Cancer Genetics 231–232 (2019) 67–79

Cancer Genetics

ORIGINAL ARTICLE

Untouchable genes in the human genome: Identifying ideal targets for cancer treatment

Ivan.P. Gorlov[a,*], Olga Y. Gorlova[a], Christopher I. Amos[b]

[a] *The Geisel School of Medicine, Department of Biomedical Data Science, Dartmouth College, HB7936, One Medical Center Dr., Dartmouth-Hitchcock Medical Center, Lebanon, NH 03756, United States;* [b] *Department of Medicine, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, United States*

Abstract

Background: Usually, genes with a higher-than-expected number of somatic mutations in tumor samples are assumed to be cancer related. We identified genes with a fewer-than-expected number of somatic mutations - "untouchable genes".

Methods: To predict the expected number of somatic mutations, we used a linear regression model with the number of mutations in the gene as an outcome, and gene characteristics, including gene size, nucleotide composition, level of evolutionary conservation, expression level and others, as predictors. Analysis of residuals from the regression model was used to compare the observed and predicted number of mutations.

Results: We have identified 19 genes with a less-than-expected number of loss-off-function (nonsense, frameshift or pathogenic missense) mutations – i.e., untouchable genes. The number of silent or neutral missense mutations in untouchable genes was equal or higher than the expected number. Many mucins, including *MUC16, MUC17, MUC6, MUC5AC, MUC5B,* and *MUC12,* are untouchable. We hypothesized that untouchable mucins help tumor cells to avoid immune response by providing a protective coat that prevents direct contact between effector immune cells, e.g., cytotoxic T-cells, and tumor cells. Survival analysis of available TCGA data demonstrated that overall survival of patients with low (below the median) expression of untouchable mucins was better compared to patients with high expression of untouchable mucins. Aside from mucins, we have identified a number of other untouchable genes.

Conclusions: Untouchable genes may be ideal targets for cancer treatment since suppression of untouchable genes is expected to inhibit survival of tumor cells.

Keywords Somatic mutations, Tumor samples, COSMIC database, Genes with the lower than expected number of somatic mutations.

Introduction

Carcinogenesis is often driven by somatic mutations in cancer-associated genes [1–3]. As a result, cancer-associated genes are among the most frequently mutated genes in tumor samples. Nevertheless, a simple counting of somatic mutations per gene to identify cancer-associated genes can be misleading because the number of mutations depends on gene characteeristics, including gene size, expression level, replication time, etc. For example, highly expressed genes tend to have a higher mutability compared to lower-expressed genes [4,5]. Genes that replicate later in the cell cycle tend to have a higher mutability compared to genes that replicate earlier [4,5]. Chromatin accessibility also has been shown to be positively associated with mutability [6]. A high frequency of $C > T$ transitions in CpG sites [7] indicates that nucleotide composition can be associated with gene mutability. These associations suggest that gene characteristics can be used to predict the number of somatic mutations in the gene. It is important to have a reliable prediction model, because the better we predict number of mutations in a gene, the better we can identify outliers – genes with an observed number of somatic mutations that is significantly different from the expected/predicted number. In our preliminary analysis, we found that the mutation-type-

Received August 10, 2018; received in revised form October 30, 2018; accepted January 18, 2019

* Corresponding author.

E-mail address: ivan.p.gorlov@dartmouth.edu

2210-7762/$ - see front matter © 2019 Elsevier Inc. All rights reserved.
https://doi.org/10.1016/j.cancergen.2019.01.005

specific models (those separately predicting the number of nonsense, missense and frameshift mutations) better explain variation in the number of somatic mutations than the pan-mutational model (paper submitted); therefore, we used mutation-type-specific prediction models in this study.

Usually, studies that are designed to identify cancer-associated genes are looking for genes with an *excess* number of somatic mutations. We hypothesized that genes with a *deficit* of loss-of-function mutations – untouchable genes – are also important. Untouchable genes do not harbor loss-of-function mutations in tumors because having functioning untouchable genes is critical for the survival of tumor cells. If identified, must-have untouchable genes may be ideal targets for cancer treatment because suppression of untouchable genes is expected to decrease tumor cell viability.

Methods and materials

Mutation data

We used mutation data from the Catalog of Somatic Mutations in Cancer (COSMIC). COSMIC is the largest repository of somatic mutations in tumor samples. Our recent (August 2018) visit to the COSMIC found that there are over 7 million unique somatic mutations detected in almost two million tumor samples. Nearly half of the mutations are detected by whole exome or whole genome sequencing. To ensure that all genes were tested the same number of times, only mutations detected by whole genome/exome sequencing were used in the analysis. All cancer types were included. Somatic mutation data from 20,147 tumor samples were analyzed. In total, there were 2833,115 missense; 183,823 nonsense; and 89,272 frameshift (FS) mutations. Mutations reported as SNPs were excluded from the analysis.

Gene characteristics used as predictors of the number of somatic mutations in the gene

The following gene characteristics were used as predictors:

(1) *Gene size.* We used NCBI Consensus coding sequence project data to retrieve sizes of the coding regions [8]. The largest transcript was used when multiple transcripts were reported for the same gene. We used a moving average to visualize the relationship between the gene size and the number of somatic mutations in the gene. In brief, genes were ranked based on size, from shortest to longest. A sliding window of 100 genes was moved along the ranked genes with one gene step. We found that this size of the sliding window was optimal for smoothing of the relationship while keeping the effects of strong outliers like *TP53* visible. The average gene size and average number of mutations were computed and displayed for each position of the window.

(2) *Number of potential sites in the gene for a given type of mutation.* The type of mutation (missense, nonsense, silent) produced by a single nucleotide substitution (SNS) depends on the type of substitution (e.g., $C > T$) and its position in a given codon. For each nucleotide position there are three possible SNSs, making the total number of all possible SNSs equal to $3 \times N$, where N is the length of the coding region of the gene in nucleotides. We predicted mutational outcomes (missense, nonsense or silent mutation) for all possible SNSs in all genes and used the number of potential sites for each type of mutation in the gene as a predictor.

(3) *Nucleotide composition.* For each gene, we estimated proportions of each of the four nucleotides and used them as predictors of the number of mutations in the gene.

(4) *Percentage of CpGs.* Mutation rate is known to be higher in CpG dinucleotides [7], therefore we used the percentage of CpGs as a predictor.

(5) *Evolutionary conservation.* It has been demonstrated that the level of evolutionary conservation of a gene correlates with its mutability [9]. We used a conservation index based on the number of known orthologs as a measure of the evolutionary conservation of each gene [10]. Orthologs for each gene were identified among 20 species with complete genome sequences: *Pan troglodytes, Macaca mulatta, Canis lupus familiaris, Bos taurus, Mus musculus, Rattus norvegicus, Gallus gallus, Xenopus tropicalis, Danio rerio, Drosophila melanogaster, Anopheles gambiae, Caenorhabditis elegans, Saccharomyces cerevisiae, Kluyveromyces lactis, Eremothecium gossypii, Schizosaccharomyces pombe, Magnaporthe oryzae, Neurospora crassa, Arabidopsis thaliana*, and *Oryza sativa*. Similar to our previous study [10], a conservation index of '1' was assigned to genes with zero or one ortholog; a conservation index of '2' was assigned to genes with two or three orthologs, and so on. Though identification of orthologs can be difficult for some genes and species because of low quality of sequencing or annotation data, this is unlikely to have a profound effect on the assessment of the level of evolutionary conservation because our estimates are based not on a single species but on 20 species with complete and annotated genome sequences.

(6) *Expression level of the gene.* The expression level of a gene negatively correlates with the density of somatic mutations in it [4,5]. Gene expression data for 1037 cancer cell lines was downloaded from the Cancer Cell Line Encyclopedia (CCLE) [11]. The average expression across CCLE cell lines was computed for each gene and used as a predictor of the number of somatic mutations in the gene.

(7) *Nucleotide diversity.* In preliminary analysis, we noted that genes with similar percentages of nucleotides (25% each) tend to have a higher density of somatic mutations. To account for this in our prediction model, we devised a single measure characterizing how strongly the proportions of four nucleotides deviate from being equal. We called this measure nucleotide diversity (ND). ND was defined as the probability that two nucleotides randomly selected from the gene coding sequence are different: $ND = 1-(P_{(A)}{}^2 + P_{(C)}{}^2 + P_{(G)}{}^2 + P_{(T)}{}^2)$, where $P_{(A)}$, $P_{(C)}$, $P_{(G)}$, and $P_{(T)}$ are the percentages of each nucleotide in the gene. ND was computed for each gene and used as a predictor.

(8) *SNP density.* One can expect that genes with a high intrinsic propensity to mutate will have a higher density of germline variants, e.g., SNPs. SNP density was computed as a ratio of the total number of unique SNPs to the size of the coding region in nucleotides. SNPs detected by the 1000 Genomes Project [12] were used to ensure that different genes were targeted the same number of times.

