BMC Bioinformatics

RESEARCH

Open Access

Integrative analysis of mutated genes and mutational processes reveals novel mutational biomarkers in colorectal cancer

Hamed Dashti[1†], Iman Dehzangi[2†], Masroor Bayati[1], James Breen[3,4,5], Amin Beheshti[6], Nigel Lovell[7], Hamid R. Rabiee[1*] and Hamid Alinejad-Rokny[8,9,10*]

*Correspondence: RABIEE@sharif.edu; h.alinejad@ieee.org
†Hamed Dashti and Iman Dehzangi contributed equally to this work
[1] Bioinformatics and Computational Biology Lab, Department of Computer Engineering, Sharif University of Technology, 11365 Tehran, Iran
[8] BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia
Full list of author information is available at the end of the article

Abstract

Background: Colorectal cancer (CRC) is one of the leading causes of cancer-related deaths worldwide. Recent studies have observed causative mutations in susceptible genes related to colorectal cancer in 10 to 15% of the patients. This highlights the importance of identifying mutations for early detection of this cancer for more effective treatments among high risk individuals. Mutation is considered as the key point in cancer research. Many studies have performed cancer subtyping based on the type of frequently mutated genes, or the proportion of mutational processes. However, to the best of our knowledge, combination of these features has never been used together for this task. This highlights the potential to introduce better and more inclusive subtype classification approaches using wider range of related features to enable biomarker discovery and thus inform drug development for CRC.

Results: In this study, we develop a new pipeline based on a novel concept called 'gene-motif', which merges mutated gene information with tri-nucleotide motif of mutated sites, for colorectal cancer subtype identification. We apply our pipeline to the International Cancer Genome Consortium (ICGC) CRC samples and identify, for the first time, 3131 gene-motif combinations that are significantly mutated in 536 ICGC colorectal cancer samples. Using these features, we identify seven CRC subtypes with distinguishable phenotypes and biomarkers, including unique cancer related signaling pathways, in which for most of them targeted treatment options are currently available. Interestingly, we also identify several genes that are mutated in multiple subtypes but with unique sequence contexts.

Conclusion: Our results highlight the importance of considering both the mutation type and mutated genes in identification of cancer subtypes and cancer biomarkers. The new CRC subtypes presented in this study demonstrates distinguished phenotypic properties which can be effectively used to develop new treatments. By knowing the genes and phenotypes associated with the subtypes, a personalized treatment plan can be developed that considers the specific phenotypes associated with their genomic lesion.

Keywords: Colorectal cancer, Cancer subtype identification, Somatic point mutations, Motif analysis, Gene prioritization, Therapeutic targets, Personalized medicine

Introduction

Cancer is the leading cause of death worldwide. Among different cancers, colorectal cancer (CRC) ranking second after lung cancer in lethality [1]. According to the Cancer Genome Atlas (TCGA), 16% of CRC samples show DNA damage mechanisms and high tumor mutational burden [1]. Accurate identification and quantification of CRC subtypes as well as understanding their underlying molecular mechanisms are essential steps to develop personalized medicine strategies that can potentially lead to effective treatment modalities and better clinical outcomes for patients. There is currently no consensus on the number of CRC subtypes. Different approaches that are proposed to classify CRCs based on mutation or gene expression have reported up to six subtypes. In 2013, Felipe De Sousa et al. [2] used a hierarchical clustering method on gene expression data from 90 CRC patients to identify three CRC subtypes. At the same time, Sadanandam et al. [3] analyzed gene expression data from a larger cohort comprising 445 patients and identified five CRC subtypes [3].

In another study on gene expression data from 566 patients, Marisa et al., applied hierarchical clustering using the Ward linkage method [4] to identify CRC subtypes. By adopting the Pearson correlation coefficient (PCC) as a distance measure on the gene expression profiles they identified six CRC subtypes [4]. Later on, Roepman et al., also used gene expression data from a cohort of 188 CRC patients, to identify three subtypes using an unsupervised clustering (hierarchical clustering) method [5].

Most recently, Guinney et al. [6] conducted a comprehensive study using gene expression data to identify CRC subtypes as a part of the International Consortium to Generate Colorectal Cancer Subtypes (CRCSC). This study identified four biologically distinct CRC subtypes called CMSs (consensus molecular subtypes [7]), each having unique clinical and molecular markers. For example, they defined CMS1 as a hyper-mutations case and CMS2 has a high copy number alteration. This study covered 87% of all studied CRC cases at ICGC as well as the other remaining uncharacterized samples. They integrated all the data from previous studies and used Random Forest (RF) algorithm (using 5972 genes) to build their ensemble model.

Genomic variations including point mutations and structural variants have been recently used to analyze human diseases as well as viruses [8–16]. There are also several studies that used point mutations and pathways instead of gene expression for cancer subtype identification [6, 12, 17–20]. Somatic mutations are widely known as important factors for different cancers. Different mutational processes and genes related to a given cancer are typically linked to distinct mechanisms of tumor development, from which subtypes can be identified. The stable nature of somatic mutations makes them good candidates for cancer subtype classification. Kuijjer et al., used mutation data sets of 6406 samples from 23 cancer types from the Cancer Genome Atlas (TCGA). They hypothesized that mutation processes in cancer can be identical in different cancer types regardless of the tumor origin which is also referred to as Pan-Cancer analysis. Their analysis identified nine subtypes across all cancers [12].

