix               | 68% Exposed           | Neutral    | 1  | 0.445       | F[Q]=55% F[R]=4% Nali=25  |
| rs36053993    | A     | G45D     | Bend                | 70% Exposed           | Disease    | 7  | 0.829       | F[G]=100% F[D]=0% Nali=25 |
| rs14461656    | A     | S136P    | Coil                | 8% Buried             | Disease    | 6  | 0.794       | F[S]=85% F[P]=8% Nali=25  |
| rs76641561    | A     | N43S     | Coil                | 50% Buried            | Neutral    | 6  | 0.202       | F[N]=20% F[S]=31% Nali=25 |
| rs34612342    | A     | Y165C    | Loop                | 29% Intermediate      | Disease    | 8  | 0.902       | F[Y]=85% F[S]=0% Nali=25  |
| rs36053993    | A     | G382D    | Bend                | 70% Exposed           | Disease    | 4  | 0.679       | SNP&GO       |

Predict disease related polymorphism on *MutY homolog (E. coli)* on *MutyH* protein structure. Neutral: Neutral Polymorphism; Disease: Disease-related Polymorphism; F[X]: Frequency of residue X in the sequence profile. Nali: Number of aligned sequences in the mutated site; NGOs: Number of Gene Ontology terms; LGOs: Total log odd score of all the Gene Ontology terms.

Table 2. The Impact in the Protein Structure Relative Stability upon Single-Site Mutation

| Name                  | Chain | Position | WT    | MT    | pH    | T    | Solvent accessibility (Acc) | Torsion angles (°) | Energy ΔΔG (kcal/mol) | Stability  |
|-----------------------|-------|----------|-------|-------|-------|------|-----------------------------|---------------------|------------------------|------------|
| *Mutyh* (Pro-Ser);(P→S) | A     | P54S     | P     | S     | 7.0   | 27   | 0.00%                        | -69.8°, 144.1°      | 1.47                   | Destabilizing          |
| *Mutyh* (Ala-Val);(A→V) | A     | A22V     | A     | V     | 7.0   | 27   | 0.00%                        | -1118.0°, 105.1°    | 0.03                   | Destabilizing          |
| *Mutyh* (Gln-Arg);(Q→S) | A     | Q63R     | Q     | R     | 7.0   | 27   | 69.36%                       | -92.5°, -49.9°      | 0.15                   | Destabilizing          |
| *Mutyh* (Gly-Asp);(G→D) | A     | G45D     | G     | D     | 7.0   | 27   | 66.97%                       | 115.9°, -145.4°     | 0.69                   | Destabilizing          |
| *Mutyh* (Pro-Ser);(S→P) | A     | S136P    | S     | P     | 7.0   | 27   | 7.70%                        | -68.3°, 140.1°      | 0.39                   | Destabilizing          |
| *Mutyh* (Asn-Ser);(N→S) | A     | N43S     | N     | S     | 7.0   | 27   | 47.73%                       | -43.6°, 153.5°      | -0.08                  | Neutral               |
| *Mutyh* (Tyr-Cys);(Y→C) | A     | Y165C    | Y     | S     | 7.0   | 27   | 45.61%                       | -60.3°, 141.1°      | 0.70                   | Destabilizing          |
| *Mutyh* (Gly-Asp);(G→D) | A     | G382D    | G     | D     | 7.0   | 27   | 57.89%                       | -92.5°, -149.3°     | 0.64                   | Destabilizing          |

Protein stability changes upon mutation. Wild type (WT); and mutant amino acid (MT), Solvent accessibility, in percent (Acc), Predicted stability energy values ΔΔG, in kcal/mol (DDGp).
Table 3. Empirical Protein-Protein Interaction of Wild and Mutant type Structure

