Biochimica et Biophysica Acta 1862 (2016) 155–165

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

journal homepage: www.elsevier.com/locate/bbadis

Exploring preferred amino acid mutations in cancer genes: Applications to identify potential drug targets

P. Anoosha, R. Sakthivel, M. Michael Gromiha *

Department of Biotechnology, Bhupat and Jyoti Mehta School of BioSciences, Indian Institute of Technology Madras, Chennai 600 036, Tamilnadu, India

ARTICLE INFO

Article history:
Received 17 September 2015
Received in revised form 24 October 2015
Accepted 11 November 2015
Available online 12 November 2015

Keywords:
Cancer
Driver mutation
Passenger
Substitution matrix
Mutation frequency

ABSTRACT

Somatic mutations developed with missense, silent, insertions and deletions have varying effects on the resulting protein and are one of the important reasons for cancer development. In this study, we have systematically analysed the effect of these mutations at protein level in 41 different cancer types from COSMIC database on different perspectives: (i) Preference of residues at the mutant positions, (ii) probability of substitutions, (iii) influence of neighbouring residues in driver and passenger mutations, (iv) distribution of driver and passenger mutations around hotspot site in five typical genes and (v) distribution of silent and missense substitutions. We observed that R→H substitution is dominant in drivers followed by R→Q and R→C whereas E→K has the highest preference in passenger mutations. A set of 17 mutations including R→Y, W→A and V→R are specific to driver mutations and 31 preferred substitutions are observed only in passenger mutations. These frequencies of driver mutations vary across different cancer types and are selective to specific tissues. Further, driver missense mutations are mainly surrounded with silent driver mutations whereas the passenger missense mutations are surrounded with silent passenger mutations. This study reveals the variation of mutations at protein level in different cancer types and their preferences in cancer genes and provides new insights for understanding cancer mutations and drug development.

1. Introduction

Several studies have shown that somatic mutations including missense, silent, nonsense, stoploss, insertions, deletions and larger forms of structural variations etc. play important roles in cancer development [1,2]. Mutations involved in cancer development and progression are named as driver mutations and functionally neutral ones as passenger mutations [3]. Currently, huge amount of cancer related mutation data is being accumulated in biological databases due to the recent advancements in sequencing technologies [4]. Specifically, cancer related clinical information data, genomic characterization data and high level sequence analysis of tumour genomes are organized into the data portal TCGA, The Cancer Gene Atlas (http://cancergenome.nih.gov/), which helps researchers to download and analyse the data. On the other hand, cancer associated mutations data from published scientific literature have been compiled and stored in COSMIC database [5]. It is very important to analyse these data to understand the distribution and nature of mutations in various cancer types.

Several investigations have been carried out to analyse cancer mutations at nucleotide level available in COSMIC database [6] and mutations in TCGA [7]. Recently, COSMIC database has been effectively utilised for understanding the characteristics of several cancer types and to uncover the mutational landscapes, which is a pressing need in cancer research [8–10]. Bozic et al. developed a mathematical model that provides a simple formula for the number of driver mutations as a function of the total number of mutations in the tumour [11] using mutations data from COSMIC database. Machine learning approaches have also been carried out to identify the mutations which contribute to cancer development. In addition, computational tools have been developed to distinguish driver mutations from passenger mutations [12–18], which is a very important task for better treatment of cancer.

In spite of these studies, systematic analysis of somatic protein mutations at a large scale has not yet been explored. In this study, we have systematically analysed the missense, silent, insertion and deletion mutations at protein level from cancer genes involved in various cancer types. We grouped the mutations which are occurring recurrently in the cosmic database as driver mutations and others as passenger mutations and developed mutation matrices for missense mutations based on their frequency of occurrence. Our analysis showed that Arginine is highly mutated in most of the cancer types in driver and passenger substitutions. We have also examined the mutant residue preference and neighbouring residue preferences at different window lengths in both driver and passenger mutations. We identified specific substitutions, such as R→Y, W→A and V→R only in driver missense mutation matrices. The analysis on the distribution of silent mutations around the missense substitutions revealed that driver missense mutations are clustered with silent driver mutations and passenger mutations are

* Corresponding author.
E-mail address: gromiha@iitm.ac.in (M. Michael Gromiha).

http://dx.doi.org/10.1016/j.bbadis.2015.11.006
0925-4439/

surrounded with silent passenger mutations. The implications of the results will be discussed.

2. Materials and methods

2.1. Dataset for analysis

We have constructed a dataset of 761,878 point mutations at protein level observed in 23,751 genes using COSMIC database v65. Among them 577,738 are missense mutations and 184,149 are silent mutations (i.e., 75.8% and 24.2% of the total point mutations, respectively). Generally, the relative frequency of a particular mutation within a gene is used as a criterion for identifying a driver mutation [19]. Hence, these mutants are classified into two groups, driver and passenger based on their occurrence in COSMIC database: if the count is more than one in individual cancer gene, then the mutants are termed as drivers and others are termed as passengers [20]. Therefore, our dataset has 34,167 (5.9%) missense driver, 543,571 (94.1%) missense passenger mutations and 11,163 (6.1%) silent driver, 172,986 (93.9%) silent passenger mutations, respectively. These mutations belong to 41 different cancer types. The datasets used for the present analysis are available at www.iitm.ac.in/bioinfo/Cosmic/.

2.2. Amino acid properties

We have collected a set of physico-chemical, energetic and conformational properties of 20 amino acid residues [21] and their binding propensities for protein–protein, protein–RNA, protein–DNA and protein–carbohydrate complexes [22,23] from the literature. These properties have been successfully used to understand the structure and function of protein and their complexes [24,25]. More details about the property values and their definitions are available at http://www.iitm.ac.in/bioinfo/fold_rate/. We related the amino acid residue preferences in cancer driver and passenger mutations (missense and silent) with these properties to understand the functional consequences of specific properties on cancer mutations.

2.3. Computation of frequency of occurrence of amino acid residues

Frequency of occurrence of each amino acid is calculated for both wild type and mutant residues for all substitutions using the following equation:

$$\text{Freq}\ (i) = n(i)/N \tag{1}$$

where, n(i) is the number of occurrences of residue i and N is the total number of substitutions.

The frequency values of 20 amino acids are normalized with amino acid composition of residues in globular proteins [26] to compute the propensity of amino acid residues for amino acid substitutions. These propensity values are used to relate with amino acid properties and binding propensity of protein–protein, protein–DNA and protein–RNA complexes.

2.4. Preference of amino acid substitutions in driver and passenger mutations

We have computed the preference of amino acid substitutions using the following equation:

$$\text{Mutation}\ (i \rightarrow j) = [n(i \rightarrow j)/N] * 100 \tag{2}$$

where $n(i \rightarrow j)$ is the number of occurrences of substitution $i \rightarrow j$ and N is the total number of substitutions. The computations have been repeated for driver and passenger missense and silent mutations as well as for all the cancer types.

2.5. Preference of neighbouring residues in driver and passenger mutations

We have computed the preferences of amino acid residues on both sides of the mutation position at window lengths of 3, 5 and 7 using specific motifs. For a tripeptide of window length 3, we constructed a motif *M*, where * is any residue and M is the mutant residue in driver or passenger mutations. The preferred tripeptides have been analysed upon the order of preference. Similar procedure has been extended on both sides for 5- and 7-residue segments. A 5 and 7-residue segments have two and three residues on both sides of the mutant, respectively.

2.6. Analysis of statistical significance

We have performed paired t-test (2-tailed, 99% confidence interval) for the driver and passenger mutation frequencies to examine the statistical significance of the results. We have used an online tool, GraphPad software (www.graphpad.com/quickcalcs/ttest1.cfm) for the analysis.

2.7. Distribution of silent and missense mutations

We have analysed the distribution of silent and missense mutations in four typical genes, TP53, APC, PTEN and PIK3CA. We calculated the number of silent around missense, silent around silent, missense around silent and missense around missense mutations with a residue interval of 10 residues such as 1–10, 11–20, 21–30, 31–40, 41–50 and >50 towards N-terminal. We have considered residues only towards N-terminal to avoid redundancy in the analysis.

3. Results and discussion

3.1. Analysis of missense mutations

We have analysed the frequency of occurrence of driver and passenger mutations in different types of cancer and the results are shown in Fig. 1a and b. In these figures, 19 different cancer types, which have more than 100 missense mutations, are included for analysis. We observed the highest frequency of driver mutations in large intestine cancer, which are associated with 6439 proteins. Further, lung and haematopoietic and lymphoid tissue (HLT) cancers have high frequency of missense mutations. On the other hand, 21 cancer types have less than 100 driver mutations and 18 of them have less than 100 passenger mutations. Such cancer types are known to occur very rare with less number of mutations [27]. For example, small-intestine cancer is a very rare cancer type, which has only 14 driver mutations. The proportion of driver and passenger mutations in different cancer types is presented in Table 1. We noticed that the probability of driver mutations is high in liver, soft issue and thyroid cancer. The driver mutations are dominant in few rare cancer types also (gastrointestinal tract, testis, pituitary etc.).