Different mutational processes can lead to somatic point mutations related to cancers. In 2015, Alexandrov et al. [21] showed that there are 30 mutational signatures in cancers, most of them associated with a specific molecular mechanism to uncover the causality behind somatic point mutations across the genome. The proposed concept

demonstrated the importance of motif context in the analysis of somatic point mutations in cancer genomics. It is also known that transcription factors bind to their specific sequence motifs. Therefore, sequence context of cancer mutations can potentially change the expression and regularization of genes including cancer driver genes. Hence, quantifying somatic mutations with respect to their gene-motifs instead of genes can provide greater discriminatory power for identification of cancer subtypes.

To the best of our knowledge, the context of mutations in highly mutated genes has not been used for cancer subtype classification and therapeutic biomarker identification. Based on the merits associated with somatic mutations, in this study we propose a new pipeline using "gene-motifs" concept to accurately identify CRC subtypes. An overview of the pipeline proposed in this study is presented in Fig. 1. To build our model, we hypothesize that subtypes can be more efficiently identified using mutated genes and the mutated sequence motifs within the genes, simultaneously. We refer to this combination as "gene-motifs". Here we first download publicly available somatic point mutations of 536 whole genome sequencing CRC individuals from the ICGC. We then used the FANTOMCAT genes list to identify the number of mutations in coding and lncRNA genes. After that, we used negative-binomial and beta-binomial distributions to identify significantly mutated genes and gene-motifs respectively to avoid bias and reduce noise in our data. Using 3131 candidate gene-motifs as features for a model-based clustering

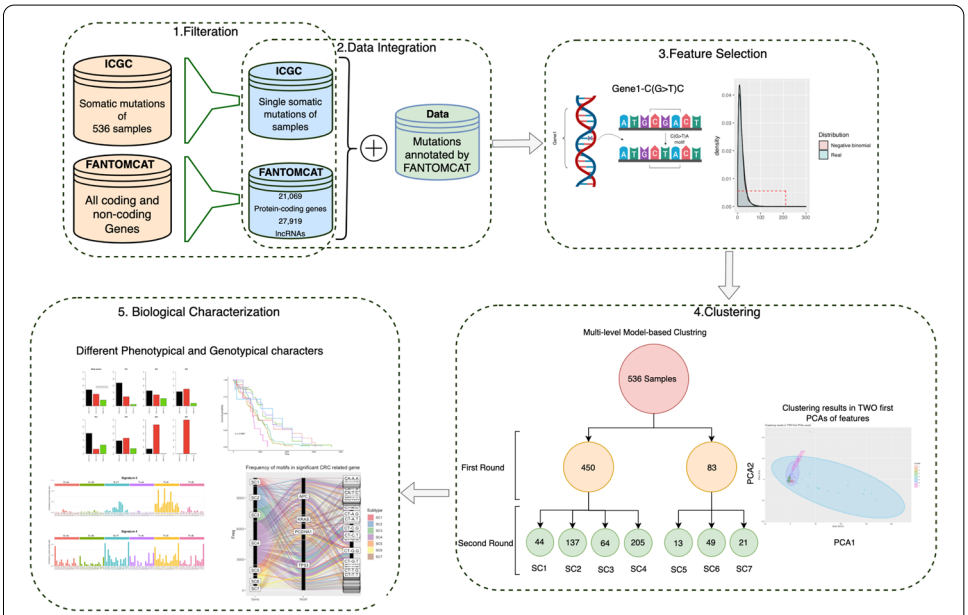

Fig. 1 An overview of our proposed pipeline. Here we first downloaded publicly available somatic point mutations of 536 whole genome sequencing CRC individuals from the ICGC. We then used the FANTOMCAT genes list to identify the number of mutations in coding and lncRNA genes. After that, we used negative-binomial and beta-binomial distributions to identify significantly mutated genes and gene-motifs, respectively. Using 3131 candidate gene-motifs as features of our model-based clustering, we identified seven CRC subtypes. Our comprehensive biological analysis showed that the identified subtypes have different mutational load in different genes. Our mutational signature analysis also showed that different combinations of signatures can be observed in each subtype. We also identified genes and conserved motifs that significantly mutated in each subtype. Finally, we performed gene ontology, pathway, and survival curve analyses

technique [22], we identify seven CRC subtypes. Our comprehensive biological analysis shows that the identified subtypes have different mutational load in different genes. Our mutational signature analysis also shows that different combinations of signatures can be observed in each subtype. We also identify novel colorectal-associated genes and specific mutated motifs, which are represented in different CRC subtypes. We also clearly demonstrate how some of the subtypes are region-based supporting the findings presented in [23].

Results

To build our unique clustering framework, we used somatic point mutations from the International Cancer Genome Consortium (ICGC) [24], which contains data from 536 patients (45% female, 55% male) with CRC across three projects (READ-US, COAD-US, COCA-CN). Typically somatic point mutations are annotated within gene-coding regions exclusively. Therefore, we also extended annotations of somatic point mutations to 27,919 lncRNAs and 21,069 protein-coding genes annotated in the FANTOM CAGE associated transcriptome (FANTOM CAT) dataset [7].

Statistical workflow to identify significant genes and gene-motifs

CRC samples have different mutational frequencies with diverse mutational profiles (Additional file 2: Figure S1). To cluster these samples, we first identified candidate genes and gene-motifs (as the feature for our clustering) that were mutated at a significantly higher level in CRC samples relative to other cancers. For this purpose, we used a Cullen and Frey graph [25] with 100-fold bootstrapping to identify best-fitted distribution to our candidate genes data. Among different distributions, the negative-binomial distribution demonstrated the best-fitted distribution (Additional file 2: Figure S2), which was validated by the Cramer-Von Mises test (using R package "VGAM" [26]—*see the method section*). Using a negative-binomial distribution as the most suitable distribution to calculate the *P*-value of mutational load for each gene, we identified 382 protein coding genes that were significantly mutated (P-value < 0.01) in the CRC samples. The rankings of the significant coding genes based on their *P*-value (i.e. candidate genes) are provided in Additional file 1: Table S1.