| Proteins              | Binding Energy Kcal/mol | Attractive VdW Kcal/mol | Repulsive VdW Kcal/mol | ACE Kcal/mol | HB Kcal/mol |
|-----------------------|-------------------------|-------------------------|------------------------|--------------|-------------|
| **Control + Control (Wild type)** |                         |                         |                        |              |             |
| MYH + MSH6            | -66.75                  | -38.82                  | 23.44                  | -11.91       | -7.45       |
| MYH + PCNA            | -59.02                  | -35.27                  | 12.73                  | -17.90       | -2.36       |
| MYH + APEX1           | -33.85                  | -41.56                  | 25.19                  | 5.40         | -3.14       |
| MYH + RPA1            | -93.60                  | -45.60                  | 31.87                  | -20.03       | -4.52       |
| **Mutant (effected) + Control (Wild type)** |                         |                         |                        |              |             |
| MYH + MSH6            | -41.43                  | -51.94                  | 36.61                  | 1.90         | -7.92       |
| MYH + PCNA            | -74.40                  | -47.41                  | 19.63                  | -6.31        | -9.97       |
| MYH + APEX1           | -57.87                  | -37.27                  | 25.50                  | -15.94       | -4.94       |
| MYH + RPA1            | -92.83                  | -48.39                  | 40.78                  | -21.52       | -5.71       |

Binding interaction in between controls (wild type) & effected (Mutant type) proteins. VdW - Van der Waals forces to the binding energy, ACE - Atomic contact energy (ACE) to the binding energy, HB - Hydrogen bonds to the binding energy.
### Table 4. Summary of MutY homolog (E. coli) SNPs within the Exonic MutyH gene regions

| Chromosome1 & Position | mRNA pos | Region | dbsNPs ID | dbSNPs Allele | Allele frequency | Peptide Allele | Codon position | Amino acid Position | Codon position | Amino acid | Function | Clinical significance |
|-----------------------|----------|--------|-----------|--------------|----------------|---------------|---------------|-------------------|----------------|-------------|----------|----------------------|
| 45795027              | 1817     |        | rs3219497 | T            | 0.0183         | R             | Q             | 2                 | 534            | Missense   | Probable-pathogenic |
| 45795043              | 1801     |        | rs3219496 | T            | 0.002          | L             | M             | 1                 | 529            | Missense   | Probable-pathogenic |
| 45796198              | 1724     | Exon 3' | rs3219494 | T            | 0.002          | G             | E             | 2                 | 503            | Missense   | Probable-pathogenic |
| 45797505              | 1230     |        | rs3219489 | G            | 0.403          | Q             | H             | 3                 | 338            | Missense   | Probable-pathogenic |
| 45799126              | 5213     |        | rs1140507 | G            | N.D            | W             | R             | 1                 | 103            | Missense   | Probable-pathogenic |
| 45800156              | 280      |        | rs3219484 | T            | 0.098          | V             | M             | 1                 | 22             | Missense   | Probable-pathogenic |
| 45797228              | 1403     |        | rs36053993| A            | 0.009          | G             | D             | 2                 | 396            | Missense   | Probable-pathogenic |
| 45796198              | 1724     |        | rs3219494 | A            | 0.002          | G             | E             | 2                 | 503            | Missense   | Probable-pathogenic |
| 45797174              | 1457     |        | rs121908383| G            | N.D            | Q             | R             | 2                 | 414            | Missense   | Probable-pathogenic |
| 45797401              | 1334     |        | rs35552891| T            | 0.010          | A             | V             | 2                 | 373            | Missense   | Probable-pathogenic |
| 45797202              | 1429     |        | rs121908382| T            | N.D            | P             | S             | 1                 | 405            | Missense   | Probable-pathogenic |
| 45798457              | 752      |        | rs34612342 | A            | 0.005          | Y             | C             | 2                 | 165            | Missense   | Probable-pathogenic |
| 45797228              | 1403     |        | rs36053993| A            | 0.009          | G             | D             | 2                 | 382            | Missense   | Probable-pathogenic |
| 45797505              | 1230     |        | rs3219489 | C            | 0.3037         | Q             | H             | 1                 | 338            | Missense   | Probable-pathogenic |
| 45795027              | 1817     |        | rs3219497 | A            | 0.0183         | R             | Q             | 2                 | 534            | Missense   | Probable-pathogenic |
| 45795043              | 1801     |        | rs3219496 | A            | 0.002          | L             | M             | 1                 | 529            | Missense   | Probable-pathogenic |
| 45796198              | 1724     |        | rs3219496 | A            | 0.002          | L             | M             | 1                 | 529            | Missense   | Probable-pathogenic |
| 45796198              | 1724     |        | rs3219494 | A            | 0.002          | G             | E             | 2                 | 503            | Missense   | Probable-pathogenic |
| 45800156              | 280      |        | rs3219484 | A            | 0.098          | V             | M             | 1                 | 22             | Missense   | Probable-pathogenic |