We have computed the frequency of occurrence of all the 20 amino acid residues using Eq. (1) and the results are shown in Fig. 2. We noticed that Arg is highly mutated in majority of the cancer types followed by Ala and Gly. We have normalized the frequency with the composition of amino acid residues [26] and we observed a similar trend with Arg. The high preference of Arg might be due to the fact that it is the most favoured residue for binding with DNA [22,28]. In addition, Arg can form multiple hydrogen bonds [29] and plays an important role as stabilizing element in proteins [30]. Further, we have related the normalized mutation frequencies of 20 amino acid residues with amino acid properties and binding propensities of protein–protein, protein–DNA, protein–RNA and protein–carbohydrate complexes. We noticed that the correlation with any of the amino acid properties is less than 0.40 and the correlation coefficients obtained with binding propensities are shown in Table 2. The highest correlation is observed with the binding propensity of protein–DNA complexes, which is consistent with the hypothesis that majority of the cancers is due to the interruption of

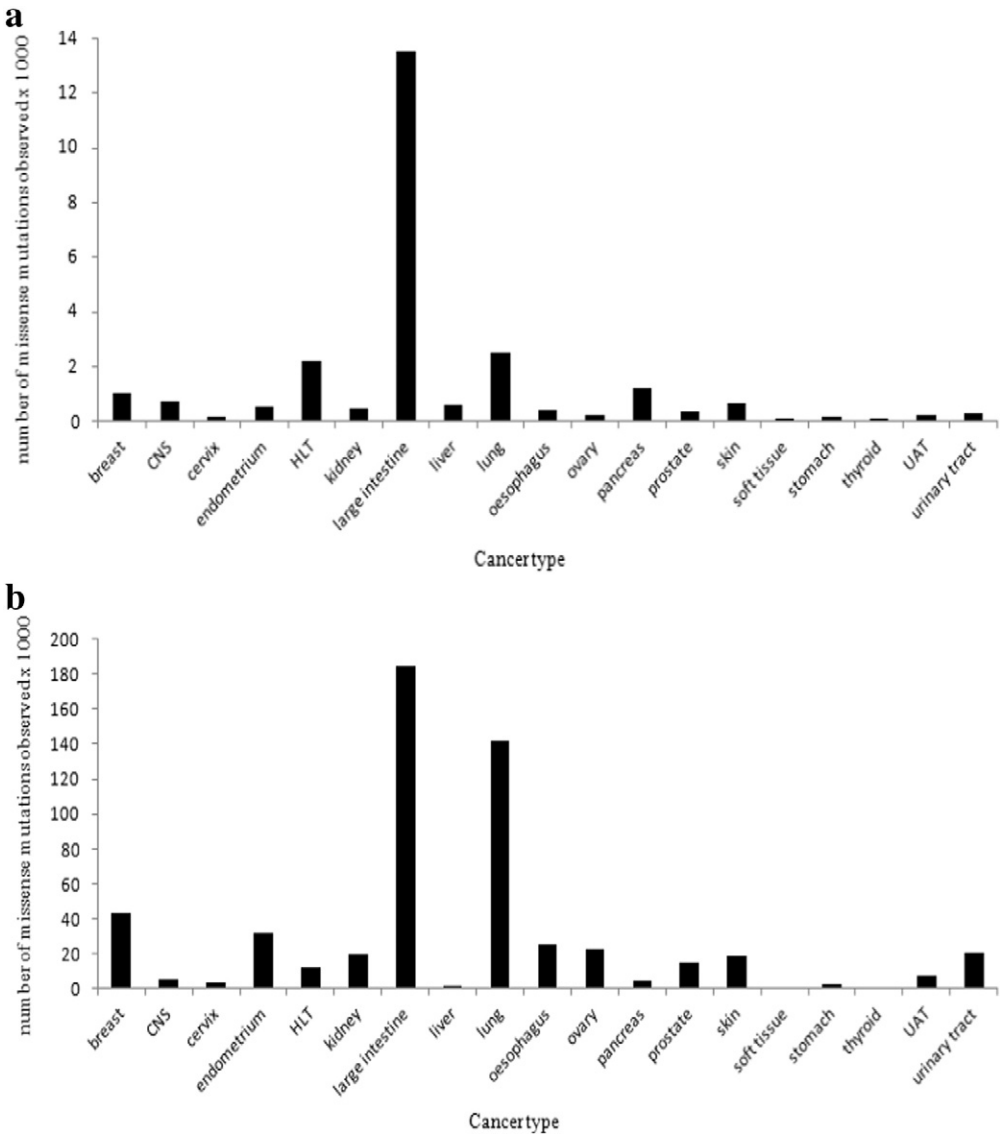

Fig. 1. a: Frequency of missense driver mutation in different cancer types. b: Frequency of missense passenger mutation in different cancer types.

Table 1
Highly mutated genes in various cancer types and the proportion of driver and passenger mutations.

Cancer type	Highly mutated genes	Ratio (%)
Breast	TP53, PIK3CA, PTEN	0.024
Central nervous system	TP53, PTEN, EGFR	0.118
Cervix	OR1S2, ZSCAN5A, MAGEA12	0.046
Endometrium	PIK3CA, PTEN, TP53	0.017
Haematopoietic and lymphoid tissue	TP53, TET2, ABL1	0.187
Kidney	VHL, MUC4, TP53	0.023
Large intestine	TP53, PIK3CA, APC	0.074
Liver	TP53, CTNNB1, HNF1A	0.357
Lung	TP53, EGFR, KRAS	0.018
Oesophagus	TP53, KRAS, NFE2L2	0.018
Ovary	TP53, PIK3CA, CTNNB1	0.010
Pancreas	TP53, CTNNB1, KRAS	0.280
Prostate	TP53, SPOP, KRAS	0.022
Skin	TP53, BRAF, KIT	0.035
Soft tissue	KIT, TP53, PDGFRA	0.356
Stomach	TP53, CTNNB1, KRAS	0.051
Thyroid	TSHR, RET, KRAS	0.255
Upper aerodigestive tract	TP53, CDKN2A, HRAS	0.035
Urinary tract	TP53, FGFR3, HRAS	0.014

Ratio = number of driver mutations/number of passenger mutations.

interactions between protein and DNA along with malfunctions in DNA repair mechanism, signal transduction etc. Specifically the well known DNA binding protein, TP53 protein is causing several types of cancers due to the mutation at the DNA binding domain.

We have further analysed the influence of Arg in different cancer types. Fig. 3 shows the mutation frequency of Arg in different cancer types in both passenger and driver mutations. Although the difference between the residue preferences in driver and passenger mutations is marginal, the preference is higher in driver mutations than passenger mutations in most of the cancer types. In cervix, oesophagus, soft tissue and thyroid cancers the preference of Arg is higher in passenger mutations than driver mutations.

On the other hand, silent driver mutations are dominated by Ser followed by Pro, Thr, Leu and Ala, whereas silent passenger mutations are dominated with Leu (Fig. 2). Met and Trp mutations are not observed in silent mutations due to the lack of synonymous codons [31].

3.2. Highly mutated genes in different cancer types

We have analysed highly mutated genes in various cancer types and the results are shown in Table 1. The most commonly mutated gene in

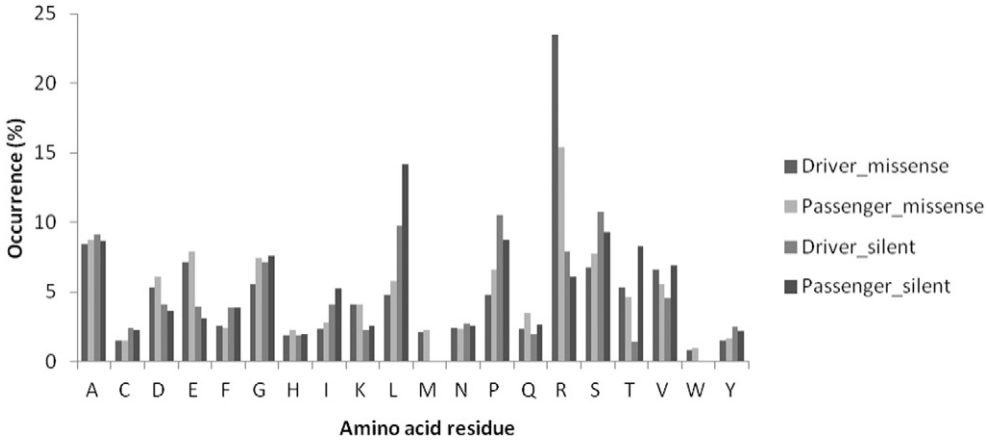

Fig. 2. Percentage of occurrence of 20 amino acids as wild type residue in driver and passenger missense and silent mutations, respectively.

Table 2
Correlation of amino acid binding propensities and cancer mutation frequencies.