Gene-motif concept

Having significantly mutated genes identified, we then tested different models to identify significantly mutated gene-motifs. The Gene-motifs refer to the sequence motifs mutated within a gene. Here, we define motif as a 3-nucleotide sequence context of the mutated nucleotide, i.e. NXN-to-NYN (where X has mutated to Y, and N: A, C, G, or T). There is in total 96 different mutations within all possible motifs in DNA [21]. Here, we determine the number of each of these 96 mutation types within each gene. As described above, we investigated different models for fitting mutational profiles in gene-motifs and based on the comparison of Poisson, negative-binomial, Weibull, Gamma, log-normal and beta-binomial distributions in a Cullen and Frey Graph (Additional file 2: Figure S3), we found that the beta-binomial distribution to be the most suitable distribution to identify significantly mutated motifs within the significant genes (e.g. gene-motif). As a result, we identified 3131 gene-motifs with *P*-values smaller than the threshold

(P-value < 0.01) as significant gene-motifs, which have been used as the feature to cluster CRC patients. The list of candidate gene-motifs is provided in Additional file 1: Table S2. An overview of the gene-motif concept is provided in Additional file 2: Figure S4.

Distinguishing subtypes based on the candidate gene-motifs

We used principal component analysis (PCA) to investigate the effectiveness and validity of our selected features. As it shown in Additional file 2: Figure S5, the similar length for the first two principal components (PCs) demonstrates significant potential discriminatory information (the variance of data along these features is almost the same) that can be obtained using this distribution, indicating the importance of our selected features.

We then used multi-level model-based clustering [22] on a samples versus gene-motif matrix to identify subtypes in CRC samples. We used this technique over other clustering methods such as hierarchical clustering due to its ability to build a formal model without requiring random initialization. Traditional clustering models, such as k-means algorithm, are randomly initialized while others like hierarchical clustering are heavily dependent on heuristics to build their model. Additionally, clustering models such as dbscan [27], hdbscan [28], and local shrinking-based clustering [29] require the user to specify the optimal number of clusters or other parameters. Whereas Model-based clustering used here does not have such requirements, enabling us to find the most suitable number of clusters based on the nature of the available data. To build this model, we first applied clustering to all patients to produce the first level of clusters and then applied the model-based clustering on each cluster separately to detect if each cluster can in turn be divided into new meaningful groups (i.e. clusters with considerable number of patients; 1% of the total number of samples). We repeated this process until no more meaningful clusters were generated (see more details in the method section).

Using the described approach, we identified two clusters with 450 and 83 patients at the first clustering level. These two clusters were then divided into smaller clusters, but still distinguishable, in the second clustering level. At this level, the first cluster was divided into four sub-clusters with 44, 137, 64 and 205 samples while the second cluster was divided into three clusters with 13, 49 and 21 samples. However, this pattern did not continue to the third level of clustering. At the third level, most of our clusters did not break down into smaller clusters. For example, in two cases (137-sample and 205-sample clusters) the clusters divided into a large cluster and few outliers. These outliers were close to the larger cluster and far from the rest of the clusters. In these cases, instead of introducing a new cluster, we investigated those samples as outliers to the same cluster. As a result, our iterative clustering method identifies seven subtypes namely, SC1, SC2, SC3, SC4, SC5, SC6 and SC7 for CRC based on somatic point mutations that were determined on 3131 candidate gene-motifs. The patients in each cluster are shown in two dimensional PCAs of features in Additional file 2: Figure S6 and listed in Additional file 1: Table S3. Interestingly, a clustered heatmap of the fraction of mutated gene-motifs in each subtype shows that most of the gene-motifs are subtype specific (Fig. 2). The frequency of 3131 gene-motifs in each subtype is provided in Additional file 1: Table S4. For the rest of this paper, we will comprehensively investigate biological interpretability of our identified subtypes.