(9) *Relative replication time.* Late-replicating genes tend to have a higher number of mutations compared to earlier-replicating genes [4,5]. We used the relative replication time data from the study by Ryba et al. [13]. We used human genome build GRCh38 to match the positions of probes with positions of the genes. When several probes were mapped to the same gene, the average replication time was used as a predictor. The closest probe was used when there were no probes in the gene. The replication time (negative for early and positive for late-replicating genes) was used as a predictor of the number of mutations in the gene.

(10) *Chromatin accessibility.* Chromatin accessibility has been shown to be associated with mutability of the chromosomal region [6]. Data from the study by Sos et al. [14] were used in chromatin accessibility analysis. The study used a transposon hypersensitive sites sequencing assay. The mean chromatin accessibility across ten cell lines was computed for each gene and used as a predictor of the number of somatic mutations in the gene.

(11) *Number of pseudogenes.* In our pilot analysis, we found that the number of pseudogenes reported for a given gene is positively associated with number of somatic mutations in it. We used the number of pseudogenes reported by the Pseudogene.org database [15].

(12) *Number of silent mutations.* Even though some silent mutations are known to be functional [16], most of them are likely to be neutral. As a result, the density of silent mutations can be used as an integrative measure of gene mutability. We computed the density of silent mutations for each gene and used it as a predictor of the number of frameshift, missense and nonsense mutations.

Statistical analysis

Stepwise multiple linear regression models were used to predict the number of mutations in each gene. The analysis of residuals from the regression model was used to identify untouchable genes for which the observed number of frameshift, nonsense or missense mutations predicted to be pathogenic was significantly lower compared to the predicted numbers. Separate prediction models were developed to predict the number of nonsense, missense pathogenic, missense neutral, silent and frameshift mutations. To stratify missense mutations as neutral and pathogenic, we used the Functional Analysis Hidden Markov Models (FATHMM) method [17]. Mutations predicted to be neutral are unlikely to affect gene function, while mutations predicted to be pathogenic are likely to have a significant deleterious effect.

Z-scores for residuals were computed separately for each mutation type. A positive *Z*-score indicates an excess and negative *Z*-score a deficit of mutations compared to the expected/predicted numbers. As a threshold for statistical significance, we used the absolute value of Bonferroni corrected *Z* values. The Bonferroni correction was based on 15,610 tests (the total number of genes used in the analysis), yielding cut-off values of ± 4.79 for the negative and positive outliers.

Table 1 Gene characteristics significant in a stepwise, best-subset multiple linear regression model for frameshift mutations that resulted from deletions.

Predictor	*T*-test	*P*-value	Beta (ß)
Gene size in nucleotides	26.27	1.93E−47	0.33
Nucleotide diversity	−11.58	5.82E−19	−0.11
Number of silent mutations in the gene	9.85	2.74E−16	0.13
Gene expression in CCLE cancer cell lines	5.54	3.15E−08	0.05
Percentage of "CpGs"	−4.26	2.04E−05	−0.06
Percentage of "G"	−3.95	7.80E−05	−0.05
Percentage of "A"	3.94	8.14E−05	0.07
Evolutionary conservation	3.59	3.36E−04	0.03
Percentage of "C"	3.44	5.94E−04	0.06
Relative replication time	2.63	8.51E−03	0.02

Survival analysis

TCGA data were used to test the hypothesis that expression of untouchable genes is associated with patient survival. Because survival data in TCGA are limited, we conducted a pan-cancer analysis of 34 TCGA cancer types. We used Gene Expression Profiling Interactive Analysis (GEPIA) [18] for pan-cancer analysis of survival. Cancer cases were stratified into high- and low-expression groups based on the median expression level of each gene.

Results

Prediction of frameshift (FS) mutations

We used separate prediction models for FS insertions and FS deletions because molecular mechanisms of insertions and deletions are likely to be different [19]. Table 1 shows predictors that remained significant in the stepwise linear regression model for FS deletions. The most significant predictor was “Gene size in nucleotides” followed by “Nucleotide diversity” and the “Number of silent mutations in the gene”. The R^2 value for the whole model was 0.24.

Table 2 shows predictors that remained significant for FS insertions. The most significant predictor was “Gene size in nucleotides” followed by “Nucleotide diversity” and the “Number of silent mutations in the gene”. The R^2 value for the whole model was 0.26.

Prediction of nonsense mutations

Table 3 shows gene characteristics that remained significant in the multiple linear regression model for predicting the number of nonsense mutations in each gene. The most significant

Table 2 Gene characteristics significant in a stepwise, best subset multiple linear regression model for frameshift mutations that resulted from insertions.

Predictor	*T*-test	*P*-value	Beta (ß)
Gene size in nucleotides	18.34	4.85E−32	0.24
Number of silent mutations in the gene	8.59	1.56E−14	0.12
Nucleotide diversity	−8.83	6.76E−14	−0.09
Gene expression in CCLE cancer cell lines	4.98	6.51E−07	0.04
Percentage of "C"	4.14	3.46E−05	0.08
Percentage of "A"	3.87	1.11E−04	0.07
Percentage of "G"	−3.19	1.45E−03	−0.04
Percentage of "CpG"	−2.66	7.78E−03	−0.04

Table 3 Gene characteristics significant in a stepwise, best subset multiple linear regression model for nonsense mutations.

Predictor	*T*-test	*P*-value	Beta (ß)
Number of potential sites for nonsense mutations	28.88	5.89E−52	0.32
Number of silent mutations in the gene	22.61	2.76E−22	0.25
Relative replication time	−4.00	6.35E−05	−0.03
Percentage of "G"	−3.49	0.000479	−0.04
Percentage of "C"	−2.16	0.030656	−0.04
Evolutionary conservation	2.07	0.03807	0.01

Table 4 Gene characteristics significant in a stepwise, best subset multiple linear regression model for silent mutations.

Predictor	*T*-test	*P*-value	Beta (ß)
Gene size in nucleotides	172.73	32.58E−324	0.79
Relative replication time	−29.32	6.72E−53	−0.14
Gene expression in CCLE cancer cell lines	−23.16	8.12E−41	−0.12
Nucleotide diversity	21.07	8.32E−37	0.12
Density of SNPs (1 K Genomes Project)	13.86	7.45E−23	0.06
Percentage of "C"	12.97	5.76E−22	0.14
Percentage of "G"	8.30	1.11E−16	0.06
Number of pseudogenes	7.53	5.28E−14	0.03
Percentage of "CpG"	7.05	1.81E−12	0.06
Percentage of "A"	3.95	7.83E−05	0.04
Chromatin accessibility	−2.96	3.05E−03	−0.01

predictor was the number of potential sites for nonsense mutations. Other significant predictors included the number of detected silent mutations and replication time. The R^2 value for the whole model was 0.40.

Prediction of silent mutations

Table 4 shows predictors that remained significant in the multiple linear regression model predicting silent mutations. Size of the gene was the most significant predictor followed by the "Relative replication time" and "Gene expression in CCLE cancer cell lines". The R^2 value for the whole model was 0.89.

Table 5 Gene characteristics significant in a stepwise, best subset multiple linear regression model for somatic mutations predicted to be neutral by FATHMM.

Predictor	*T*-test	*P*-value	Beta (ß)
Number of silent mutations in the gene	79.86	2.86E−148	0.74
Evolutionary conservation	−38.74	8.32E−70	−0.21
Nucleotide diversity	−13.95	1.51E−24	−0.10
SNP density	13.88	6.81E−24	0.08
Percentage of "CpG"	−12.99	1.75E−22	−0.14
Relative replication time	5.80	6.6E−09	0.03
Number of pseudogenes	4.05	5.04E−05	0.02

Table 6 Gene characteristics significant in a stepwise, best subset multiple linear regression model for somatic mutations predicted to be pathogenic by FATHMM.

Predictor	*T*-test	*P*-value	Beta (ß)
Number of silent mutations in the gene	62.92	2.47E−116	0.56
Evolutionary conservation	27.66	1.86E−50	0.15
Gene size in nucleotides	25.78	4.79E−46	0.22
Relative replication time	−10.49	2.46E−17	−0.06
SNP density	−10.37	5.34E−17	−0.05
Percentage of "CpG"	7.12	1.11E−12	0.07
Nucleotide diversity	6.48	9.57E−11	0.04
Percentage of "G"	−5.09	3.58E−07	−0.04
Percentage of "C"	−5.02	5.12E−07	−0.06
Number of pseudogenes	−2.41	0.015904	−0.01

Prediction of the numbers of neutral and pathogenic missense mutations

Table 5 shows gene characteristics that remained significant in the multiple linear regression model for neutral missense mutations. The most significant predictor was the number of reported silent mutations in the gene. Other significant predictors included "Evolutionary conservation", "Nucleotide diversity", "SNP density" and "Percentage of 'CpG'". The R^2 value for the whole model was 0.87.