Mutation type	Mutation class	P–C	P–P	P–RNA	P–DNA
Missense mutations	Driver	−0.09	0.21	0.50	0.58
	Passenger	−0.13	0.11	0.46	0.49
Insertions	Driver	−0.16	−0.16	0.02	−0.06
	Passenger	0.04	0.10	0.20	0.09
Deletions	Driver	−0.20	−0.21	−0.03	−0.01
	Passenger	−0.39	−0.28	−0.07	−0.06

P–C: protein–carbohydrate; P–P: protein–protein: P–RNA: protein–RNA; P–DNA: protein–DNA.

many cancer types is TP53 and it plays potential roles in most of the cancer types as well as a target for cancer treatment [32,33]. Further, OR1S2, PIK3CA, VHL, EGFR, KIT and TSHR showed many mutants and these genes have been studied well to investigate the mechanism and treatment of cancer [34–36]. A very recent analysis of somatic mutations in exome sequences has identified 33 genes that were not previously known to be significantly mutated in cancer, including genes related to proliferation, apoptosis, genome stability, chromatin regulation, immune evasion, RNA processing and protein homeostasis [37]. Oncogenes such as RHEB, RHOA, SOS1, ELF3, SGK1 and MYOCD are involved in several cancer types, among which notably, RHEB and RHOA encode small GTPases, where recurrent mutations affect their function. Another gene known as ARHGAP35 (previously called GRLF1), which plays a role in apoptosis resides in a small genomic region that is focally deleted in many tumours. These genes are previously not well known in cancer progression.

3.3. Construction of driver and passenger mutation matrices

We have computed the frequency of occurrence of missense mutations in different cancer genes using Eq. (2) and developed 20 × 20 amino acid mutation matrices for driver and passenger mutations. The missense mutation matrices for driver and passenger mutations are presented in Table 3a and b, respectively. Each value in the matrix corresponds to the percentage of a particular missense mutation observed in the available data in COSMIC database. We have assessed the statistical significance of driver and passenger mutation frequencies, using t-test. The relationship between the number of topmost differing mutation types and p-values is shown in Fig. 4. The p-value is calculated for the topmost mutation types in both driver and passenger matrices by adding one substitution in each iteration and we noticed that the topmost 47 substitution types have the p-value of less than 0.01, which are considered for further analysis.

We observed that R→H substitution occurs with the highest frequency in driver mutations followed by R→C, R→Q and in passenger mutations the frequently occurring mutation is E→K substitution. It is also noted that the matrices are dominated with specific mutations and there is no mutation for several combinations, for example, A→C, A→H in driver mutations and D→K in passenger mutations. Among 380 possible missense mutations, 164 substitutions are not observed in both driver and passenger mutation matrices, 31 are observed only in passenger mutations and 17 substitutions are observed only in driver mutations (Table 4). This analysis suggests that the unique substitutions observed in driver mutation matrices may have significant role in cancer development and progression.

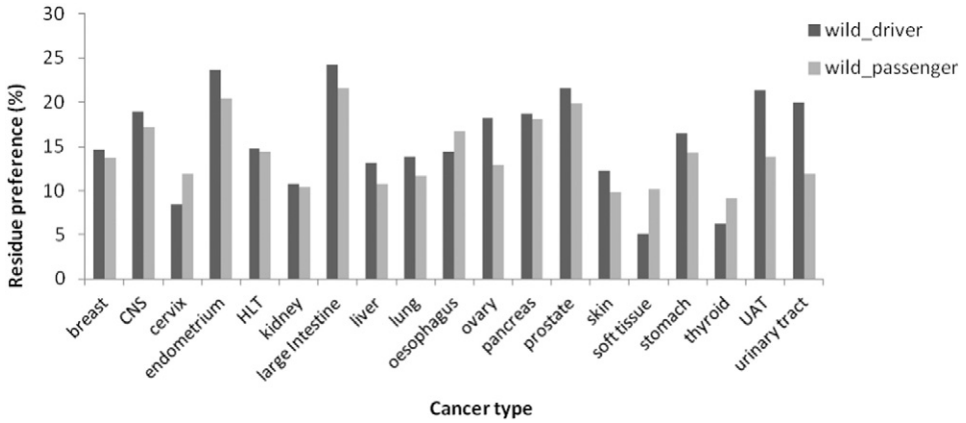

Fig. 3. Frequency of Arginine mutation in different cancer types.

Table 3
Frequency of occurrence of missense mutations.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
(a) Driver mutations																				
A	–	0	0.31	0.187	0.003	0.167	0	0.003	0	0.003	0	0	0.237	0	0	0.544	**3.717**	**3.316**	0	0
C	0.003	–	0	0	0.252	0.12	0	0	0	0	0	0	0	0	0.442	0.184	0	0	0.105	0.416
D	0.1	0	–	0.366	0.006	**1.455**	0.263	0.006	0.003	0	0	**2.11**	0	0	0	0	0	0.196	0	0.811
E	0.138	0	**1.06**	–	0	**1.835**	0.003	0	**3.433**	0	0	0	0	0.462	0	0	0	0.237	0	0
F	0.003	0.228	0	0	–	0	0	0.091	0.003	**1.586**	0	0	0	0	0	0.38	0	0.22	0	0.07
G	0.225	0.518	0.738	0.597	0.009	–	0	0.006	0	0.003	0	0.012	0.003	0	**1.405**	**1.03**	0	0.764	0.263	0.003
H	0	0	0.07	0	0	0	–	0	0	0.12	0	0.225	0.114	0.225	0.679	0	0	0	0	0.477
I	0	0	0.003	0	0.149	0	0.003	–	0.041	0.132	0.246	0.155	0	0	0.018	0.167	0.761	0.664	0	0
K	0	0	0	0.843	0.003	0	0	0.05	–	0.003	0.094	**1.34**	0	0.161	**1.08**	0	0.541	0	0	0
L	0.003	0	0	0	0.84	0	0.108	0.588	0.003	–	0.304	0	**1.613**	0.19	0.389	0.193	0	0.498	0.029	0
M	0	0	0	0	0	0	0	0.937	0.059	0.132	–	0	0.003	0	0.032	0	0.378	0.585	0	0
N	0	0	0.477	0	0	0	0.149	0.158	0.38	0	0	–	0	0	0	0.948	0.149	0	0	0.123
P	0.208	0	0	0	0.012	0	0.41	0	0.009	**1.902**	0	0.003	–	0.263	0.19	**1.264**	0.524	0	0	0
Q	0	0	0	0.205	0	0	0.644	0	0.34	0.158	0	0	0.149	–	0.86	0	0	0	0	0
R	0	**5.002**	0	0	0	0.922	**5.696**	0.644	0.337	0.948	0.225	0	0.31	**4.99**	–	0.536	0.129	0	**3.755**	0.006
S	0.094	0.342	0	0	0.641	0.702	0	0.211	0.003	**1.583**	0	0.489	**1.34**	0	0.524	–	0.214	0	0.023	0.615
T	**1.522**	0	0	0.003	0	0	0	0.588	0.135	0	**2.221**	0.205	0.278	0	0.082	0.284	–	0	0	0
V	**1.548**	0	0.12	0.14	0.222	0.345	0	**2.154**	0.003	0.5	**1.607**	0	0	0.003	0.003	0	0	–	0	0
W	0.003	0.234	0	0	0.003	0.059	0	0	0.003	0.17	0	0	0	0	0.293	0.059	0	0	–	0
Y	0	0.685	0.126	0	0.126	0	0.316	0	0	0	0	0.138	0	0	0	0.117	0	0	0	–
(b): Passenger mutations																				
A	–	0	0.566	0.329	0.003	0.32	0	0.002	0	0.002	0	0.001	0.351	0	0.001	**1.24**	**3.053**	**2.874**	0	0
C	0	–	0	0	0.403	0.08	0	0	0	0.001	0	0	0	0	0.229	0.268	0	0	0.089	0.482
D	0.141	0.001	–	0.613	0.002	0.705	0.789	0	0	0	0	**2.166**	0	0	0	0.001	0	0.293	0	**1.36**
E	0.232	0	**1.676**	–	0	0.661	0	0.002	**3.582**	0.006	0.001	0	0	**1.434**	0.001	0.001	0.001	0.339	0	0
F	0	0.299	0	0	–	0	0	0.17	0	**1.293**	0	0	0.001	0	0	0.257	0	0.325	0	0.117
G	0.478	0.847	0.902	0.913	0.009	–	0	0.003	0.013	0.019	0.001	0.004	0.001	0	**1.455**	0.936	0.001	**1.386**	0.481	0.001
H	0	0	0.16	0	0	0	–	0	0	0.164	0	0.413	0.109	0.36	0.418	0	0.001	0	0	0.667
I	0	0	0	0	0.236	0	0	–	0.049	0.216	0.655	0.208	0	0	0.026	0.157	0.566	0.717	0	0
K	0.001	0	0	0.531	0	0	0	0.1	–	0.001	0.2	**1.757**	0	0.292	0.577	0	0.652	0	0.001	0
L	0	0	0	0	**1.274**	0.001	0.169	0.969	0.001	–	0.669	0	0.835	0.244	0.38	0.183	0	**1.034**	0.039	0
M	0	0	0	0	0	0	0	**1.174**	0.095	0.194	–	0.003	0	0	0.061	0	0.317	0.4	0	0
N	0	0	0.329	0	0.001	0	0.226	0.216	0.55	0	0	–	0	0	0	0.656	0.218	0	0	0.16
P	0.425	0	0	0	0.015	0	0.804	0.002	0.006	**1.955**	0	0.006	–	0.436	0.347	**1.733**	0.919	0.001	0	0
Q	0	0	0	0.527	0	0	**1.257**	0	0.689	0.323	0	0	0.18	–	0.53	0	0	0	0	0
R	0	**2.73**	0	0	0.002	0.606	**3.019**	0.629	0.61	**1.074**	0.322	0.002	0.317	**2.877**	–	0.747	0.417	0	**2.069**	0
S	0.171	0.883	0	0	**1.185**	0.358	0	0.481	0.001	**1.354**	0	0.696	0.574	0	0.674	–	0.435	0	0.054	0.907
T	0.911	0	0	0	0	0	0	0.85	0.307	0.001	**1.212**	0.411	0.255	0	0.152	0.532	–	0	0	0
V	0.895	0	0.121	0.164	0.447	0.257	0.001	**1.365**	0.001	**1.041**	**1.229**	0	0	0	0	0.001	0	–	0	0
W	0	0.395	0	0	0.007	0.037	0	0	0	0.264	0.001	0	0	0	0.21	0.054	0	0	–	0.001
Y	0	0.708	0.088	0	0.2	0	0.429	0	0	0.001	0	0.153	0	0	0	0.088	0	0	0	–

Mutations, which have the frequency of more than 1 are shown in bold.