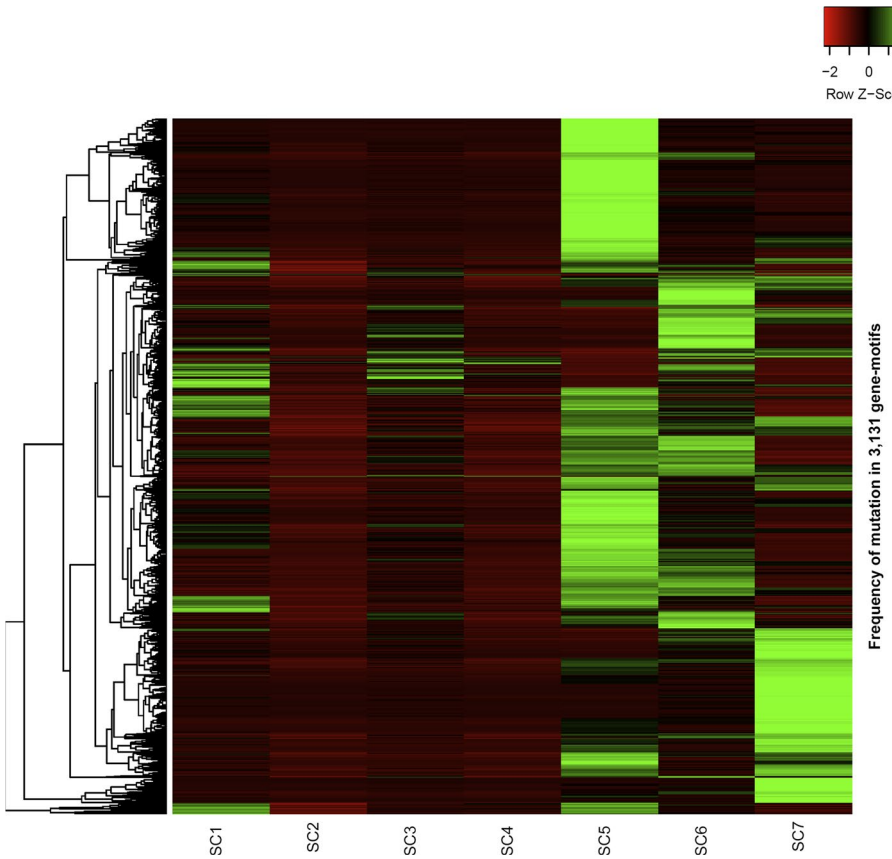

Fig. 2 Clustered heatmap of 3131 gene-motifs. Hierarchical clustering of 3131 gene-motifs in seven CRC subtypes based on their mutational frequency. Intensity represents the frequency of mutation in each subtype

Biological interpretation of CRC subtypes to identify subtype-specific biomarkers

In this section, we comprehensively investigate the biological characterization of each CRC subtype to identify subtypes-specific genes, specific conserved motifs, and gene pathways which underlie colorectal cancer.

Gene association as the biomarker of each subtype

After determining each subtype, we comprehensively investigated the biological interpretability of each subtype by identifying subtype-specific genes. The 382 candidate genes revealed that many of them have been previously associated to CRC and/or other cancers. For example, we identified *PCDHGA6*, *TTN, TP53, FRG1BP, AP2A2, MUC6*, *H3BNH8, DOCK3, CACNA1D, SERPINB, ZBED6, ZBED6, NXPH1, NKAIN2,* and *EIF3H* genes that are known to be associated with CRC cancer [15, 16, 30–33]. We also used Fisher's exact test to identify subtype-specific genes and also those genes that significantly mutated in multiple subtypes. As a result, we identified many common and subtype-specific genes, some of which are known to be associated with CRC or other cancers (Additional file 1: Table S5). Interestingly, this analysis enables us to identify unique mutated genes in a specific subtype, in which none of the other subtypes have mutation in these genes (Additional file 1: Table S5).

For example, *PCDHGA6* and *TTN* appeared as two of the significant genes in SC1. *PCDHGA6* was mutated in 42 of 44 patients of SC1 samples (95%), while mutated in 144 patients of 489 patients (29%), in other clusters. Most of the *PCDHG* family (e.g. *PCDHGB3* and *PCDHGA6*) are calcium-dependent cell-adhesion proteins [34], significantly mutated in SC1 patients and previously reported as diagnostic markers for cervical cancer [35]. Wang et al. reported that some of the *MCDH* family can be used as the diagnostic markers for cervical cancer [35]. The suitable guide for treatment as well as determining biomarker for this type of cancer may be prepared by assessment of *PCDH* family genes. *TTN,* which produces the very large protein called titin that has several functions within sarcomeres, was significantly mutated in above 90% of patients in SC1. One of the main function of this gene is to provide structure, flexibility, and stability of cell structures, which is associated to cancer [36]. We also found out that 90% of SC1 patients have been mutated in *CATG00000081433*, which is a FANTOMCAT-specific coding gene. Importantly, this gene encompasses multiple colorectal cancer associated GWAS (Genome-wide association study) SNPs. This gene is also dynamically expressed in an induced pluripotent stem cell model of neuron differentiation dataset [7].

TP53 is the most mutated gene in SC2, which has been previously associated with a variety of cancers, including CRC [31]. In addition, adenomatous polyposis coli (*APC)* was also found within this sub-cluster that is known as driver gene for colorectal, pancreatic, desmoid, hepatoblastoma, and glioma cancers [13]. *APC* is a tumor suppressor gene that encodes a multi-functional protein that is involved in several processes, including regulation of β-catenin/Wnt signalling, intercellular cell adhesion, proliferation, apoptosis and differentiation [37].

In SC3, *FRG1BP*, a gene previously associated with prostate cancer [38], was mutated in ~ 50% of samples in SC3 compared to less than 15% of samples in other clusters. *AP2A2* is another important gene that mutated in 54% of SC3 samples and in less than 23% of samples who are in other subtypes. *AP2A2*, as a component of the adaptor protein complex 2 (AP-2), is involved in clathrin-dependent endocytosis in which proteins are incorporated into vesicles surrounded by clathrin (clathrin-coated vesicles) [39]. More recently, a role for *AP2A2* has been suggested in asymmetric cell division and self-renewal of hematopoietic stem and progenitor cell [40]. Another important gene is *MUC6* which has mutated in 54% of SC3 samples and less than 24% in other samples. *MUC6* is shown to be associated with pancreatic ductal carcinoma and small intestine cancer [41]. It is also involved in "Defective *GALNT12* causes colorectal cancer 1 (*CRCS1*)" and "Defective *GALNT3* causes familial hyperphosphatemic tumoral calcinosis (*HFTC*)" pathways based on the Reactome Pathways database [42].

H3BNH8 is the most mutated gene in SC4 samples (mutated in 61% of SC4 samples). This gene is an important paralog of APC (homologous genes that have diverged within one species); these genes arise during gene duplication where one copy of the gene receives a mutation that gives rise to a new gene with a new function, though the function is often related to the role of the ancestral gene. *TP53* and *APC* are other associated genes to this subtype. These genes are also significantly mutated in SC2 samples.

DOCK3 is mutated in 100% of SC5 samples while it is mutated in only 5% of samples in other subtypes. *DOCK3* is known as a modifier of cell adhesion (MOCA) and presenilin-binding protein (PBP) [43, 44], which has an important role in melanoma [45].

