Journal of Computational Biology and Bioinformatics Research Vol. 5(1), pp. 1-5, April 2013
Available online at http://www.academicjournals.org/JCBBR
DOI: 10.5897/JCBBR12.012
ISSN 2141-2227 ©2013 Academic Journals

Full Length Research Paper

MLH1 gene: An *in silico* analysis

Amitha Joy*, Jubil, C. A, Syama, P. S and Rohini Menon

Department of Biotechnology, Sahrdaya college of Engineering and Technology, Kodakara-680684, Thrissur, India.

Accepted 7 February, 2013

The MLH1 gene responsible for colon cancer has been examined to identify functional consequences of single-nucleotide polymorphisms (SNPs). 16 SNPs have been identified in the MLH1 gene in which all are found to be nonsynonymous. Non synymous SNPs are relevant in many of the human inherited diseases since they change the aminoacid sequence of the protein. 56% of the identified nsSNPs have been reported as damaging. In the analysis of SNPs using SIFT, UTRscan, FastSNP and PolyPhen-2, it was recognized that rs41295284 and rs35001569 were responsible for the alteration in levels of expression. It has been concluded that among all SNPs of MLH1 gene, the mutation in rs41295284 and rs35001569 have the most significant effect on functional variation.

Key words: Single nucleotide polymorphism, non-synonymous, colon cancer.

INTRODUCTION

A single nucleotide polymorphism (SNP) is a source variance in a genome. A SNP is a single base mutation in DNA. SNPs are the most simple form and most common source of genetic polymorphism in the human genome (90% of all human DNA polymorphisms (Smith, 2002). A SNP in a coding region may have two different effects on the resulting protein: Synonymous, the substitution causes no amino acid change to the protein it produces; non-synonymous, the substitution results in an alteration of the encoded amino acid. One half of all coding sequence SNPs result in non-synonymous codon changes (Smith, 2002). A non-synonymous single nucleotide polymorphism (nsSNP) occurring in a coding gene may cause an amino acid substitution in the corresponding protein product, thus affecting the phenotype of the host organism. Non-synonymous variants constitute more than 50% of the mutations known to be involved in human inherited diseases (Kumar et al., 2009).

Familial colorectal cancer (CRC) is a major public health problem by virtue of its relatively high frequency. Hereditary non-polyposis colorectal cancer (HNPCC), also called Lynch syndrome, accounts for approximately 5-8% of all CRC patients. Among these, 3% are mutation positive that is, caused by germline mutations in the DNA mismatch repair genes that have so far been implicated (*MLH1*, *MSH2*, *MSH6*, *PMS1*, and *PMS2*) (Henry and Albert, 1999).

MLH1, MSH2 and MSH6 genes play an important role in repairing mistakes made in DNA replication in colon cancer [9]. In the present study, the role of the SNPs of MutL homolog 1 (MLH1) in disease mutations is discussed. MLH1 is a human gene located on the short (p) arm of chromosome 3 and base pair from 37,034,840 to base pair 37,092,336 and cytogenetic location: 3p21.3. Its Locus ID (NCBI) is 4292. This gene was identified as a locus frequently mutated in hereditary non polyposis colon cancer (HNPCC) (US National Library of Medicine).