We have also compared the frequency of driver mutations with well known PAM250 [38] and BLOSUM62 [39] matrices, which are mostly used in protein sequence alignment. Interestingly, most of the substitutions are not similar and are not likely to be present over a long period of time. This analysis supports our findings that the frequently occurring driver mutations are specific to cancer. We have checked the correspondence of driver and passenger mutation frequencies with the amino acid replaceability matrix, which is derived from neighbourhood selectivity in sequences [40]. The authors showed a good correlation between the amino acid replaceability matrix and naturally accepted point mutations. Interestingly, we found that most of the topmost passenger substitutions (e.g. E→K, V→L and E→Q) identified in the current study are similar to those in that matrix and this observation supports that passenger mutations are neutral compared to driver mutations.

Further, we have correlated driver and passenger mutation matrix frequencies presented in Table 3a and b, with 87 available mutation matrices in AA Index database [41] to identify similar substitution matrices published in the literature. The highest correlation ($r = 0.48$) was obtained between driver matrix mutation frequencies and Genetic code matrix [42]. This analysis shows that few substitutions are likely to be preferred more than others due to the requirement of very less changes in their codons to be replaced by another coding for different amino acid. Further, we have examined the occurrence of preferred unique driver and passenger mutations (Table 3) with topmost mutants in 87 substitution matrices. We observed that matrices for alignment score for distant homologs and structural alignments using entropy have 87.5% of topmost driver mutants. On the other hand, these matrices also contain 46% and 64% of topmost passenger mutants, respectively. Hence, available mutation matrices alone are not sufficient to differentiate drivers from passenger mutations and the developed matrices

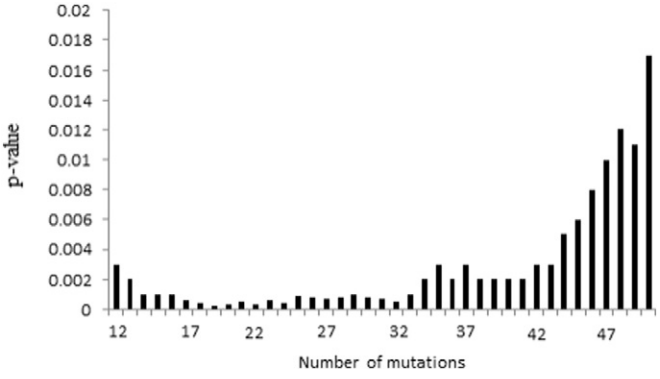

Fig. 4. The relationship between the number of topmost differing mutation types and p-values.

Table 4
Unique driver and passenger mutations.

	Amino acid substitutions
Driver mutations	C→A, D→I, D→K, E→H, F→A, F→K, I→D, I→H, K→F, L→A, M→P, R→Y, T→E, V→Q, V→R, W→A, W→K
Passenger mutations	A→N, A→R, C→L, D→C, D→S, E→I, E→L, E→M, E→R, E→S, E→T, F→P, G→K, G→M, G→T, H→T, K→A, K→W, L→G, M→N, N→F, P→I, P→V, R→F, R→N, T→L, V→H, V→S, W→M, W→Y, Y→L

Table 5
Topmost ten substitutions in highly mutated genes.

TP53	PTEN	EGFR	APC	KIT	PIK3CA	VHL	CTNNB1	BRAF	KRAS
R→P	**E→G**	L→P	**E→G**	N→S	**E→G**	L→P	A→T	**G→R**	**G→D**
P→L	D→G	E→K	S→P	E→K	**G→D**	V→D	S→F	S→P	G→S
P→S	Y→C	G→S	S→L	D→G	**K→R**	L→R	S→P	G→E	G→V
V→G	L→P	A→T	Q→R	**P→L**	**A→V**	R→P	**E→G**	G→V	A→T
R→L	**K→R**	R→C	S→G	V→A	D→N	**G→D**	G→V	**A→V**	G→N
R→H	F→L	**P→L**	**P→L**	D→N	H→R	**E→G**	S→C	**G→D**	G→C
C→F	G→V	V→M	E→V	V→I	N→S	**G→R**	E→K	I→T	**G→R**
R→G	A→T	V→A	T→A	**K→R**	**P→L**	D→G	**A→V**	S→F	A→P
K→R	H→R	**A→V**	K→E	F→L	**G→R**	E→K	**G→D**	Q→R	**A→V**
C→Y	**G→R**	**E→G**	P→H	**E→G**	S→P	V→L	T→A	**K→R**	Q→R

Mutations which are present in majority of the genes are shown in bold.

(Table 3) might provide insights for better understanding of cancer related mutations.

3.4. Analysis of topmost ten substitutions in highly mutated genes

We have analysed the occurrence of topmost ten driver substitutions in each of the highly mutated genes viz. TP53, PTEN, EGFR, APC, KIT, PIK3CA, VHL, CTNNB1, BRAF and KRAS and the results are presented in Table 5. We observed that, most of these substitutions such as E→G, K→R, P→L, A→V, G→R and G→D present in majority of the genes have high frequencies in driver mutation matrix (Table 3a). Thus, it is evident that the constructed mutation matrix is not influenced by individual genes and could serve as a useful tool for analysing various substitutions as potential drivers. However, we also noticed that few of the topmost ten substitutions are unique to individual genes (e.g. I→T in BRAF, and C→Y in TP53) but found to have low frequency in driver mutation matrix (Table 3a). Hence, availability of significant amount of mutation data for individual genes could aid in the development of gene specific analysis [43].

3.5. Cancer-specific driver mutations

It has been reported that the frequency of missense mutation varies across individual cancer types [44–46]. To gain deep insights on the frequency of mutations in individual cancer types, we have considered the topmost 10 driver missense mutations (Table 3a) for which the overall frequency of occurrence in all cancer types together is >2 and are statistically significant (Fig. 4). The results showed that these mutations are tissue specific as seen in Table 6a. Although R→H has the highest overall frequency of occurrence but it is specific to only few cancer types such as endometrium, large intestine, stomach and pancreas. Similarly, E→K is specific only to skin and urinary tract cancer, and its frequency of occurrence in other cancer types is low. Further, analysis on individual cancer type showed that 9 out of topmost 10 mutations from driver matrix in lung and liver cancers have frequencies less than 2. On the other hand, in endometrium and large intestine cancer types, 70% of the topmost substitutions have frequency greater than 2. We also noticed that there is no mutation from the list of the topmost 10 preferred ones in few specific cancer types (e.g. R→Q, A→V, T→M and V→I in soft tissue, and A→T in ovary and thyroid). These observations reveal that the preferred mutations are specific to cancer site and are referred as "cancer specific drivers".

Further, we have evaluated the preference of amino acid substitutions in all the considered cancer types and the topmost 10 missense mutations in each cancer type are presented in Table 6b. We observed that highly frequent R→H substitution is prevalent only in few cancer types such as large intestine, stomach, pancreas, oesophagus etc. There are noticeable differences in the occurrence of substitutions in different cancer types. For example, in lung and ovary cancer, G→V substitutions are prevalent whereas in few of the other cancer types such as thyroid, soft tissue and urinary aerodigestive tract (UAT), G→D substitutions are dominant. We also noticed that D→E is highly occurring in liver cancer with a frequency of 3.5 whereas its overall frequency of occurrence in all cancer types is 0.37. Similar trends are observed for K→T and E→D in oesophagus, S→N in cervix, A→D in prostate etc. We infer from these results that the driver mutations are specific to cancer site which strongly agree with the previous reports [45].

3.6. Preference of neighbouring residues in driver and passenger mutations

Previous studies have shown that there is a range of non-random pairing of residues in the neighbouring positions in primary structure of proteins and it is different for each of the amino acids [47]. This might play a very important role in short range interactions, which are

Table 6a
Frequency of topmost 10 preferred missense mutations in different cancer types[a].