CACNA1D is another cancer related gene which is Calcium voltage-gated channel subunit alpha1 D [46] that is identified as an associated gene in samples of SC5. *SERPINB4* is another important cancer-related gene, which is mutated in 53% of SC5 samples and no mutation in other subtypes. The protein of *SERPINB4* can inactivate granzyme M, an enzyme that kills tumor cells [47]. *SERPINB4* has been previously identified as associated gene with loci 18q deletion syndrome and squamous cell carcinoma [48]. *ZBED6* is another important gene which has been mutated in 63% of SC5 and almost no mutation in other subtypes. *ZBED6* is known as a repressor of *IGF2* that influences cell proliferation and development that affects cell cycle and growth of human CRC cells [49].

ZNF717 as one of the important genes that mutated in 85% of SC6 samples and only in 10% of samples in other subtypes. This gene involves many cell activities such as cell proliferation, differentiation and apoptosis, and in regulating viral replication and transcription [50]. *ZNF717* has been recently suggested as a candidate gene in the African-American population with CRC [50]. Interestingly, 90% of SC6 samples mutated in genes that are related to immunoglobulins such as the *IGHM* and *IGHV3* families.

Finally, we found *NXPH1* is mutated in all SC7 samples and mutated only in 2% of samples in other subtypes. This protein forms a very tight complex with alpha neurexins, a group of proteins that promote adhesion between dendrites and axons [51]. In addition, we identified 100% of SC7 samples also mutated in protein coding gene *CATG00000086870*, which was recently discovered by the FANTOMCAT consortia. Interestingly, this gene mutated in less than 1% of samples, who are in other subtypes. Moreover, using FANTOM5 expression atlas [52] we found this gene is highly expressed for colorectal tissue. *NKAIN2* is known as a tumor suppressor in Chinese prostate cancer [53]. This gene also mutated in all 21 Chinese samples of SC7 but only 2% of samples in other subtypes. *EIF3H* as a CRC potential diagnostic biomarker [54] is mutated in all SC7 samples but in only 3% of samples in other subtypes. *EIF3H* is a cancer-related gene and suggested as a CRC potential diagnostic biomarker [54]. We provided a ranked list (based on significant *P*-value) of the candidate genes in each subtype as the Additional file 1: Table S5.

Mutational signature

Alexandrov et al. identified that mutational signatures in cancer genomes demonstrate different biological mechanisms [21, 55]. They identified 30 distinct mutational signatures of all cancers (mostly from whole exome sequencing) in which four (i.e. 1, 5, 6 and 10) are associated with CRC [21]. These CRC mutational signatures are associated to age of patients (signature 1), defective DNA mismatch repair (signature 6), and altered activity of the error-prone polymerase POLE (signature 10) [56]. To find which mutational processes are active in the CRC subtypes, we used the CANCERSIGN tool [57] to identify mutational signatures (*see the method section*). Using whole genome sequencing data from ICGC CRC samples, we identify seven signatures overall for the CRC patients (Fig. 3a).

Using the Pearson correlation, we calculated the correlation between identified signatures in our study and Alexandrov's signatures (in Additional file 2: Figure S7). As shown in this figure, there are strong correlations between the Alexandrov's signatures, and those signatures identified in this study. A linear combination of signatures 1 and 3 in

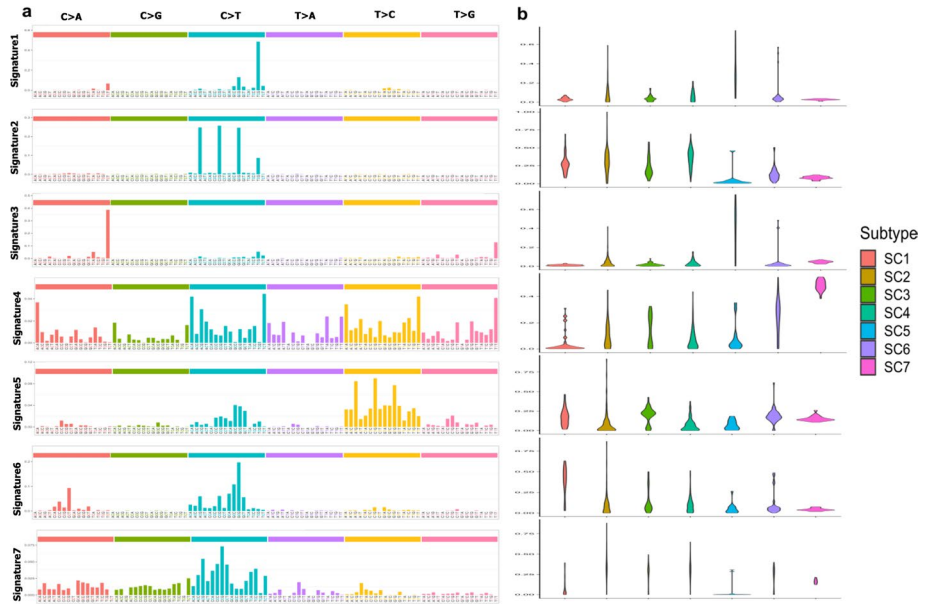

Fig. 3 Cancer mutational signature analysis. We used the CANCERSIGN tool to identify mutational signatures in each subtype. Using CANCERSIGN and WGS somatic mutation form the CRC samples, we identified seven mutational signatures where were very similar with those identified by Alexandrov et al. [21] except the signatures 1 and 3 that the linear combination of them can represent the signature 10 COSMIC (Alexandrov et al.). The exposure matrix for identified signatures are also provided as the Additional file 1: Table S11. The evaluation plot of deciphering 3-mer mutational signatures also provided in Additional file 2: Figure S11