Table 6 shows gene characteristics that remained significant in the multiple linear regression model for prediction of pathogenic missense mutations. The most significant predictor was the number of silent mutations reported in the gene. Other significant predictors included "evolutionary conservation", "Gene size in nucleotides", "SNP density" and "Relative replication time". The R^2 value for missense mutations predicted to be pathogenic was 0.85 for the whole model.

Untouchable genes

Our major goal was to identify genes with a less-than-expected number of the loss-of-function mutations – i.e., untouchable genes. A gene was considered to be untouchable if the *Z*-score for any of loss-of-function (FS deletions, FS insertions, nonsense or missense pathogenic) mutations was <-4.79. Nineteen untouchable genes were identified. Table 7

Table 7 Residual *Z* scores for untouchable genes.

Gene	Pathogenic mutations			FS deletions			FS insertions			Nonsense			Neutral			Silent		
	Obs.	Pred.	*Z*-score	Obs.	Pred.	*Z*-score	Obs.	Pred.	*Z*-score	Obs.	Pred.	*Z*-score	Obs.	Pred.	*Z*-score	Obs.	Pred.	*Z*-score
MUC16	0	2852.2	**−30.0**	29	112.4	**−9.8**	20	42.6	**−5.24**	170	223.7	−3.5	5185	1876.7	**47.0**	1209	1241.3	3.6
FLG	7	1366.3	**−14.3**	31	40.0	−1.1	10	15.9	−1.36	127	114.7	0.8	3459	1156.9	**32.7**	719	329.0	**13.5**
MUC17	13	1333.2	**−13.9**	11	43.4	−3.8	4	17.1	−3.04	58	90.4	−2.1	3088	1153.5	**27.5**	694	393.3	**12.2**
MUC6	156	1073.9	**−9.7**	27	30.5	−0.4	19	12.3	1.56	44	66.5	−1.5	3372	952.4	**34.3**	595	209.3	**14.0**
HRNR	13	898.8	**−9.3**	14	26.2	−1.4	4	10.5	−1.50	68	73.3	−0.3	2332	799.0	**21.8**	483	243.6	**8.9**
DSPP	31	906.3	**−9.2**	2	21.7	−2.3	1	8.7	−1.78	11	51.5	−2.6	928	790.0	2.0	510	71.8	**14.2**
MUC5AC	185	1023.8	**−8.8**	41	45.0	−0.5	10	17.3	−1.69	43	74.9	−2.1	1485	767.5	**10.2**	52	35.2	0.8
OBSCN	1104	1918.1	**−8.6**	62	70.1	−1.0	15	26.7	−2.71	105	155.1	−3.2	1708	1330.9	**5.4**	862	731.4	**6.1**
MUC5B	173	981.5	**−8.5**	40	45.5	−0.7	9	17.5	−1.96	41	71.0	−1.9	1420	731.6	**9.8**	465.0	480.8	−0.4
FLG2	43	684.2	**−6.8**	16	21.3	−0.6	1	8.3	−1.70	93	64.7	1.8	1309	567.8	**10.5**	349	199.4	**5.3**
MUC12	0	630.5	**−6.6**	7	39.5	−3.8	10	15.0	−1.16	24	61.2	−2.4	1078	436.3	**9.1**	256.0	423.2	**−6.0**
AHNAK	594	1197.0	**−6.3**	30	46.1	−1.9	11	17.2	−1.43	54	93.7	−2.6	1032	756.7	3.9	508	422.3	2.5
ABCA13	518	1092.0	**−6.0**	34	39.8	−0.7	22	15.1	1.60	84	100.0	−1.0	1464	741.5	**10.3**	471	376.4	2.6
RP1L1	147	697.3	**−5.8**	36	22.4	1.6	7	8.7	−0.40	56	55.2	0.1	1318	581.3	**10.5**	369	207.6	**5.8**
MKI67	53	597.6	**−5.7**	40	27.8	1.4	15	10.7	1.00	52	59.2	−0.5	980	410.2	**8.1**	254.0	237.7	0.6
PLIN4	14	521.9	**−5.3**	12	14.5	−0.3	5	5.6	−0.14	9	31.0	−1.4	828	491.9	**4.8**	305	135.4	**6.4**
FSIP2	84	575.3	**−5.2**	15	44.2	−3.4	6	16.1	−2.34	30	102.5	**−4.8**	353	170.4	2.6	88	487.1	**−15.4**
MAGEC1	0	476.5	**−5.0**	12	12.9	−0.1	6	5.5	0.12	40	34.3	0.4	1421	504.2	**13.0**	283	106.6	**6.0**
PKD1L1	104	571.1	**−4.9**	28	21.0	0.8	5	8.0	−0.69	40	49.9	−0.6	861	425.1	**6.2**	260	236.6	0.9

Negative outliers (genes with a less-than-expected number of mutations) are shown in blue and positive outliers (genes with a more-than-expected number of mutations) are shown in red.

shows *Z*-scores for the untouchable genes. The scores were derived from prediction models for different types of mutations, with significant *Z*-scores.

Fig. 1 shows the distributions of *Z*-scores for FS deletions (a), FS insertions (b), nonsense mutations (c), pathogenic missense mutations (d), silent mutations (e), and neutral missense mutations (f). Census cancer genes tend to have larger *Z*-scores for FS, nonsense and pathogenic mutations, indicating that a residual analysis can also be used to detect cancer-associated genes.

Mucins as untouchable genes

We found that many mucins tend to have negative *Z*-scores for loss-of-function mutations: FSs, nonsense, and pathogenic missense mutations. The deficit of loss-of-function mutations is not a result of overall low mutability of mucins. In fact, mucins have a higher-than-expected number of silent and neutral mutations.

Table 8 shows average *Z*-scores for mucins, census cancer genes and all other genes in the human genome. The data are stratified by mutation types. All other genes in the human genome were used as a reference group in this analysis. The *Z*-scores for frameshift, nonsense, and pathogenic missense mutations for cancer census genes are higher compared to the *Z*-scores for other genes. On the contrary, mucins show a deficit of loss-of-function and pathogenic missense mutations (negative *Z*-scores). At the same time, mucins show an excess of silent and neutral mutations (positive *Z*-scores).

These findings can be further illustrated by comparison of the predicted and observed number of mutations (Fig. 2). Census cancer genes tend to have a higher-than-expected number of nonsense and pathogenic missense mutations, but not silent and neutral missense mutations. Mucins tend to have a lower-than-expected number of nonsense and pathogenic missense mutations, but not silent and neutral missense mutations.

Survival analysis

Direct contact between immune and tumor cells is required for the immune system to kill tumor cells. Mucins cover tumor cells with a protective coat, preventing a direct contact with effector immune cells [20,21]. One can expect, therefore, that downregulation of mucins will "undress" tumor cells, exposing them to the immune system. To test this hypothesis, we compared the overall survival of patients with high (above median) and low (below median) mucin expression. All untouchable genes were included in the analysis. Mucins were included regardless of their "untouchable" status. For seven genes (*DSPP, MUC5AC, MUC7, MUC17, MUC22, MAGEC1,* and *FLG2)* survival analysis was not possible because of a lack of survival and/or gene expression data. As a result, a total of twenty genes was analyzed. Table 9 shows the log rank test and hazard ratio for the analyzed genes. All untouchable mucins showed a significant association with overall survival: low mucin expressing patients had a better survival compared to higher expressing patients (see Fig. 3). Results were ambiguous for the mucins without a deficit of pathogenic mutations: *MUC15, MUC20*, and *MUCL1* were not associated

Table 8 *Z*-scores for mucins, census cancer genes and other genes in the human genome. Other genes were used as a reference group in the comparative analysis. *Z*-scores significantly different from *Z*-scores of other genes are shown in bold.