Cancer type	R→H	R→C	R→Q	R→W	A→T	E→K	A→V	T→M	V→I	D→N
Breast	2.60	2.31	2.79	1.92	1.54	3.85	1.54	1.15	1.25	1.35
CNS	4.40	2.98	3.12	1.42	2.98	3.12	**3.97**	1.13	1.84	1.99
Cervix	1.93	0.65	1.29	0.65	**3.87**	2.58	1.29	0.65	1.94	1.94
Endometrium	**5.73**	3.58	**6.26**	**5.90**	2.50	2.68	2.86	0.89	0.36	1.07
HLT	3.11	3.11	2.34	1.76	2.70	2.21	2.39	1.17	2.21	1.44
Kidney	1.90	1.27	2.33	2.11	1.90	1.69	2.11	0.21	0.63	1.27
Large intestine	5.45	**4.75**	5.71	3.81	3.70	2.66	3.51	1.52	1.59	1.77
Liver	2.08	1.21	1.04	1.73	1.56	1.39	0.52	0.52	0.52	1.73
Lung	1.63	1.31	1.03	0.99	1.31	2.90	1.23	0.48	0.75	1.39
Oesophagus	4.67	2.44	2.22	2.44	1.33	3.33	1.11	0.89	1.33	1.33
Ovary	2.12	1.27	2.97	2.97	0	2.54	0.42	0.85	0.85	0.42
Pancreas	5.39	4.48	3.32	3.15	2.90	3.40	2.82	**2.82**	**2.82**	1.66
Prostate	5.47	1.22	4.56	3.95	2.13	3.04	1.82	0.91	2.43	1.22
Skin	0.91	4.42	2.29	1.98	1.37	**8.69**	1.22	0.31	0.46	**3.20**
Soft tissue	2.21	1.47	0	0.74	0.74	2.21	0	0	0	0.74
Stomach	5.52	1.38	2.07	2.76	2.07	2.07	1.38	0	0	2.07
Thyroid	0.78	1.55	2.33	0.55	0	4.65	3.88	0	1.55	0
UAT	3.44	2.67	2.29	1.53	0.38	3.44	0.38	0.38	0.38	1.15
Urinary tract	2.80	2.10	2.80	3.15	0.35	7.34	3.15	0.35	0.35	1.75

[a] The highest frequency in each cancer type is shown in bold.

Table 6b
Topmost 10 missense mutations in different cancer types[a].

Cancer type	Missense mutation									
Breast	R→W	V→M	P→L	E→K	R→Q	R→H	K→N	R→C	P→S	G→R
CNS	R→H	A→V	G→R	R→Q	E→K	A→T	R→C	P→L	G→S	P→S
Cervix	N→S	A→T	F→L	**S→N**	I→T	E→K	**T→I**	G→S	S→G	G→D
Endometrium	R→Q	R→W	R→H	G→V	R→C	A→V	E→K	A→T	P→L	Y→C
HLT	R→H	R→C	A→T	A→V	R→Q	E→K	V→I	P→L	L→P	V→M
Kidney	L→P	R→Q	S→R	R→W	**L→F**	A→V	G→V	A→T	R→H	P→S
Large intestine	R→Q	R→H	R→C	R→W	**E→G**	A→T	A→V	V→A	L→P	T→A
Liver	**D→E**	Y→C	P→L	R→L	R→S	R→H	G→V	G→R	R→W	C→F
Lung	G→V	E→K	R→L	V→L	D→Y	**G→C**	M→I	P→T	**A→S**	R→H
Oesophagus	R→H	**K→T**	**E→D**	L→R	S→R	V→M	E→K	**C→F**	H→R	R→W
Ovary	G→V	Y→C	R→Q	R→W	R→G	C→Y	R→P	S→F	E→K	D→V
Pancreas	R→H	R→C	E→K	R→Q	R→W	A→T	T→M	V→I	A→V	P→L
Prostate	R→H	R→Q	R→W	**A→D**	E→K	G→S	R→L	V→I	A→T	H→R
Skin	E→K	P→S	S→F	P→L	R→C	G→E	D→N	G→R	R→Q	H→Y
Soft tissue	G→D	G→S	Y→C	G→R	D→Y	N→K	G→V	**Q→H**	**V→D**	E→K
Stomach	R→H	G→V	G→D	G→S	V→M	R→W	Y→C	P→S	A→T	**P→R**
Thyroid	G→D	E→K	G→C	G→R	C→R	T→I	A→V	G→S	Q→R	G→V
UAT	G→D	E→K	R→H	P→L	Y→C	R→L	R→G	V→L	G→S	H→L
Urinary tract	E→K	R→G	H→R	R→W	Y→C	A→V	R→Q	R→H	**K→E**	G→R

[a] Unique mutations are shown in bold.

essential for protein stability and function. Hence, we have examined the preference of neighbouring residues (N- and C-termini) of mutant positions at window lengths 3, 5 and 7. Interestingly, neighbouring residue preferences vary with the type of cancer types and associated proteins. Overall analysis showed that aromatic residues are less preferred in the neighbouring positions of mutant residue whereas Gly and Ser are highly preferred in N and C termini of driver mutant positions and passenger mutant positions, respectively.

Mutation at a single amino acid might cause conformational changes in the neighbouring residues and a tripeptide is known to be the smallest unit that captures the bending of the main chain of a protein [48]. In our current analysis, we observed that two tripeptides are dominantly present on either side of the driver mutation site in nearly 40% (16 of 41) of cancer types. In all those cases the mutated amino acid is flanked by AVG and GVG in N-terminal and C-terminal sides, respectively. Interestingly, both the tripeptides are constituted of hydrophobic amino acids. One possible reason is that the mutation might cause disruption of hydrophobic interactions mediated by these peptides by changing the conformation and consequently alter the normal function of the protein.

3.7. Distribution of driver and passenger mutations near hotspot mutation

We have collected hotspot mutations data from previous experimental reports [49–53], for topmost five highly mutated genes in all cancer types. Further, we have examined the distribution of driver and passenger mutations around the hotspot mutation in respective genes. The list of hotspot mutations from the five typical proteins is presented in Table 7. We have computed the number of driver and passenger mutations at different window lengths (3–13) from hotspot site in the protein sequence. Interestingly, we observed that driver mutations are highly accumulated in the hotspot mutation region, whereas the number of passenger mutations is comparatively low. This analysis is strongly supported by the previous reports [54], suggesting that the frequency of mutations around hotspot might provide a clue for identifying functional mutants in cancer progression. Fig. 5 demonstrates the distribution of driver and passenger mutations around hotspot in five typical genes.

Table 7
List of hotspot mutants in topmost five highly mutated genes.

Gene	Hotspot mutations
TP53	R175H, R273H, R248W
PIK3CA	E542K, E545K, H1047R
CTNNB1	S37A
KRAS	G12S
KIT	D816H

3.8. Insertions and deletions

Dataset of cancer associated mutations includes only in-frame indels with 2647 deletions among which 426 are drivers and 2221 are passenger mutations and 1078 insertions with 270 drivers and 808 passengers. The occurrence of deletions is higher than insertions and single residue indels are more frequent. The genes EGFR and KIT have high number of driver deletions and driver insertions respectively. Fig. 6 shows the frequency of each amino acid in deletions and insertions, respectively. Analysis showed that glutamic acid is highly preferred in both driver deletions and insertions. We correlated amino acid frequencies in insertions and deletions with amino acid binding propensities and the results are given in Table 2.

3.9. Distribution of silent and missense mutations

Recent studies suggested that silent mutations play significant roles in human cancers and can frequently act as drivers in cancer progression [55,56]. Although these mutations do not alter the protein sequence they may affect other properties such as speed or accuracy of mRNA translation, mRNA folding, splicing mechanisms [57–59], etc. and lead to cancer through diverse mechanisms.

We have analysed the distribution of silent and missense mutations in cancer causing genes at protein level and the results for four typical genes (TP53, PIK3CA, PTEN and APC) are presented in Fig. 7. We noticed that the missense mutations clustered together in a linear protein sequence as reported in the literature [60,61], In addition, the distribution of silent mutations is mostly mixed with missense mutations than silent mutations alone. In TP53 gene, we observed a distinct pattern of distribution in different ranges of amino acid positions (1–10, 11–20, 21–30 and 41–50 residues from the mutant), in which missense mutations are highly distributed around missense mutations than silent mutations (Fig. 7a). On the other hand, silent mutations are preferred to mix with the missense mutations than silent mutations themselves. This might be due to the presence of more than 80% silent and missense mutations in the protein. In the case of APC and PIK3CA, this distribution is unclear as the protein sequence length is very high compared to the number of mutations. However, we observed that missense around missense mutations is high and distribution of silent mutations around missense is