our study (Fig. 3a) represented COSMIC signature 10 which is also associated with CRC. Signature 2 in our study was a common signature between all subtypes (Fig. 3b) and Signature 4 in our study is similar to signature 5 of Alexandrov (associated to CRC but no known etiology), while it has a greater exposure in SC5. In addition, Signature 5 in our study is similar to signatures 16 and 26 Alexandrov which are not associated to CRC. The etiology of signature 16 is unknown, while signature 26 is possibly associated with defective DNA mismatch repair [56]. This signature has greater exposure in SC4 and SC5. Signature 6 in our study is similar to signature 6 Alexandrov, which is associated with CRC. This signature has greater exposure in SC3 and SC7 and possibly is associated with defective DNA mismatch repair [56]. Finally, signature 7 in our study is similar to signature 19 Alexandrov and is associated with pilocytic astrocytoma. Note that the exposure of signature 7 is greater in SC2, SC3 and SC6 (see Fig. 3b for more details). A detailed investigation of mutational load in coding and non-coding genes is also presented and discussed in the next section.

Mutational rates of protein coding and lncRNA genes

We then investigated the mutation rate in protein-coding genes and lncRNAs of the clusters, which can provide important insight into their overall impact on different CRC subtypes. This can help us to understand if our identified subtypes are different with respect to their mutational load in coding and non-coding genes. As shown in Additional file 2: Figures S8 and S9, the mutation rate in coding and non-coding genes are different between subtypes. Interestingly, our analysis shows that SC5 is a hyper-mutated

subtype (Additional file 2: Figure S8). In addition, as shown in Additional file 2: Figure S10, the mutational rate of long non-coding genes in subtype SC7 are much higher than other subtypes, suggesting a potential role for long non-coding RNA genes in this subtype.

We also investigate the difference between our identified subtypes with respect to the mutational load in different functional consequence or variant types (*see the method section*). Interestingly, we observe that mutations in some of the subtypes are enriched in different consequence types (Additional file 2: Figure S11). For example, missense variants have a greater fraction in SC1, SC2, SC3 and SC4, while intron variants have a greater fraction in subtypes SC5, SC6 and SC7 (Additional file 2: Figure S11). Interestingly, there is almost no missense variant in SC7, and our analysis also shows that most of the mutations in SC7 are intron variant (57.7%) and intergenic region variant (21.6%). Moreover, unlike other subtypes, there is no synonymous variant in SC7. In SC6, most of the mutations are in intron and downstream gene. A summary of this analysis is presented in Table 1.

Gene-motif differential analysis in each subtype

Our gene association analysis revealed that there are several genes that are significantly mutated in multiple subtypes. This observation encouraged us to investigate if mutations in these genes have different motif preferences in each subtype. To do this, we investigated the mutations in tri-nucleotide motifs in the significant genes for each subtype. We found a considerable number of genes were significantly mutated in multiple subtypes, and that the mutations happened in different tri-nucleotide motifs. An example of this phenomena is *TP53* which was identified as a significant gene in both SC1

Table1 Consequence types of mutations in the CRC subtypes

Consequence	SC1	SC2	SC3	SC4	SC5	SC6	SC7
3_prime_UTR_variant	3.08	3.21	2.5	2.94	2.94	2.28	0.38
5_prime_UTR_premature_start_codon_gain_variant	0.19	0.16	0.18	0.23	0.16	0.14	0.02
5_prime_UTR_variant	0.66	0.7	0.82	0.69	0.77	0.64	0.11
downstream_gene_variant	19.94	17.54	18.68	17.79	15.17	17.22	9.26
exon_variant	10.81	10.69	9.68	10.82	8.66	8.03	0.91
initiator_codon_variant	0	0.01	0	0	0	0	0
intergenic_region	0.18	0.06	1.45	0.15	2.67	5.56	21.58
intragenic_variant	0	0	0	0	0	0	0
intron_variant	14.65	15.46	24.61	12.63	27.78	31.56	57.41
missense_variant	23.4	25.98	16.5	25.95	20.87	11.97	0.38
splice_acceptor_variant	0.32	0.23	0.26	0.31	0.2	0.15	0
splice_donor_variant	0.33	0.36	0.27	0.43	0.11	0.14	0.01
splice_region_variant	1.18	1.22	1.04	0.94	1.16	1	0.06
start_lost	0.03	0.03	0.02	0.03	0.01	0.02	0
stop_gained	1.21	1.96	0.88	2.03	2.71	0.79	0.02
stop_lost	0.02	0.01	0.01	0.01	0.03	0.01	0
stop_retained_variant	0.01	0.02	0.01	0.01	0.01	0.01	0
synonymous_variant	10.74	10.56	8.96	11.9	5.92	7.02	0.22
upstream_gene_variant	13.26	11.8	14.14	13.13	10.83	13.47	9.64

and SC5. However, the gene was identified to have different gene-motif patterns in each of these subtypes. The preference tri-nucleotide motif of mutations in SC1 is CT-G.G, however, the preference tri-nucleotide motif of mutations in SC5 is CT-T.G. *KRAS* is another example where most of its mutations occurred in CT-A.C and CT-G.C for SC1; CA-A.C and CT-A.C for SC2; CT-A.C, CA-A.C, CT-G.C for SC3; CT-G.C, CA-A.C, CT-A.C for SC4; CT-G.C, CT-G.T, and TG-A.T for SC5; CT-A.C, CT-G.C, CA-A.C for SC6; and CT-A.C, CA-A.C, CA-C.A for SC7 (Fig. 4a). A full list of genes that significantly mutated in multiple subtypes but in different tri-nucleotide motifs, is provided in Additional file 1: Table S6.