Z-score for	Mucins ($N=14$)					Census cancer genes ($N=380$)					Other genes ($N=15{,}221$)				
	Min	Max	Median	Mean	SE	Min	Max	Median	Mean	SE	Min	Max	Median	Mean	SE
FS deletions	−9.84	0.66	−0.69	**−1.82**	0.73	−2.83	59.63	0.08	**1.28**	0.24	−3.98	12.89	−0.14	−0.03	0.01
FS insertions	−5.24	1.56	−0.91	**−1.31**	0.45	−1.53	53.7	0.04	**1.29**	0.24	−2.71	23.22	−0.13	−0.03	0.01
Nonsense	−3.46	0.16	−1.02	**−1.17**	0.29	−2.64	75.77	0.08	**1.26**	0.29	−4.67	4.38	−0.03	−0.03	0.01
Pathogenic	−3.75	66.63	0.36	1.23	0.22	−14.31	9.72	−0.03	**−0.03**	0.01	−30.02	0.68	−2.66	−6.01	2.22
Silent	−9.87	14.03	−0.11	**0.85**	1.68	−5.25	11.93	−0.02	**0.25**	0.09	−15.37	14.19	−0.07	0.01	0.01
Neutral	−1.31	46.96	2.63	**10.21**	4.05	−9.83	7.15	−0.17	**−0.36**	0.07	−7.86	32.68	0.01	0.01	0.01

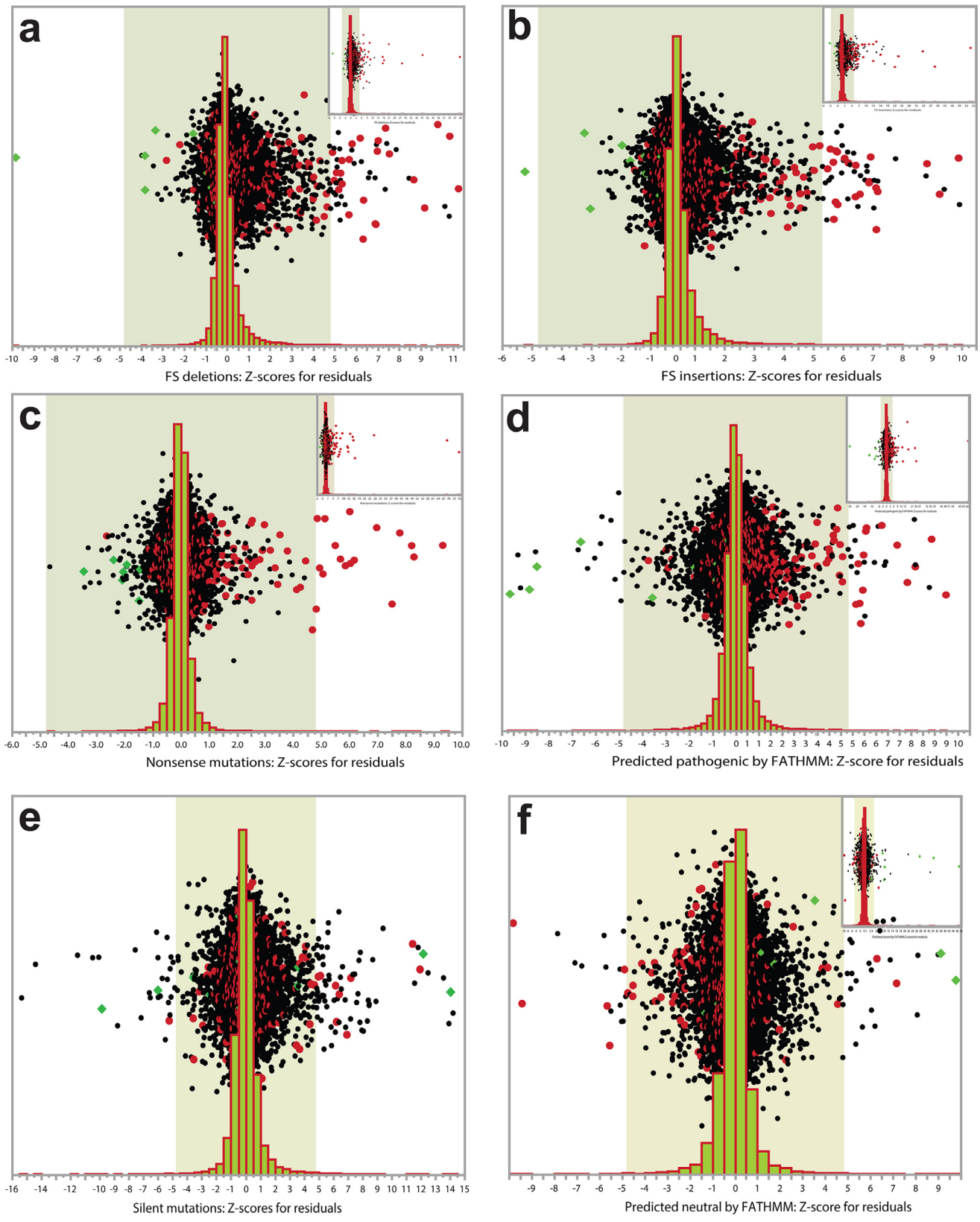

Fig. 1 Distribution of *Z*-scores from residual analysis. (a) Frameshift deletions; (b) frameshift insertions; (c) Nonsense mutations; (d) pathogenic missense mutations; (e) silent mutations; (f) neutral missense mutations. Mucins are shown by green diamonds, census cancer genes by red circles, and all other genes by black circles. For visualization purposes, the distributions were trimmed to remove extreme outliers. Inlets show the complete distributions, including strong outliers. Shaded areas encompass the *Z*-scores within the ± 4.79 region.

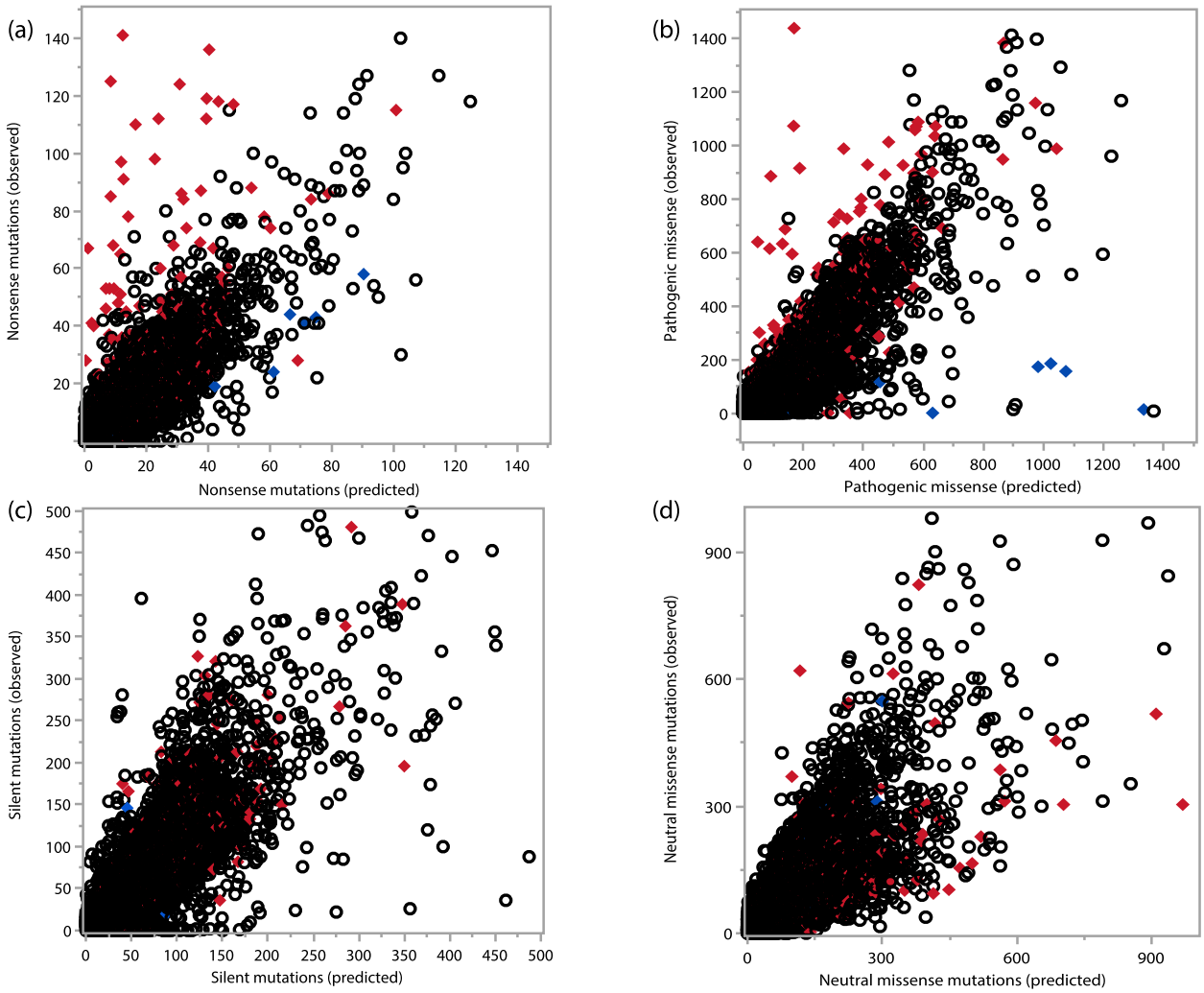

Fig. 2 The predicted and observed numbers of mutations. Each dot represents a gene. Red diamonds show census cancer genes, blue diamonds show mucins, and black circles show all other genes. (a) nonsense mutations; (b) pathogenic missense mutations; (c) silent mutations; and d) neutral missense mutations.

with survival, while *MUC1, MUC4,* and *MUC21* were. The results of survival analysis were also ambiguous for non-mucin genes. We found that *HRNR, ABCA13, RP1L1*, and *MKI67* were associated with survival, while *FLG, OBSCN, AHNAK, PLIN4, FSIP2*, and *PKD1L1* were not.