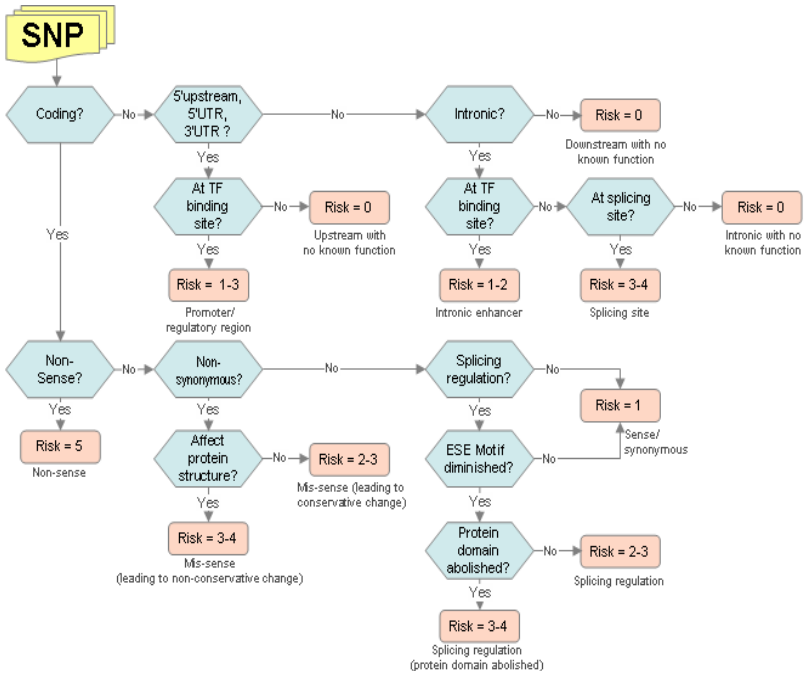
Computational techniques have been used to characterize the polymorphs and predict their involvement in the disease by studying all mutations of MLH1 gene with their variation in individuals (Namboori et al., 2011). Single-nucleotide polymorphisms (SNPs) can be prioritized and classified according to their functional impact based on prediction using bioinformatics and computational tools. The present study analysed the MLH1gene mutations using the tools Sorting intolerant from tolerant (SIFT), PolyPhen-2, functional analysis and selection tool for single nucleotide polymorphisms (FASTSNP) and UTRscan.

MATERIALS AND METHODS

SIFT

SIFT is a sequence homology-based tool that sorts intolerant from

*Corresponding author. E-mail: amithajoy@sahrdaya.ac.in. Tel: 09633899665.



Ranking	Risk Definition
0	No effect
1	Very low
2	Low
3	Medium
4	High
5	Very High

Figure 1. SNP prioritization based on the predicted functional effects.

tolerant amino acid substitutions and predicts whether an amino acid substitution in a protein will have a phenotypic effect (http://sift-dna.org). SIFT is based on the premise that protein evolution is correlated with protein function. Positions important for function should be conserved in an alignment of the protein family, whereas unimportant positions should appear diverse in an alignment (Kumar et al., 2009; Pauline and Henikoff, 2002; Pauline and Henikoff, 2001; Pauline and Henikoff, 2003; Pauline and Henikoff, 2006). The algorithm makes an in-depth search of the protein repositories to find the tolerance of each deviation from the conserved pattern (Namboori et al., 2011). This probability factor helps us to predict the effect of the deviation, that is whether it is deleterious or not. The cutoff value of tolerance has been fixed as 0.05. Hence, if the value is more than or equal to 0.05, the corresponding deviation can be treated as tolerating, while the tolerance value less than 0.05 predicts the change to be harmful.

A library of MLH1 sequences were prepared by providing the NCBI database of SNP by applying appropriate limits like homo-sapiens, chromosome 3, cited in Pubmed, etc. The corresponding rsids of the obtained result were compiled. SIFT analysis of the selected rsids were done using the online software. The SIFT program works on the hypothesis that most of the conserved regions of amino acids are retained in normal protein molecules. The observed changes in these positions may lead to malfunctioning of the protein molecules, and in most cases, deviations are likely to be deleterious (Namboori et al., 2011).

FastSNP server

FastSNP is a web server that allows users to efficiently identify the SNPs most likely to have functional effects (http://fastsnp.ibms.sinica.edu.tw). FastSNP prioritizes SNPs according to 12 phenotypic risks and putative functional effects, such as changes to the transcriptional level, pre-mRNA splicing, protein structure, etc (Hsiang-Yu et al., 2006). The SNP prioritization result is based on the predicted functional effects and their estimated risk proposed by Tabor et al. (2002) as shown in the flowchart (Figure 1).

UTRscan server

UTRscan is a pattern matcher which searches protein or nucleotide (DNA, RNA, tRNA) sequences in order to find UTR motifs (http://itbtools.ba.itb.cnr.it/utrscan). It is able to find, in a given sequence, motifs that characterize 3'UTR and 5'UTR sequences. Such motifs are defined in the UTRSite Database, a collection of functional sequence patterns located in the 5'- or 3'-UTR sequences. The UTRsite entries describe the various regulatory elements present in UTR regions and whose functional role has been established on experimental basis. Each UTRsite entry is constucted on the basis of information reported in the literature and revised by scientists experimentally working on the functional characterization of the relevant UTR regulatory element (UTRdb and UTRsite, 2010; UTRdb and UTRsite, 2005; UTRdb and UTRsite, 2000).