Mutation rate in transcripts

We then investigated the difference between our identified subtypes with respect to the mutational load in different transcripts of the coding genes. Our analyses revealed that for many of the candidate coding genes, the mutations occurred in specific transcripts of the genes (Fig. 4b and Additional file 2: Figure S12). Interestingly, 68.11% of coding genes were significantly mutated in multiple subtypes. However, the mutations enriched in different transcripts of the genes (Additional file 1: Table S7). For example, gene *TTN* is significantly mutated in all CRC subtypes except SC7. Interestingly, 62% of samples in SC5 mutated in transcript *ENST00000436599* which is much higher than other subtypes (7%, 1%, 6%, 5% and 12% for subtypes SC1, SC2, SC3, SC4, and SC6, respectively). In addition, for gene *PCDHA2* that is significantly mutated in subtypes SC1, SC5, and SC7 we observe different patterns. The De Novo mutations for this gene in SC5 are enriched in ENST00000378132 and ENST00000520672, while in SC1 and SC7 the mutations are enriched in transcript ENST00000526136 (Additional file 2: Figure S12). We provide a

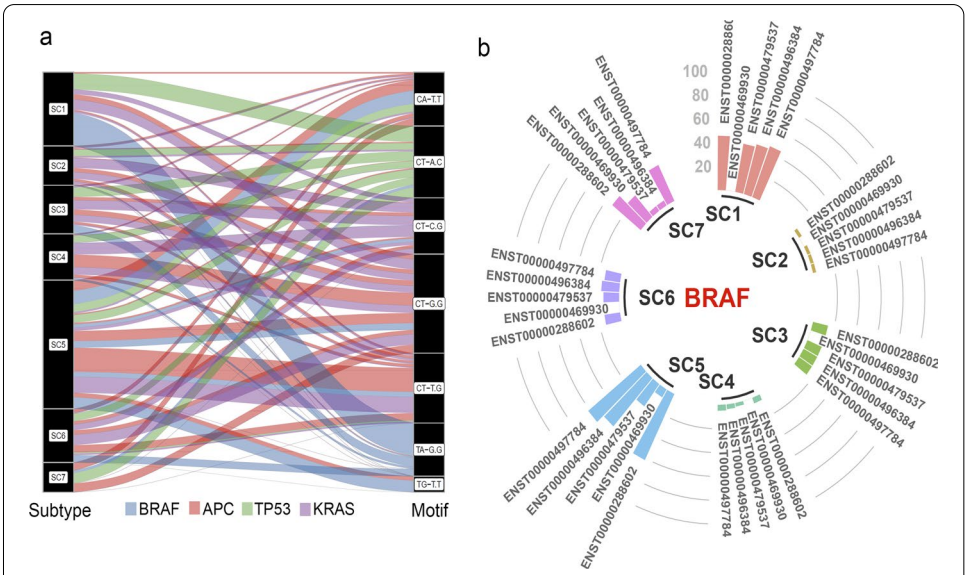

Fig. 4 Enrichment of mutations in different motifs and transcripts. **a** Investigating the mutations in tri-nucleotide motifs of genes that become significant in multiple subtypes revealed a motif context specificity of mutations in each subtype. For example, *TP53* becomes a significant gene in both SC1 and SC5. However, the preference tri-nucleotide motif of mutations in SC1 is CT-G.G, while in SC5, it is CT-T.G. **b** Our analysis also shows that for some of the candidate genes, the mutations are enriched in a specific transcript of the gene

list of candidate genes with enrichment of De Novo mutations in different transcripts of the gene in Additional file 1: Table S7.

Gene ontology and pathway analyses

Next, we conducted gene ontology (GO) and gene pathway analyses to investigate whether candidate genes in the identified subtypes are associated with any specific gene ontology or pathway term [58–60]. To do this, we chose 50 highly significant genes in each subtype and performed GO and pathway analyses using WebGestalt [61] (*see method section*) for each subtype. As shown in Fig. 5a several unique and common GO terms have been identified for each subtype. For example, "cellular response