Many untouchable mucins are membrane tethered

To assess cellular localization of untouchable and non-untouchable mucins, we used data from the COMPARTMENTS database [22]. COMPARTMENTS maintains conclusions about cellular localization of a gene product based on manually curated reviews of literature and data from high-throughput immunohistochemistry studies. The confidence scores (CSs) of COMPARTMENTS are an integrated measure of evidence for localization of the protein in each of the pre-specified (total eleven) subcellular compartments. CSs can be between zero (no evidence for localization in a specific cellular compartment) and five (very strong evidence for localization of the protein in a given compartment). The results of the analysis are shown in Table 10. All untouchable mucins, including *MUC4,* which shows a nominally significant deficit of pathogenic mutations, have the highest confidence score for localization in a cellular membrane. Out of 6 non-untouchable mucins, only two show strong evidence for cellular membrane localization.

Untouchable mucins tend to be less amplified in tumor samples compared to mucins that do not show a deficit of pathogenic or loss-of-function mutations

We used COSMIC whole genome sequencing data to identify samples with Copy Number Variation (CNV) gains or losses. For CNV assessment, COSMIC uses a strict definition of gain: $\geq$ 5 copies if average genome ploidy $\leq$2.7 or $\geq$ 9 copies if average genome ploidy $\geq$2.7. We found that untouchable mucins show CNV gains in a smaller percentage of samples

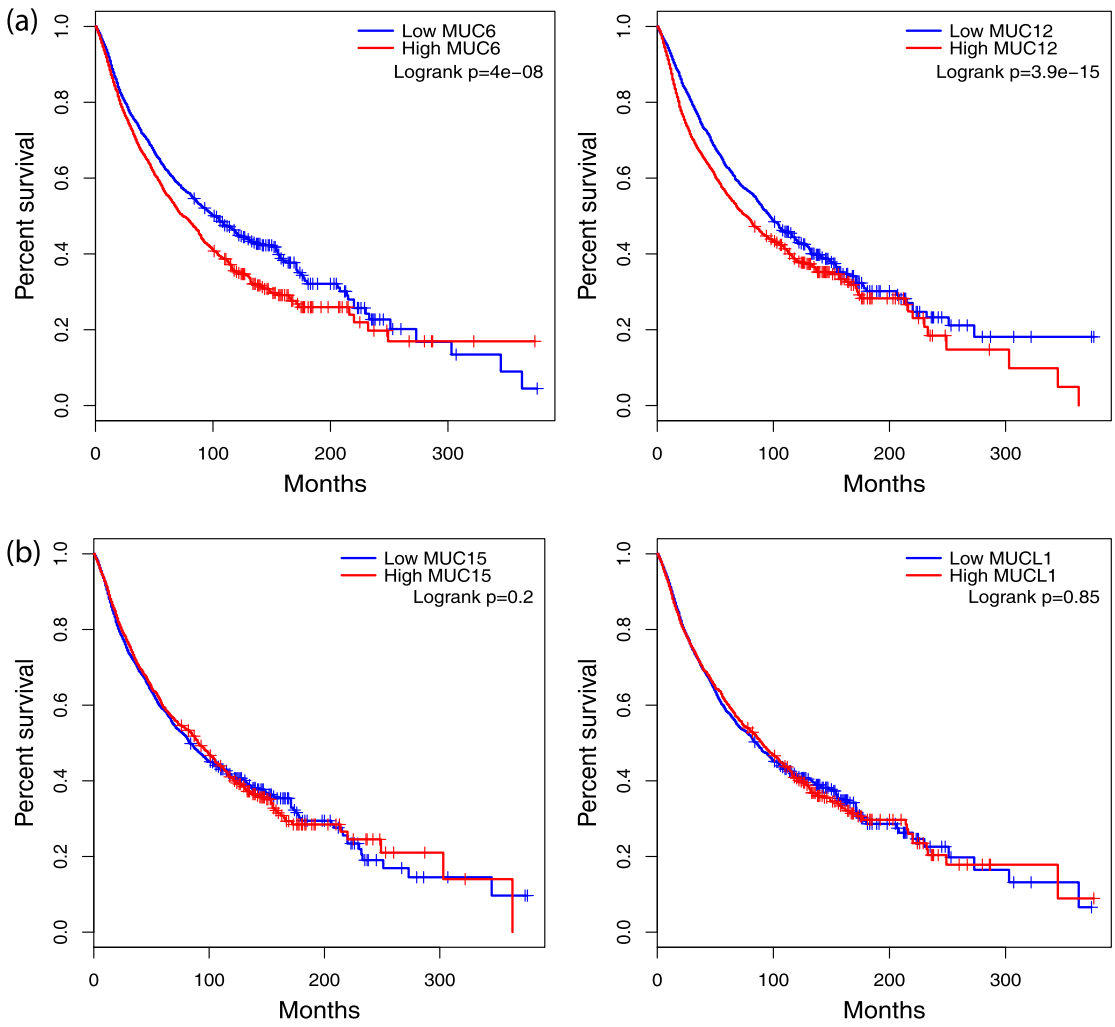

Fig. 3 Survival curves for two mucin genes with (MUC6 and *MUC12*) and without (*MUC15, MUCL1*) a deficit of loss-of-function or pathogenic mutations. High expression cases are shown in red and low expressing cases are shown in blue.

compared to non-untouchable mucins—those without a deficit of pathogenic mutations. The mean number of samples with a gain of mucin for untouchable mucins was 23±8, and the mean number of samples with a loss of mucin for untouchable mucins was 198±51. The difference is significant (t-test3.4, df=10, $p=0.008$). Untouchable and non-untouchable mucins do not differ by CNV losses: 8 ± 2 and 9 ± 6, correspondingly.

Discussion

We identified nineteen genes in the human genome with a deficit of loss-of-function mutations. Six out of nineteen untouchable genes are mucins. Mucins are known to be associated with cancer initiation and progression through regulation of inflammation, cell adhesion or proliferation (see [20,23,24] for review). Expression of mucins is the prognostic marker for survival in lung, [25] breast, [26] colorectal, [27] and prostate cancer [28].

We found that mucin 16 (MUC16) exhibits the strongest deficit of pathogenic mutations, with almost 3000 predicted pathogenic missense mutations and zero observed pathogenic missense mutations (Table 7). The observed numbers of frameshift and nonsense mutations were also lower than expected. The observed number of silent mutations does not differ from the expected number, and the observed number of neutral missense mutations was higher than expected. The association of *MUC16* with cancer has been known for almost 40 years. The first report on the CA125 antigen encoded by *MUC16* was published in 1981 [29]. CA125 is used biomarker for detection and screening of ovarian cancer [30,31]. Other untouchable mucins, including *MUC17* [32], *MUC6, MUC5AC* [33,34], *MUC5B* [35,36], MUC12 [37], and *MUC4* [38–40], are known to be associated with cancer risk and development.

Untouchable mucins may provide a critical survival advantage to tumor cells. We hypothesized that loss-of-function or pathogenic mutations in untouchable mucins expose tumor cells to a host's immune system, which leads to elimination of the mutant clone, and that is the reason why we do not detect loss-of-function or pathogenic mutations in untouchable genes, while neutral and silent mutations are allowed. Mucins form a physical barrier for antigen-presenting cells or immune effector cells (natural killer cells, cytotoxic T lymphocytes). Those immune cells need to come into contact with tumor

Table 9 P-values from log ratio (LR) test, and hazard ratio (HR) from survival analysis.

Gene name	Mucin	Untouchable	LR P	HR
FLG	No	Yes	6.80E−01	1
HRNR	No	Yes	3.00E−02	0.89
OBSCN	No	Yes	3.40E−01	0.95
AHNAK	No	Yes	4.90E−01	1
ABCA13	No	Yes	1.00E−14	1.5
RP1L1	No	Yes	5.00E−08	1.3
MKI67	No	Yes	1.00E−14	1.7
PLIN4	No	Yes	8.80E−01	0.98
FSIP2	No	Yes	8.90E−01	1
PKD1L1	No	Yes	7.20E−01	1
MUC16	Yes	Yes	1.60E−03	1.2
MUC6	Yes	Yes	4.00E−07	1.2
MUC5B	Yes	Yes	5.00E−07	1.2
MUC12	Yes	Yes	3.90E−14	1.3
MUC4	Yes	No	1.10E−11	1.3
MUC21	Yes	No	5.60E−08	1.3
MUC15	Yes	No	5.60E−01	0.95
MUC20	Yes	No	6.80E−02	1.1
MUCL1	Yes	No	9.50E−01	0.99
MUC1	Yes	No	8.60E−03	1.2

Table 10 Cellular localization of mucins. CM – cellular membrane, EC – extracellular localization.