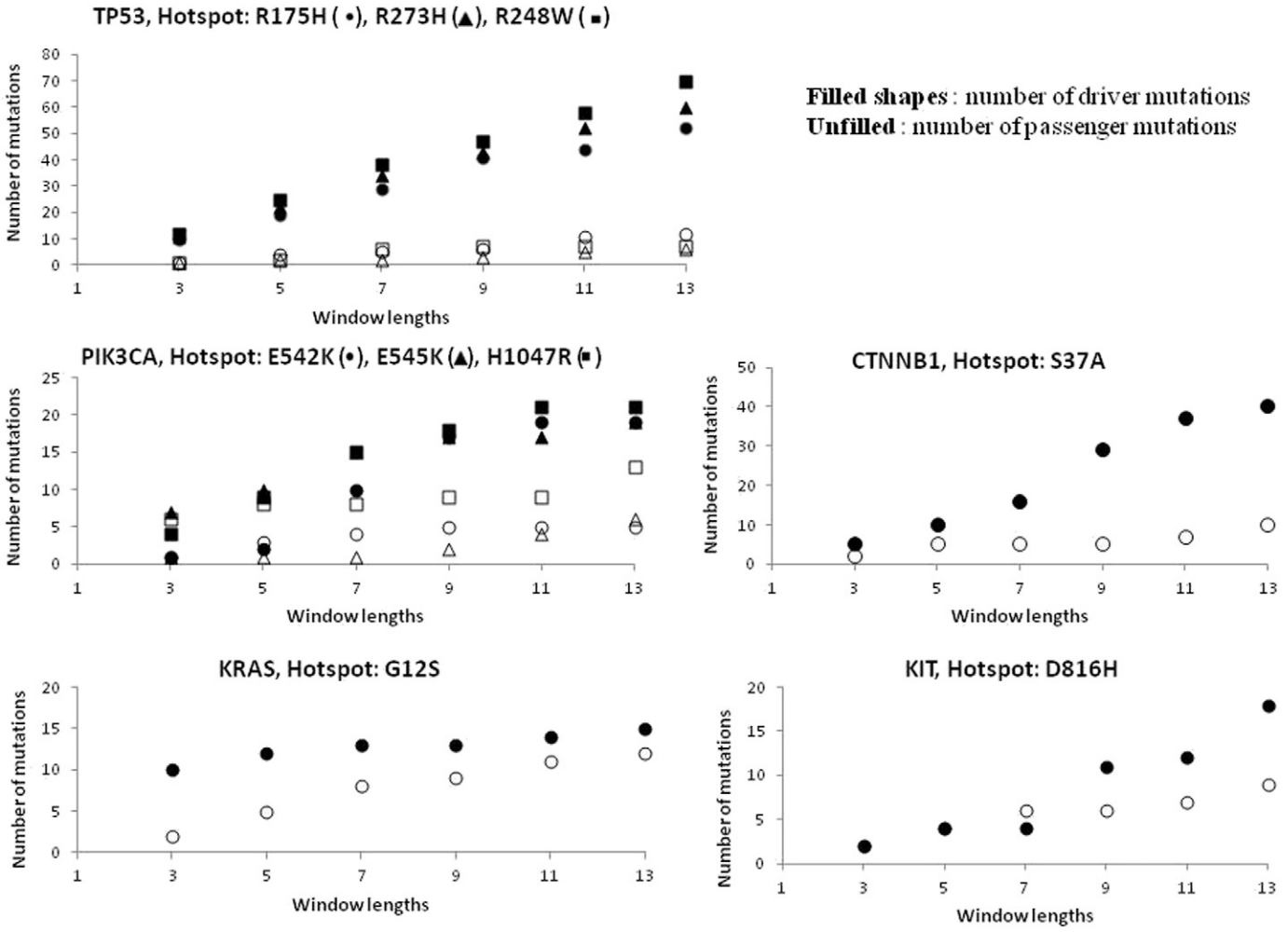

Fig. 5. Distribution of driver and passenger mutants around hotspot mutation site.

more than the cluster of silent mutations. In PTEN and PIK3CA genes, the number of missense around missense mutations is high near the mutant position and the trend is moving downwards at distant residues (say 30–50 in PIK3CA). This analysis reveals that although the protein length and number of mutations are different in all these four genes, the missense mutations are clustered together and silent mutations are mostly clustered with missense mutations in the sequence.

Further, we analysed the distribution of silent mutations around missense mutations at different position ranges (1 to 50 in steps of 10) in the sequence. Interestingly we observed that the driver missense mutations are mostly surrounded by driver silent mutations and passenger missense mutations are surrounded by passenger silent mutations. Further, we examined the distribution of silent around missense mutations for known driver and passenger mutations in top three

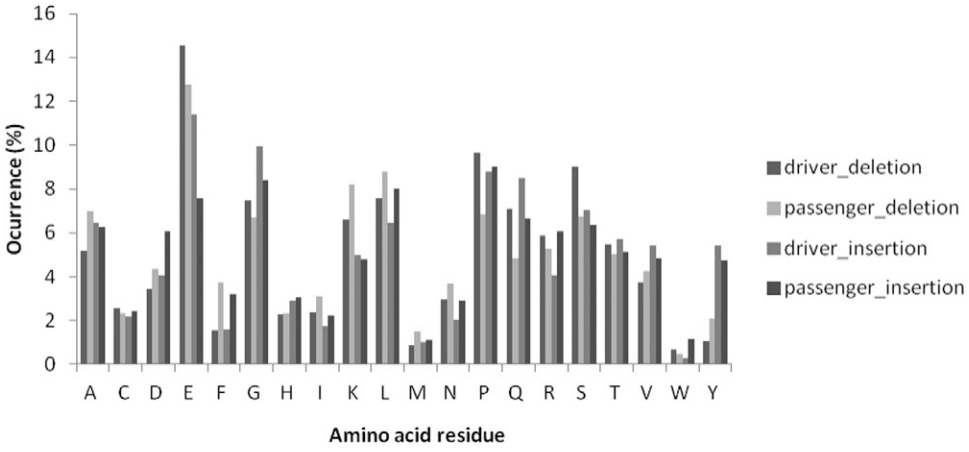

Fig. 6. Frequency of occurrence of 20 amino acids in insertions and deletions.

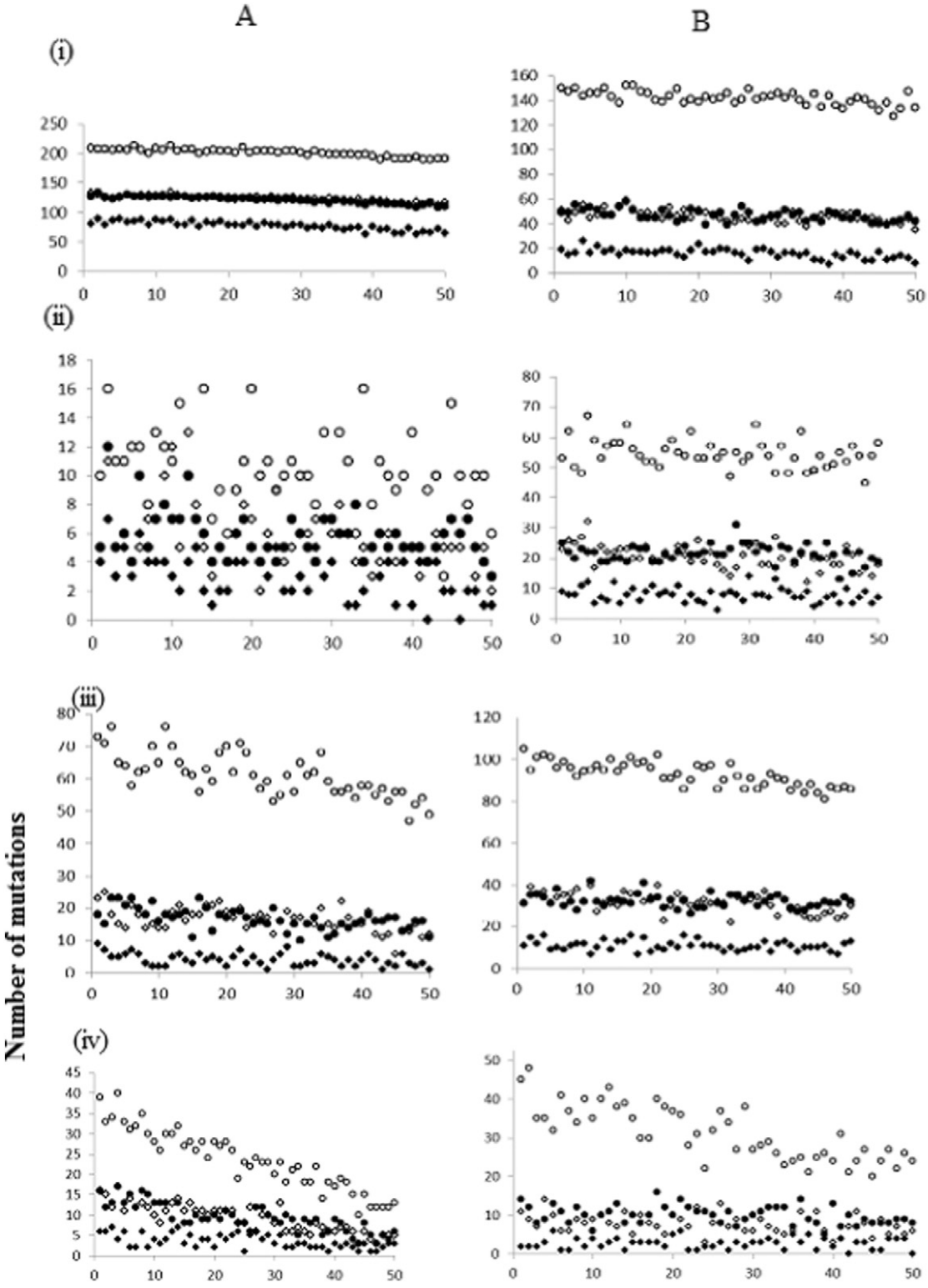

Fig. 7. Distribution of silent and missense mutations in different genes ●: silent around missense, ○: missense around missense, ◇: missense around silent, and ♦: silent around silent mutations. (A) Driver and (B) Passenger. (i) TP53, (ii) APC, (iii) PTEN and (iv) PIK3CA.

genes reported in the literature [62,63] and the results are presented in Table 8. We found that the distribution of missense around missense or silent around silent mutations is distinct at the sequence position range 1–10. In driver mutations, the occurrence of silent driver and silent passenger around missense driver mutations is 89% and 12% respectively. On the other hand, for the passenger mutations, the occurrence is 14% and 86% respectively. Further in TP53, it is observed that driver missense mutations are surrounded with driver silent mutations in almost all position ranges. For example, R175H is a well studied hotspot mutation in TP53 (50), which have 7 driver silent mutations in the range of 1–10 from position 175 in the sequence towards N-terminal (residues 165–175). 7 out of 10 possible positions around R175 are occupied

Table 8
Distribution of silent mutations around missense mutations.