PolyPhen-2

Functional activity of protein was also investigated with the tool PolyPhen-2, which works on structure and multiple alignments with homologous proteins (http://genetics.bwh.harvard.edu/pph2/). The same SNPs were used for this characterization. PolyPhen-2 makes use of a set of features consisting of Dictionary of Secondary Structure in Proteins database to extract secondary structure, “solvent accessible surface area” and “phi–psi” dihedral angle. The calculated parameters include “normed accessible surface area” change in accessible surface propensity resulting from the substitution, change in residue side chain volume, region of the phi–psi map from the Ramachandran plot, and normalized B factor (temperature factor) for the residue (Namboori et al., 2011). With all these parameters, the algorithm computes the “position-specific independent count” (PSIC) score of each SNP. The PSIC score

Table 1. SIFT analysis of the SNPs.

SNP	Amino acid change	Prediction	Score
rs41295282	S93G	Damaging	0.02
rs1295280	G22A	Damaging	0
rs28930073	D132H	Damaging	0
rs11541859	E89Q	Damaging	0.01
rs63750549	G638*	Damaging	0.02
rs63750540	K461*	Tolerated	0.3
rs41295284	L607H	Damaging	0
rs35831931	V716M	Tolerated	0.09
rs35338630	H264D	Tolerated	0.1
rs35045067	Y646C	Damaging	0
rs35001569	K618E	Damaging	0
rs34213726	K443Q	Tolerated	0.53
rs2020873	H718Y	Damaging	0
rs2020872	I32V	Tolerated	0.32
rs1800149	L729V	Tolerated	0.13
rs1799977	I219V	Tolerated	0.27

Table 2. UTRScan analysis of the SNPs

SNP	Number of signal matches	Regulatory element
rs41295284	5	IRES, ADH_DRE, uORF, MBE, PAS.
rs35001569	4	ADH_DRE, uORF, MBE, PAS.
rs35831931	3	IRES, K-BOX, uORF.
rs28930073	3	BRD-BOX, uORF, MBE.
rs2020873	3	IRES, K-BOX, uORF.
rs63750540	2	uORF, MBE.
rs11541859	2	IRES, uORF.
rs1799977	2	uORF, MBE.
rs63750549	1	uORF.
rs41295282	1	uORF.
rs35338630	1	uORF.
rs35045067	1	uORF.
rs2020872	1	uORF.
rs34285587	0	_
rs34213726	0	_
rs1800149	0	_

IRES, internal ribosome entry site (IRES),;ADH_DRE , alcohol dehydrogenase 3'UTR downregulation control element (ADH_DRE); uORF, Upstream Open Reading Frame (uORF); MBE, Musashi binding element (MBE);PAS, polyadenylation Signal (PAS); BRD-BOX, Brd-Box (Brd); K, BOX - K-Box (KB).

differences between the variations due to different SNPs have been calculated. As this difference increases, the possibility of functional impact on the variation increases. If the PSIC score difference is ≥0.9, the variation can be treated as probably damaging.

RESULTS AND DISCUSSION

The computational analysis of MLH1 gene using the tools SIFT, UTRscan, FastSNP and Polyphen2 led to the following conclusions. SIFT analysis predicted 9 out of 16 mutations as damaging as shown in Table 1. The SNPs rs41295284 and rs35001569 with score 0 for both the mutations, correlated with the results of other analysis. In the rsid rs41295284, amino acid leucine was mutated to histidine at 607th position and in rs35001569 amino acid lysine was mutated to glutamic acid at 618th position. Table 2 shows functional significance analysis using UTRscan which predicted the rsid rs41295284 and rs35001569 as damaging since they contained the regulatory elements: internal ribosome entry site, alcohol

Table 3. FastSNP analysis of the SNPs.