Gene	Localization	
	CM	EC
MUC16	5	5
MUC17	5	4
MUC6	5	5
MUC5AC	5	5
MUC5B	4	5
MUC12	5	2
MUC4	5	5
MUC21	4	5
MUC15	4	4
MUC20	5	4
MUC7	4	5
MUCL1	4	5
MUC1	5	5

cells directly to kill them. Consistent with this hypothesis, we found that expression of untouchable mucins is negatively associated with patient survival.

There is more than one way by which mucins may influence cancer development. Mucins may regulate the expression of the genes directly involved in cell proliferation and survival [41]. Missense mutations in mucin genes may impact immune response and survival through creation of neoantigenic epitopes [42,43]. Mucins produce multiple splice isoforms [44,45] and have overlapping function with compensatory upregulation of the lost or downregulated genes [46,47].

We focused on loss-of-function and neutral mutation in mucins. We found that a number of mucins (dubbed untouchable in this study) show a significant deficit of loss-of-function mutations, suggesting that having functional mucins is important for survival of tumor cells. We found that lower expression of mucins is associated with a better overall survival in pan-cancer setting. Some studies however reported the opposite effect of mucin expression on patient survival [48,49]. The difference can be attributed to interplay between different mechanisms by which mucins influence cancer development. For example, tumor cells may become more vulnerable to immune system either due to the loss of protective physical barrier (loss-of-function mutations in untouchable mucins) or due to acquiring neoantigen-generating missense mutations. In the latter case higher expression of mutant mucin with neoantigenic properties may lead to a stronger immune response and a better patient survival. Future studies focused on a joint analysis of loss-of-function and neoantigenic mutations will provide a better understanding of the role of mucin somatic mutations in cancer development and progression.

A recent review by Dhanisha et al. [50] provides an excellent description of mucins, including review of studies where mucins were used as targets for cancer treatment. There are currently 24 Phase 1–3 clinical trials targeting mucins. Fifteen studies are targeting *MUC1,* three studies are investigating *MUC4,* three studies focus on *MUC13*, and three studies analyze *MUC16*. Only one (*MUC16*) untouchable mucin is included in ongoing clinical trials.

We found that untouchable mucins that are extremely rare show CNV gains, while non-untouchable mucins are likely to be amplified. Amplification could allow non-untouchable mucins to accumulate mutations in amplified copies of the gene while maintaining enough amounts of functional wild-type gene sequences. It is possible, therefore, that both (untouchable and non-untouchable) mucins are important for tumor survival, but mucins with CNV gains can allow to have loss-of-function mutations.

A number of non-mucin genes with a deficit of loss-of-function mutations were also detected. Below we provide brief descriptions of non-mucin untouchable genes.

FLG – filaggrin. FLG is a filament-associated protein that aggregates keratin filaments. FLG is a key structural element of the epidermal barrier. Up to ten percent of people of European origin carry one of two loss-of-function mutations in FLG, designated R501X and 2282del4 [51]. Both of these mutations lead to premature termination of transcription and homozygosity is associated with the complete absence of processed filaggrin in the epidermis [51]. Loss-of-function FLG mutations increase the risk for atopic dermatitis [52] and allergic rhinitis [53]. There is no evidence that loss-of-function filaggrin mutations are associated with cancer risk [54].

HRNR – Hornerin, FLG3. This protein is the component of the epidermal cornified cell envelope, which serves as the upper physical barrier of epidermis. Elevated expression of hornerin has been shown to be associated with breast cancer progression [55].

DSPP – Dentin Sialophosphoprotein. This preproprotein is secreted by odontoblasts and is proteolytically processed to generate two principal proteins of the dentin extracellular matrix of the tooth, dentin sialoprotein and dentin phosphoprotein. *DSPP* gene-silencing inhibits key tumorigenic activities in the human oral cancer cell line [56].

OBSCN – Obscurin. OBSCN is important in the organization of myofibrils. There is no published evidence on the role of *OBSCN* in cancer.

FLG2 – Filaggrin Family Member 2. FLG2 is functionally similar to FLG. The protein encoded by *FLG2* is required for skin cornification. Defects in this gene are associated with

skin diseases. There is no published evidence on the role of *FLG2* in cancer.

AHNAK – Desmoyokin. AHNAK is a structural scaffold protein important in maintaining the blood-brain barrier. The gene plays a role in cell migration. A number of publications indicate the involvement of the *AHNAK* in cancer development, in the progression of breast cancer [57], and in glioma [58].

ABCA13 – ATP Binding Cassette Subfamily A Member 13. ABC transporters are ubiquitously expressed membrane-bound proteins. There is no published evidence on the role of this protein in cancer.

RP1L1 – Retinitis Pigmentosa 1 Like 1. RP1L1 regulates microtubule polymerization. We are not aware of any studies linking RP1L1 to cancer risk or progression.

MKI67 – Marker Of Proliferation Ki-67. This gene plays a role in maintaining chromosomes and preventing them from collapsing into a single chromatin mass [59]. There is no published evidence on the role of *MKI67* in cancer.

PLIN4 – Perilipin 4. PLIN4 plays an important role in the regulation of lipolysis [60,61]. There is no published evidence that indicates this gene is associated with cancer.

FSIP2 – Fibrous Sheath Interacting Protein 2. This gene encodes a protein that is a structural component of the sperm fibrous sheath. Currently, there is no published evidence that this gene is cancer-associated.

MAGEC1 – MAGE Family Member C1. MAGEC1 is a member of the melanoma antigen gene family. MAGE antigens are recognized by cytolytic T lymphocytes. There is published evidence on the role of MAGEC1 in cancer, including myeloma [62], prostate cancer [63], breast cancer [64], and melanoma [65].

PKD1L1 – Polycystin 1 Like 1. PKD1L1 is a member of the polycystin protein family and serves as a cation channel in the cellular membrane. There is no published evidence on the role of this gene in cancer.

Untouchable genes are likely to represent genes of two types: genes that provide protection against immune response, and essential housekeeping genes that are required for survival of any cell, including tumor cells. Untouchable mucins are examples of the genes in the first category, and *FLG, MKI67, OBSCN* and *PKD1L1* are examples of the genes in the second category. The four listed genes are ubiquitously expressed essential genes [66]. Targeting essential housekeeping genes is likely to be challenging because it will be associated with significant side effects. Targeting untouchable mucins seems to be more feasible and less likely to be associated with side effects.

Disclosure

There is no conflict of interest for any of the authors.

Authors' contributions

Conception and design: I. Gorlov, O. Gorlova, C. Amos.

Development of methodology: I. Gorlov, O. Gorlova, C. Amos.

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): I. Gorlov, O. Gorlova, C. Amos.

Writing, review, and/or revision of the manuscript: I. Gorlov, O. Gorlova, C. Amos.

Acknowledgments

This work was supported in part by the National Institutes of Health P01 CA206980-01A1, **U19 CA148127** and 1R56LM12371-01A1. This work is also supported by RR170048 grant from the Cancer Prevention Research Institute of Texas.

The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

We are grateful to Lamar Moss for help with preparation of the manuscript for publication.

References

[1] Black RC, Khurshid H. NSCLC: an update of driver mutations, their role in pathogenesis and clinical significance. R I Med J 2013;2015(98):25–8.

[2] Pon JR, Marra MA. Driver and passenger mutations in cancer. Annu Rev Pathol 2015;10:25–50.

[3] Watson IR, Takahashi K, Futreal PA, Chin L. Emerging patterns of somatic mutations in cancer. Nat Rev Genet 2013;14:703–18.

[4] Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, Dicara D, Ramos AH, Lawrence MS, Cibulskis K, Sivachenko A, Voet D, Saksena G, Stransky N, Onofrio RC, Winckler W, Ardlie K, Wagle N, Wargo J, Chong K, Morton DL, Stemke-Hale K, Chen G, Noble M, Meyerson M, Ladbury JE, Davies MA, Gershenwald JE, Wagner SN, Hoon DS, Schadendorf D, Lander ES, Gabriel SB, Getz G, Garraway LA, Chin L. A landscape of driver mutations in melanoma. Cell 2012;150:251–63.