Mutation	Gene	Mutation type	1–10 Silent_d/Silent_p	11–20 Silent_d/Silent_p	21–30 Silent_d/Silent_p	31–40 Silent_d/Silent_p
R175H	TP53	Driver	7/0	9/0	7/1	9/1
R248Q	TP53	Driver	6/2	4/3	7/2	4/3
R273H	TP53	Driver	8/0	8/2	8/0	5/2
R249S	TP53	Driver	7/1	3/4	8/1	4/4
R282W	TP53	Driver	8/2	7/5	9/3	8/0
H1047L	PIK3CA	Driver	4/0	1/2	3/2	5/1
E542K	PIK3CA	Driver	3/1	1/0	0/1	1/0
E545K	PIK3CA	Driver	3/0	0/1	1/1	1/0
R130G	PTEN	Driver	5/1	1/3	3/2	0/3
G127E	PTEN	Driver	3/0	3/3	1/3	0/3
E62D	TP53	Passenger	1/4	1/2	0/0	0/1
I112V	PIK3CA	Passenger	1/4	0/0	0/1	0/1
D92Y	PTEN	Passenger	0/4	0/5	0/0	2/3

Silent_d: number of silent driver mutations (recurrent silent mutations).
Silent_p: number of silent passenger mutations (non-recurrent silent mutations).
1–10, 11–20, 21–30, 31–40, 41–50 are sequence positions from the mutant position

with driver silent mutations and there are no passenger silent mutations. In other position ranges 11–20, 21–30 and 31–40 from R175 in the sequence, the number of driver silent mutations is 9, 7 and 9, respectively whereas only 2 passenger silent mutations were observed in all the ranges from R175 with 40 possible positions. These results indicate that 32 among 40 positions are occupied with driver silent mutations whereas the passenger silent mutations are only two. In case of mutations in other genes, this pattern is more prominent in the position range of 1–10 (Table 8). On the other hand, for the passenger mutations, the number of passenger silent mutations is dominant around the mutation position. Based on these results, we suggest that the pattern of driver silent and passenger silent mutations around mutation position provides deep insights to understand the role of silent mutations in cancer progression.

4. Conclusions

We have analysed the missense mutations, insertions and deletions present in various cancer types in the available data in COSMIC database and developed amino acid mutation matrices for both driver and passenger missense mutations, which shows the frequency of mutations at protein level in cancer genes and also gives a comprehensive comparison between the driver and passenger missense substitutions. From the current analysis, we have identified few substitutions which are unique to driver mutations and further this information can be utilised for screening potential targets for cancer therapy. We also examined the distribution of driver and passenger mutations in topmost five highly mutated genes, at different window lengths from the hotspot mutant site in the sequence, which showed that driver mutations are highly accumulated near hotspot. Further, we analysed the distribution of silent mutations around missense mutations and found an interesting pattern where recurrent silent mutations mostly surround the driver missense mutations in the sequence. This observation potentiates the importance of silent mutations in cancer progression as suggested by earlier studies. The present study would help in better understanding the mutations observed in different cancer types and depicting their role in carcinogenesis. Along these directions, analysis on co-residue mutations for understanding different cancer pathways is on progress.

Abbreviations

CNS	Central Nervous System
HLT	Haematopoietic and Lymphoid Tissue
UAT	Urinary Aerodigestive Tract
TCGA	The Cancer Genome Atlas
COSMIC	Catalogue of Somatic Mutations in Cancer

Transparency document

The Transparency document associated with this article can be found, in the online version.

Acknowledgements

We thank the reviewers for their constructive comments. We acknowledge the Bioinformatics facility and Indian Institute of technology Madras for computational facilities. PA thanks Department of Science and Technology (DST), India for providing research fellowship.

References

[1] C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, C. Hunter, G. Bignell, Patterns of somatic mutation in human cancer genomes, Nature 446 (2007) 153–158.
[2] L.D. Wood, D.W. Parsons, S. Jones, J. Lin, T. Sjoblom, B. Vogelstein, The genomic landscapes of human breast and colorectal cancers, Science 318 (2007) 1108–1113.
[3] B.J. Raphael, J.R. Dobson, L. Oesper, F. Vandin, Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine, Genome Med. 6 (2014) 5.
[4] E.R. Mardis, R.K. Wilson, Cancer genome sequencing: a review, Hum. Mol. Genet. 18 (2009) R163–R168.
[5] S.A. Forbes, G. Bhamra, S. Bamford, E. Dawson, C. Kok, J. Clements, A. Menzies, J.W. Teague, P.A. Futreal, M.R. Stratton, The catalogue of somatic mutations in cancer (COSMIC), Curr. Protoc. Hum. Genet. (2008) (Chapter 10: Unit 10.11).
[6] P. Iengar, An analysis of substitution, deletion and insertion mutations in cancer genes, Nucleic Acids Res. 40 (2012) 6401–6413.
[7] C. Kandoth, M.D. McLellan, F. Vandin, K. Ye, B. Niu, C. Lu, M. Xie, Q. Zhang, J.F. McMichael, M.A. Wyczalkowski, M.D. Leiserson, C.A. Miller, J.S. Welch, M.J. Walter, M.C. Wendl, T.J. Ley, R.K. Wilson, B.J. Raphael, L. Ding, Mutational landscape and significance across 12 major cancer types, Nature 502 (2013) 333–339.
[8] G.R. Abecasis, A. Auton, L.D. Brooks, M.A. DePristo, R.M. Durbin, R.E. Handsaker, H.M. Kang, G.T. Marth, G.A. McVean, An integrated map of genetic variation from 1092 human genomes, Nature 491 (2012) 56–65.
[9] B. Vogelstein, N. Papadopoulos, V.E. Velculescu, S. Zhou, L.A. Diaz Jr., K.W. Kinzler, Cancer genome landscapes, Science 339 (2013) 1546–1558.
[10] G.R. Ritchie, I. Dunham, E. Zeggini, P. Flicek, Functional annotation of noncoding sequence variants, Nat. Methods 11 (2014) 294–296.
[11] I. Bozic, T. Antal, H. Ohtsuki, H. Carter, D. Kim, S. Chen, M.A. Nowak, Accumulation of driver and passenger mutations during tumor progression, Proc. Natl. Acad. Sci. 107 (2010) 18545–18550.
[12] J.S. Kaminker, Y. Zhang, C. Watanabe, Z. Zhang, CanPredict: a computational tool for predicting cancer associated missense mutations, Nucleic Acids Res. 35 (2007) 595–598.
[13] H. Carter, S. Chen, L. Isik, S. Tyekucheva, V.E. Velculescu, K.W. Kinzler, B. Vogelstein, R. Karchin, Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations, Cancer Res. 69 (2009) 6660–6667.
[14] P.C. Ng, S. Henikoff, SIFT: predicting amino acid changes that affect protein function, Nucleic Acids Res. 31 (2003) 3812–3814.
[15] C. Ferrer-Costa, J.L. Gelpí, L. Zamakola, I. Parraga, X. de la Cruz, M. Orozco, PMUT: a web-based tool for the annotation of pathological mutations on proteins, Bioinformatics 21 (2005) 3176–3178.
[16] P.D. Thomas, A. Kejariwal, N. Guo, H. Mi, M.J. Campbell, A. Muruganujan, B. Lazareva-Ulitsky, Applications for protein sequence-function evolution data: mRNA/protein