SNP	Possible functional effects for the top ranking	Lower risk	Upper risk
rs2020873	Missense (non-conservative)	3	4
rs35338630	splicing site	3	4
rs63750540	Nonsense	5	5
rs41295282	Missense (conservative)	2	3
rs1799977	Missense (conservative); Splicing regulation	2	3
rs1800149	Missense (conservative); Splicing regulation	2	3
rs2020872	Promoter/regulatory region	1	3
rs11541859	Missense (conservative); Splicing regulation	2	3
rs28930073	Missense (conservative); Splicing regulation	2	3
rs34213726	Missense (conservative)	2	3
rs34285587	Missense (conservative); Splicing regulation	2	3
rs35001569	Missense (conservative); Splicing regulation	2	3
rs35045067	Missense (conservative); Splicing regulation	2	3
rs35831931	Missense (conservative); Splicing regulation	2	3
rs41295284	Missense (conservative); Splicing regulation	2	3
rs63750549	Nonsense	5	5

Table 4. Polyphen2 analysis of SNPs.

SNP	Mutation effect	Score
rs2020873	Probably damaging	0.973
rs41295282	benign	0.019
rs28930073	Probably damaging	1
rs41295284	Probably damaging	1
rs35831931	Probably damaging	0.973
rs35045067	Probably damaging	1
rs34213726	benign	0.022
rs2020872	benign	0.012
rs1800149	benign	0.00

dehydrogenase 3'UTR downregulation, Upstream open reading frame (uORF), Musashi binding element (MBE) and polyadenylation Signal (PAS). Various studies have shown that the transcriptional regulation is biologically important and the alteration in the transcriptional components leads to disease.

FastSNP results predicted the rsids rs41295284 and rs35001569 as medium risk ranking ones as shown in Table 3. Further, Table 4 shows the polyphen2 analysis which predicted only nine results among which was the rsid rs41295284 as 'probably damaging one' with score 1. The mutations observed in nsSNP are of three types, missense, nonsense, and frameshift. Besides the coding regions, SNPs may also be found in mRNA untranslated regions (UTRs) and promoter regions, which may affect the gene expression, transcription factor binding, sequence of RNA which is noncoding, and gene splicing (Namboori et al., 2011).

Conclusion

Hence the combined approach using SIFT, UTRscan, FastSNP and PolyPhen2 predicts that the mutation rs41295284 and rs35001569 are the most deleterious among the mutations for MLH1 gene causing colon cancer characterized by the mutation amino acid leucine to histidine. The recognition of these SNPs as deleterious ones provides insight into cancer biology and presents as anticancer therapeutic targets and diagnostic markers.

Since the missense mutations are nucleotide substitutions that change an amino acid in a protein, the deleterious effects of these mutations are commonly attributed to their impact on primary amino acid sequence and protein structure.

REFERENCES

Henry TL, Albert C (1999). Genetic susceptibility to non-polyposis colorectal cancer. J. Med. Genet. 36:801–818.

Hsiang-Yu Y, Jen-Jie C, Wen-Hsien T, Chia-Hung L, Chuan-Kun L, Yi-Jung L, Hui-Hung W, Adam Y, Yuan-Tsong C, Chun-Nan H (2006). FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res. 34:635-641.

Kumar P, Henikoff S, Pauline C (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, Nat. Protoc. 4:8-9.

Kumar P, Henikoff S, Pauline C (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat. Protoc. 4:1073-81.

Pauline C, Henikoff S (2002). Accounting for Human Polymorphisms Predicted to Affect Protein Function. Genome Res. 12:436-446.

Pauline C, Henikoff S (2001). Predicting Deleterious Amino Acid Substitutions. Genome Res. 11:863-874.

Pauline C, Henikoff S (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 131:3812-3814.

Pauline C, Henikoff S (2006). Predicting the Effects of Amino Acid Substitutions on Protein Function. Annu. Rev. Genomics Hum Genet.

7:61-80.
Namboori PK, Vineeth KV, Rohith V, Hassan I, Lekshmi S, Akhila S, Nidheesh M (2011). The ApoE gene of Alzheimer's disease (AD). Funct. Integr. Genomics 11:519–522.
Smith K (2002). Genetic Polymorphism and SNPs Genotyping, Haplotype Assembly Problem Haplotype Map. Functional Genomics and Proteomics
Tabor HK, Risch NJ, Myers RM (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. Nat. Rev. Genet. 3(5):391-397
UTRdb, UTRsite (2010). A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res., 38:75-80.
UTRdb, UTRsite (2005). A collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res., 33:141-146.
UTRdb, UTRsite (2000). Specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Nucleic Acids Res., 28: 193-196.