[5] Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, Kiezun A, Hammerman PS, McKenna A, Drier Y, Zou L, Ramos AH, Pugh TJ, Stransky N, Helman E, Kim J, Sougnez C, Ambrogio L, Nickerson E, Shefler E, Cortes ML, Auclair D, Saksena G, Voet D, Noble M, DiCara D, Lin P, Lichtenstein L, Heiman DI, Fennell T, Imielinski M, Hernandez B, Hodis E, Baca S, Dulak AM, Lohr J, Landau DA, Wu CJ, Melendez-Zajgla J, Hidalgo-Miranda A, Koren A, McCarroll SA, Mora J, Lee RS, Crompton B, Onofrio R, Parkin M, Winckler W, Ardlie K, Gabriel SB, Roberts CW, Biegel JA, Stegmaier K, Bass AJ, Garraway LA, Meyerson M, Golub TR, Gordenin DA, Sunyaev S, Lander ES, Getz G. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–218.

[6] Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, Garg K, John S, Sandstrom R, Bates D, Boatman L, Canfield TK, Diegel M, Dunn D, Ebersol AK, Frum T, Giste E, Johnson AK, Johnson EM, Kutyavin T, Lajoie B, Lee BK, Lee K, London D, Lotakis D, Neph S, Neri F, Nguyen ED, Qu H, Reynolds AP, Roach V, Safi A, Sanchez ME, Sanyal A, Shafer A, Simon JM, Song L, Vong S, Weaver M, Yan Y, Zhang Z, Zhang Z, Lenhard B, Tewari M, Dorschner MO, Hansen RS, Navas PA, Stamatoyannopoulos G, Iyer VR, Lieb JD, Sunyaev SR, Akey JM, Sabo PJ, Kaul R, Furey TS, Dekker J, Crawford GE, Stamatoyannopoulos JA. The accessible chromatin landscape of the human genome. Nature 2012;489:75–82.

[7] Millar CB, Guy J, Sansom OJ, Selfridge J, MacDougall E, Hendrich B, Keightley PD, Bishop SM, Clarke AR, Bird A. Enhanced CpG mutability and tumorigenesis in MBD4-deficient mice. Science 2002;297:403–5.
[8] Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, Dicuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. The consensus coding sequence (CCDS) project: identifying a common protein–coding gene set for the human and mouse genomes. Genome Res 2009;19:1316–23.
[9] Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 2012;151:1431–42.
[10] Gorlov IP, Moore JH, Peng B, Jin JL, Gorlova OY, Amos CI. SNP characteristics predict replication success in association studies. Hum Genet 2014;133:1477–86.
[11] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P Jr, de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 2012;483:603–7.
[12] Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature 2015;526:68–74.
[13] Ryba T, Battaglia D, Chang BH, Shirley JW, Buckley Q, Pope BD, Devidas M, Druker BJ, Gilbert DM. Abnormal developmental control of replication-timing domains in pediatric acute lymphoblastic leukemia. Genome Res 2012;22:1833–44.
[14] Sos BC, Fung HL, Gao DR, Osothprarop TF, Kia A, He MM, Zhang K. Characterization of chromatin accessibility with a transposome hypersensitive sites sequencing (THS-seq) assay. Genome Biol 2016;17:20.
[15] Karro JE, Yan Y, Zheng D, Zhang Z, Carriero N, Cayting P, Harrrison P, Gerstein M. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. Nucleic Acids Res 2007;35:D55–60.
[16] Fernandez-Calero T, Cabrera-Cabrera F, Ehrlich R, Marin M. Silent polymorphisms: can the tRNA population explain changes in protein properties? Life (Basel) 2016;6.
[17] Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 2018;34:511–13.
[18] Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res 2017;45:W98–W102.
[19] Sancar A, Lindsey-Boltz LA, Unsal-Kacmaz K, Linn S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. Annu Rev Biochem 2004;73:39–85.
[20] Hollingsworth MA, Swanson BJ. Mucins in cancer: protection and control of the cell surface. Nat Rev Cancer 2004;4:45–60.
[21] Anandkumar A, Devaraj H. Tumour immunomodulation: mucins in resistance to initiation and maturation of immune response against tumours. Scand J Immunol 2013;78:1–7.
[22] Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI, Schneider R, Jensen LJ. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database (Oxford) 2014;2014 bau012.
[23] Kufe DW. Mucins in cancer: function, prognosis and therapy. Nat Rev Cancer 2009;9:874–85.
[24] Segal-Eiras A, Croce MV. Breast cancer associated mucin: a review. Allergol Immunopathol (Madr) 1997;25:176–81.
[25] Lakshmanan I, Ponnusamy MP, Macha MA, Haridas D, Majhi PD, Kaur S, Jain M, Batra SK, Ganti AK. Mucins in lung cancer: diagnostic, prognostic, and therapeutic implications. J Thorac Oncol 2015;10:19–27.
[26] Cai X, Sheng J, Tang C, Nandakumar V, Ye H, Ji H, Tang H, Qin Y, Guan H, Lou F, Zhang D, Sun H, Dong H, Zhang G, Liu Z, Dong Z, Guo B, Yan H, Yan C, Wang L, Su Z, Li Y, Jones L, Huang XF, Chen SY, Wu T, Lin H. Frequent mutations in EGFR, KRAS and TP53 genes in human lung cancer tumors detected by ion torrent DNA sequencing. PLoS One 2014;9:e95228.
[27] Itzkowitz SH, Bloom EJ, Kokal WA, Modin G, Hakomori S, Kim YS. Sialosyl-Tn. A novel mucin antigen associated with prognosis in colorectal cancer patients. Cancer 1990;66:1960–6.
[28] Cozzi PJ, Wang J, Delprado W, Perkins AC, Allen BJ, Russell PJ, Li Y. MUC1, MUC2, MUC4, MUC5AC and MUC6 expression in the progression of prostate cancer. Clin Exp Metastasis 2005;22:565–73.
[29] Bast RC Jr, Feeney M, Lazarus H, Nadler LM, Colvin RB, Knapp RC. Reactivity of a monoclonal antibody with human ovarian carcinoma. J Clin Invest 1981;68:1331–7.
[30] Felder M, Kapur A, Gonzalez-Bosquet J, Horibata S, Heintz J, Albrecht R, Fass L, Kaur J, Hu K, Shojaei H, Whelan RJ, Patankar MS. MUC16 (CA125): tumor biomarker to cancer therapy, a work in progress. Mol Cancer 2014;13:129.
[31] Karam AK, Karlan BY. Ovarian cancer: the duplicity of CA125 measurement. Nat Rev Clin Oncol 2010;7:335–9.
[32] Kitamoto S, Yokoyama S, Higashi M, Yamada N, Matsubara S, Takao S, Batra SK, Yonezawa S. Expression of MUC17 is regulated by HIF1alpha-mediated hypoxic responses and requires a methylation-free hypoxia responsible element in pancreatic cancer. PLoS One 2012;7:e44108.
[33] Javanbakht M, Akhavanmoghadam J, Talaei AJ, Aghyani M, Mozafari M, Khedmat L, Mohebbi M. Differential expression of two genes Oct-4 and MUC5AC associates with poor outcome in patients with gastric cancer. Clin Exp Pharmacol Physiol 2017;44:1099–105.
[34] Kaur S, Smith LM, Patel A, Menning M, Watley DC, Malik SS, Krishn SR, Mallya K, Aithal A, Sasson AR, Johansson SL, Jain M, Singh S, Guha S, Are C, Raimondo M, Hollingsworth MA, Brand RE, Batra SK. A combination of MUC5AC and CA19-9 improves the diagnosis of pancreatic cancer: a multicenter study. Am J Gastroenterol 2017;112:172–83.
[35] Lahdaoui F, Messager M, Vincent A, Hec F, Gandon A, Warlaumont M, Renaud F, Leteurtre E, Piessen G, Jonckheere N, Mariette C, Van Seuningen I. Depletion of MUC5B mucin in gastrointestinal cancer cells alters their tumorigenic properties: implication of the Wnt/beta-catenin pathway. Biochem J 2017;474:3733–46.
[36] Nagashio R, Ueda J, Ryuge S, Nakashima H, Jiang SX, Kobayashi M, Yanagita K, Katono K, Satoh Y, Masuda N, Murakumo Y, Hachimura K, Sato Y. Diagnostic and prognostic significances of MUC5B and TTF-1 expressions in resected non-small cell lung cancer. Sci Rep 2015;5:8649.