expression analysis and coding SNP scoring tools, Nucleic Acids Res. 34 (2006) 645–650.
[17] B. Li, V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, S.D. Mooney, P. Radivojac, Automated inference of molecular mechanisms of disease from amino acid substitutions, Bioinformatics 25 (2009) 2744–2750.
[18] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov, S.R. Sunyaev, A method and server for predicting damaging missense mutations, Nat. Methods 7 (2010) 248–249.
[19] J. Zhang, J. Liu, J. Sun, C. Chen, G. Foltz, B. Lin, Identifying driver mutations from sequencing data of heterogeneous tumors in the era of personalized genome sequencing, Brief. Bioinform. 15 (2014) 244–255.
[20] F. Gnad, A. Baucom, K. Mukhyala, G. Manning, Z. Zhang, Assessment of computational methods for predicting the effects of missense mutations in human cancers, BMC Genomics 14 (2013) S7.
[21] M.M. Gromiha, A statistical model for predicting protein folding rates from amino acid sequence with structural class information, J. Chem. Inf. Model. 45 (2005) 494–501.
[22] M.M. Gromiha, K. Fukui, Scoring function based approach for locating binding sites and understanding recognition mechanism of protein–DNA complexes, J. Chem. Inf. Model. 51 (2011) 721–729.
[23] M.M. Gromiha, K. Veluraja, K. Fukui, Identification and analysis of binding site residues in protein–carbohydrate complexes using energy based approach, Protein Pept. Lett. 21 (2014) 799–807.
[24] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Importance of mutant position in Ramachandran plot for predicting protein stability of surface mutations, Biopolymers 64 (2002) 210–220.
[25] M.M. Gromiha, Importance of native-state topology for determining the folding rate of two-state proteins, J. Chem. Inf. Comput. Sci. 43 (2003) 1481–1485.
[26] M.M. Gromiha, Protein Bioinformatics: From Sequence to Function, Elsevier Publishers, 2010.
[27] I. Reynolds, P. Healy, D.A. McNamara, Malignant tumours of the small intestine, Surgeon (2014) (pii: S1479-666X).
[28] S. Ahmad, M.M. Gromiha, A. Sarai, Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information, Bioinformatics 20 (2004) 477–486.
[29] C.L. Borders Jr., J.A. Broadwater, P.A. Bekeny, J.E. Salmon, A.S. Lee, A.M. Eldridge, V.B. Pett, A structural role for arginine in proteins: multiple hydrogen bonds to backbone carbonyl oxygens, Protein Sci. 3 (1994) 541–548.
[30] N.T. Mrabet, A. Van den Broeck, I. Van den Brande, P. Stanssens, Y. Laroche, A.M. Lambeir, G. Matthijssens, J. Jenkins, M. Chiadmi, H. van Tilbeurgh, F. Rey, J. Janin, W.J. Quax, I. Lasters, M. De Maeyer, S.J. Wodak, Arginine residues as stabilizing elements in proteins, Biochemistry 31 (1992) 2239–2253.
[31] W.H. Li, C.I. Wu, C.C. Luo, A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes, Mol. Biol. Evol. 2 (1985) 150–174.
[32] A. Petitjean, M.I. Achatz, A.L. Borresen-Dale, P. Hainaut, M. Olivier, TP53 mutations in human cancers: functional selection and impact on cancer prognosis and outcomes, Oncogene 26 (2007) 2157–2165.
[33] A. Vazquez, E.E. Bond, A.J. Levine, G.L. Bond, The genetics of P53 pathway, apoptosis and cancer therapy, Nat. Rev. Drug Discov. 7 (2008) 979–987.
[34] D. Karunagaran, E. Tzahar, R.R. Beerli, X. Chen, D. Graus-Porta, B.J. Ratzkin, R. Seger, N.E. Hynes, Y. Yarden, ErbB-2 is a common auxiliary subunit of NDF- and EGF-receptors: implications for breast cancer, EMBO J. 15 (1996) 254–264.
[35] M. Ayyoub, L. Memeo, F.E. Alvarez, C. Colarossi, R. Costanzo, E. Aiello, D. Martinetti, D. Valmori, Assessment of MAGE-A expression in resected non-small cell lung cancer in relation to clinicopathological features and mutational status of EGFR and KRAS, Cancer Immunol. Res. 2 (2014) 943–948.
[36] H. Sun, Y. Yang, L. Yang, B. Su, G. Jiang, K. Fei, D. Lu, Snapback primer mediated clamping PCR for detection of EGFR and KRAS mutations in NSCLC patients by high resolution melting analysis, Biomed. Res. Int. 2014 (2014) 407537.
[37] M.S. Lawrence, P. Stojanov, C.H. Mermel, J.T. Robinson, L.A. Garraway, T.R. Golub, M. Meyerson, S.B. Gabriel, E.S. Lander, G. Getz, Discovery and saturation analysis of cancer genes across 21 tumour types, Nature 505 (2014) 495–501.
[38] W.C. Barker, L.K. Ketcham, M.O. Dayhoff, A comprehensive examination of protein sequences for evidence of internal gene duplication, J. Mol. Evol. 10 (1978) 265–281.
[39] S. Henikoff, J.G. Henikoff, Amino acid substitution matrices from protein blocks, Proc. Natl. Acad. Sci. 89 (1992) 10915–10919.
[40] E. Tüdös, M. Cserzö, I. Simon, Predicting isomorphic residue replacements for protein design, Int. J. Pept. Protein Res. 36 (1990) 236–239.
[41] S. Kawashima, M. Kanehisa, AAindex: amino acid index database, Nucleic Acids Res. 28 (2000) 374.
[42] S.A. Benner, M.A. Cohen, G.H. Gonnet, Amino acid substitution during functionally constrained divergent evolution of protein sequences, Protein Eng. 7 (1994) 1323–1332.
[43] P. Anoosha, L.T. Huang, R. Sakthivel, D. Karunagaran, M.M. Gromiha, Discrimination of driver and passenger mutations in epidermal growth factor receptor in cancer, Mutat. Res. Fundam. Mol. Mech. Mutagen. 780 (2015) 24–34.
[44] S. Benvenuti, M. Frattini, S. Arena, C. Zanon, V. Cappelletti, D. Coradini, A. Bardelli, PIK3CA cancer mutations display gender and tissue specificity patterns, Hum. Mutat. 29 (2008) 284–288.
[45] K. Blighe, Cancer mutations and their tissue-specific nature, J. Cancer Sci. Ther. 6 (2014) 009–011.
[46] O.M. Sieber, S.R. Tomlinson, I.P. Tomlinson, Tissue, cell and stage specificity of (epi) mutations in cancers, Nat. Rev. Cancer 5 (2005) 649–655.
[47] M. Cserzö, I. Simon, Regularities in the primary structure of proteins, Int. J. Pept. Protein Res. 34 (1989) 184–195.
[48] S. Anishetty, R. Anishetty, G. Pennathur, Understanding mutations and protein stability through tripeptides, FEBS Lett. 580 (2006) 2071–2080.
[49] W.A. Freed-Pastor, C. Prives, Mutant p53: one name, many proteins, Genes Dev. 26 (2012) 1268–1286.
[50] K. Oda, D. Stokoe, Y. Taketani, F. McCormick, High frequency of coexistent mutations of PIK3CA and PTEN genes in endometrial carcinoma, Cancer Res. 65 (2005) 10669–10673.
[51] G. Smith, R. Bounds, H. Wolf, R.J.C. Steele, F.A. Carey, C.R. Wolf, Activating K-Ras mutations outwith 'hotspot' codons in sporadic colorectal tumours—implications for personalised cancer medicine, Br. J. Cancer 102 (2010) 693–703.
[52] C.M. Li, C.E. Kim, A.A. Margolin, M. Guo, J. Zhu, J.M. Mason, B. Tycko, CTNNB1 mutations and overexpression of Wnt/β-catenin target genes in WT1-mutant Wilms' tumors, Am. J. Pathol. 165 (2004) 1943–1953.
[53] I.C. De-Beauchêne, A. Allain, N. Panel, E. Laine, A. Trouvé, P. Dubreuil, L. Tchertanov, Hotspot mutations in KIT receptor differentially modulate its allosterically coupled conformational dynamics: impact on activation and drug sensitivity, PLoS Comput. Biol. 10 (2014), e1003749.
[54] J. Roszik, S.E. Woodman, HotSpotter: efficient visualization of driver mutations, BMC Genomics 15 (2014) 1044.
[55] Z.E. Sauna, C. Kimchi-Sarfaty, Understanding the contribution of synonymous mutations to human disease, Nat. Rev. Genet. 12 (2011) 683–691.
[56] F. Supek, B. Miñana, J. Valcárcel, T. Gabaldón, B. Lehner, Synonymous mutations frequently act as driver mutations in human cancers, Cell 156 (2014) 1324–1335.
[57] J.L. Parmley, J.V. Chamary, L.D. Hurst, Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers, Mol. Biol. Evol. 23 (2006) 301–309.
[58] D. Drummond, C.O. Wilke, Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution, Cell 134 (2008) 341–352.
[59] D.B. Goodman, G.M. Church, S. Kosuri, Causes and effects of N-terminal codon bias in bacterial genes, Science 342 (2013) 475–479.
[60] R.A. King, M.M. Mentink, W.S. Oetting, Non-random distribution of missense mutations within the human tyrosinase gene in type I (tyrosinase-related) oculocutaneous albinism, Mol. Biol. Med. 8 (1991) 19–29.
[61] D. Pérez-Caballero, C. González-Rubio, M.E. Gallardo, M. Vera, M. Lopez-Trascasa, S. Rodríguez de Córdoba, P. Sánchez-Corral, Clustering of missense mutations in the C-terminal region of factor H in atypical hemolytic uremic syndrome, Am. J. Hum. Genet. 68 (2001) 478–484.
[62] B. Karakas, K.E. Bachman, B.H. Park, Mutation of the PIK3CA oncogene in human cancers, Br. J. Cancer 94 (2006) 455–459.
[63] I. Rodríguez-Escudero, M.D. Oliver, A. Andrés-Pons, M. Molina, V.J. Cid, R. Pulido, A comprehensive functional analysis of PTEN mutations: implications in tumor-and autism-related syndromes, Hum. Mol. Genet. 20 (2011) 4132–4142.