[37] Matsuyama T, Ishikawa T, Mogushi K, Yoshida T, Iida S, Uetake H, Mizushima H, Tanaka H, Sugihara K. MUC12 mRNA expression is an independent marker of prognosis in stage II and stage III colorectal cancer. Int J Cancer 2010;127:2292–9.
[38] Carraway KL, Theodoropoulos G, Kozloski GA, Carothers Carraway CA. Muc4/MUC4 functions and regulation in cancer. Future Oncol 2009;5:1631–40.
[39] Mercogliano MF, Inurrigarro G, De Martino M, Venturutti L, Rivas MA, Cordo-Russo R, Proietti CJ, Fernandez EA, Frahm I, Barchuk S, Allemand DH, Figurelli S, Deza EG, Ares S, Gercovich FG, Cortese E, Amasino M, Guzman P, Roa JC, Elizalde PV, Schillaci R. Invasive micropapillary carcinoma of the breast overexpresses MUC4 and is associated with poor outcome to adjuvant trastuzumab in HER2-positive breast cancer. BMC Cancer 2017;17:895.
[40] Urey C, Andersson B, Ansari D, Sasor A, Said-Hilmersson K, Nilsson J, Andersson R. Low MUC4 expression is associated with survival benefit in patients with resectable pancreatic cancer receiving adjuvant gemcitabine. Scand J Gastroenterol 2017;52:595–600.
[41] Katoh M. Multilayered prevention and treatment of chronic inflammation, organ fibrosis and cancer associated with canonical WNT/betacatenin signaling activation (Review). Int J Mol Med 2018;42:713–25.
[42] Balachandran VP, Luksza M, Zhao JN, Makarov V, Moral JA, Remark R, Herbst B, Askan G, Bhanot U, Senbabaoglu Y, Wells DK, Cary CIO, Grbovic-Huezo O, Attiyeh M, Medina B, Zhang J, Loo J, Saglimbeni J, Abu-Akeel M, Zappasodi R, Riaz N, Smoragiewicz M, Kelley ZL, Basturk O, Gonen M, Levine AJ, Allen PJ, Fearon DT, Merad M, Gnjatic S, Iacobuzio–Donahue CA, Wolchok JD, DeMatteo RP, Chan TA, Greenbaum BD, Merghoub T, Leach SDAustralian Pancreatic Cancer Genome I, Garvan Institute of Medical R, Prince of Wales H, Royal North Shore H, University of G, St Vincent's H, Institute QBMR, University of Melbourne CfCR, University of Queensland IfMB, Bankstown H, Liverpool H, Royal Prince Alfred Hospital COBL, Westmead H, Fremantle H, St John of God H, Royal Adelaide H, Flinders Medical C, Envoi P, Princess Alexandria H, Austin H, Johns Hopkins Medical I, Cancer AR-NCfARo. Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer. Nature 2017;551:512–16.
[43] Haugstad KE, Hadjialirezaei S, Stokke BT, Brewer CF, Gerken TA, Burchell J, Picco G, Sletmoen M. Interactions of mucins with the Tn or Sialyl Tn cancer antigens including MUC1 are due to GalNAc-GalNAc interactions. Glycobiology 2016;26:1338–50.
[44] Kumar S, Cruz E, Joshi S, Patel A, Jahan R, Batra SK, Jain M. Genetic variants of mucins: unexplored conundrum. Carcinogenesis 2017;38:671–9.
[45] Sousa AM, Grandgenett PM, David L, Almeida R, Hollingsworth MA, Santos-Silva F. Reflections on MUC1 glycoprotein: the hidden potential of isoforms in carcinogenesis. APMIS 2016;124:913–24.
[46] Balague C, Gambus G, Carrato C, Porchet N, Aubert JP, Kim YS, Real FX. Altered expression of MUC2, MUC4, and MUC5 mucin genes in pancreas tissues and cancer cell lines. Gastroenterology 1994;106:1054–61.
[47] Verma M, Davidson EA. Mucin genes: structure, expression and regulation. Glycoconj J 1994;11:172–9.
[48] Fu H, Liu Y, Xu L, Chang Y, Zhou L, Zhang W, Yang Y, Xu J. Low expression of mucin-4 predicts poor prognosis in patients with clear-cell renal cell carcinoma. Medicine (Baltimore) 2016;95:e3225.
[49] Stojnev S, Ristic-Petrovic A, Velickovic LJ, Krstic M, Bogdanovic D, Khanh do T, Ristic A, Conic I, Stefanovic V. Prognostic significance of mucin expression in urothelial bladder cancer. Int J Clin Exp Pathol 2014;7:4945–58.
[50] Dhanisha SS, Guruvayoorappan C, Drishya S, Abeesh P. Mucins: structural diversity, biosynthesis, its role in pathogenesis and as possible therapeutic targets. Crit Rev Oncol Hematol 2018;122:98–122.
[51] Smith FJD, Irvine AD, Terron-Kwiatkowski A, Sandilands A, Campbell LE, Zhao Y, Liao H, Evans AT, Goudie DR, Lewis-Jones S, Arseculeratne G, Munro CS, Sergeant A, O'Regan G, Bale SJ, Compton JG, DiGiovanna JJ, Presland RB, Fleckman P, McLean WHI. Loss-of-function mutations in the gene encoding filaggrin cause ichthyosis vulgaris. Nature Genetics 2006;38:337.
[52] Rodriguez E, Baurecht H, Herberich E, Wagenpfeil S, Brown SJ, Cordell HJ, Irvine AD, Weidinger S. Meta-analysis of filaggrin polymorphisms in eczema and asthma: robust risk factors in atopic disease. J Allergy Clin Immunol 2009;123:1361–70 e7.
[53] van den Oord RA, Sheikh A. Filaggrin gene defects and risk of developing allergic sensitisation and allergic disorders: systematic review and meta-analysis. BMJ 2009;339:b2433.
[54] Skaaby T, Husemoen LL, Thyssen JP, Meldgaard M, Thuesen BH, Pisinger C, Jorgensen T, Carlsen K, Johansen JD, Menne T, Szecsi PB, Stender S, Linneberg A. Filaggrin loss-of–function mutations and incident cancer: a population-based study. Br J Dermatol 2014;171:1407–14.
[55] Choi J, Kim DI, Kim J, Kim BH, Kim A. Hornerin Is Involved in Breast Cancer Progression. J Breast Cancer 2016;19:142–7.
[56] Joshi R, Tawfik A, Edeh N, McCloud V, Looney S, Lewis J, Hsu S, Ogbureke KU. Dentin sialophosphoprotein (DSPP) gene-silencing inhibits key tumorigenic activities in human oral cancer cell line, OSC2. PLoS One 2010;5:e13974.
[57] Chen B, Wang J, Dai D, Zhou Q, Guo X, Tian Z, Huang X, Yang L, Tang H, Xie X. AHNAK suppresses tumour proliferation and invasion by targeting multiple pathways in triple-negative breast cancer. J Exp Clin Cancer Res 2017;36:65.
[58] Zhao Z, Xiao S, Yuan X, Yuan J, Zhang C, Li H, Su J, Wang X, Liu Q. AHNAK as a Prognosis Factor Suppresses the Tumor Progression in Glioma. J Cancer 2017;8:2924–32.
[59] Cuylen S, Blaukopf C, Politi AZ, Muller-Reichert T, Neumann B, Poser I, Ellenberg J, Hyman AA, Gerlich DW. Ki-67 acts as a biological surfactant to disperse mitotic chromosomes. Nature 2016;535:308–12.
[60] Wolins NE, Skinner JR, Schoenfish MJ, Tzekov A, Bensch KG, Bickel PE. Adipocyte protein S3-12 coats nascent lipid droplets. J Biol Chem 2003;278:37713–21.
[61] Tansey JT, Sztalryd C, Hlavin EM, Kimmel AR, Londos C. The central role of perilipin a in lipid metabolism and adipocyte lipolysis. IUBMB Life 2004;56:379–85.
[62] Shires K, Wienand K. Cancer testis antigen MAGE C1 can be used to monitor levels of circulating malignant stem cells in the peripheral blood of multiple myeloma patients. J Cancer Res Clin Oncol 2016;142:2383–96.
[63] Mattila H, Schindler M, Isotalo J, Ikonen T, Vihinen M, Oja H, Tammela TL, Wahlfors T, Schleutker J. NMD and microRNA expression profiling of the HPCX1 locus reveal MAGEC1 as a candidate prostate cancer predisposition gene. BMC Cancer 2011;11:327.
[64] Hou S, Sang M, Zhao L, Hou R, Shan B. The expression of MAGE-C1 and MAGE-C2 in breast cancer and their clinical significance. Am J Surg 2016;211:142–51.
[65] Curioni-Fontecedro A, Pitocco R, Schoenewolf NL, Holzmann D, Soldini D, Dummer R, Calvieri S, Moch H, Mihic-Probst D, Fitsche A. Intratumoral heterogeneity of MAGE-C1/CT7 and MAGE-C2/CT10 expression in mucosal melanoma. Biomed Res Int 2015;2015:432479.
[66] Luo H, Lin Y, Gao F, Zhang CT, Zhang R. DEG 10, an update of the database of essential genes that includes both protein–coding genes and noncoding genomic elements. Nucleic Acids Res 2014;42:D574–80.