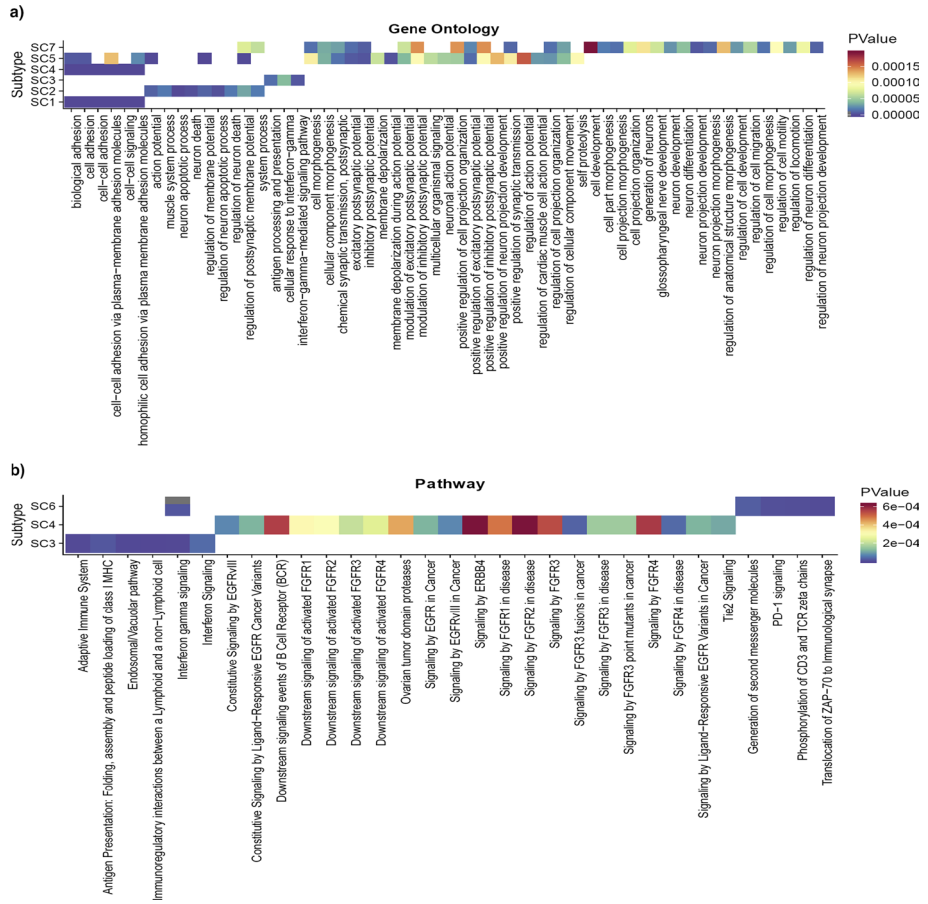

Fig. 5 Gene ontology and pathway analysis. **a** Gene Ontology analysis. Applying the WebGestalt tool [61] on 50 highly mutated coding genes in each subtype revealed several unique and common gene ontologies in the CRC subtypes. For example, "cellular response to interferon-gamma (GO:0071346)" is only associated to SC3. Also "Cell development (GO:0048468)" and "regulation of cell motility (GO:2000145)" ontologies are uniquely associated to SC7. A full list of ontology terms that were significantly associated with each subtype is provided in Additional file 1: Table S9. **b** Gene pathway analysis. Our gene pathway analysis also revealed 30 pathways that were associated with subtypes SC3, SC4, and SC6. Again, we identified several unique pathways associated to different subtypes. For example, SC4 has a unique pathway called "Signaling by FGFR3 point mutants in cancer" that has been previously associated with urothelial, breast, endometrial, squamous lung cancers and ovarian cancer [62]. More details about gene pathway analysis is provided as the Additional file 1: Table S9

to interferon-gamma (GO:0071346)" was only associated with SC3, "cell development (GO:0048468)" and "regulation of cell motility (GO:2000145)" unique to SC7, and "self-proteolysis (GO:0097264)" uniquely associated with SC5. Conversely, "cell adhesion (GO:0007155)" is an example of common gene ontologies enriched in subtypes SC1, SC4, and SC5, and "regulation of postsynaptic membrane potential (GO:0060078)" common in SC2, SC5, and SC7. A full list of ontology terms that significantly associated with each subtype is provided in Additional file 1: Table S8. In addition, pathway analysis also identified 30 pathways that are associated with subtypes SC3, SC4, and SC6 (Fig. 5b). Similar to GO terms, we also identified several unique pathways associated with different subtypes. For instance, SC4 has a unique pathway called "Signaling by *FGFR3* point mutants in cancer" that has been previously associated with urothelial, breast, endometrial, squamous lung cancers, and ovarian cancer [62]. Additionally, "Interferon gamma signaling" are represented in SC3 and SC6, which can be used in immunotherapy for CRC [63]. More details are provided as the Additional file 1: Table S9.

Clinical report and survival analysis

We also examined clinical data, such as gender, ethnicity, and regions that were available for a subset of the CRC samples. Our findings demonstrate interesting patterns in our data allocated in each subtype. For example, 52.3% of the patients in SC7 are female while 76.9% and 60.9% of patients are male in SC5 and SC3, respectively (see more detail in Fig. 6a and Additional file 1: Table S10). Geographical differences were also observed in each subtype. For instance, 21 patients in SC7 are from China and 68.2% of patients in SC1 are from USA (e.g., COAD project; Fig. 6b). Additionally, our analysis also showed

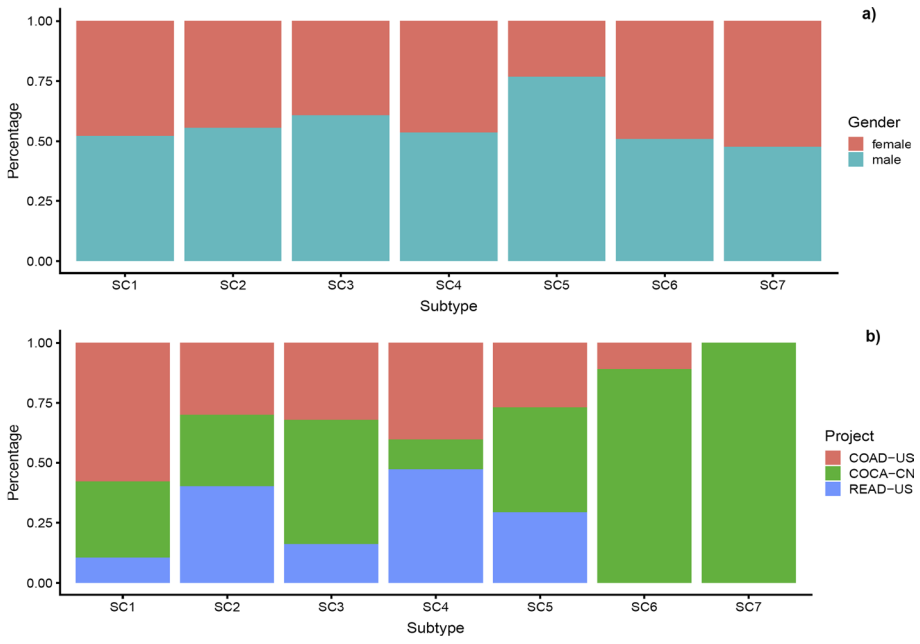

Fig. 6 Gender and geographical distributions of the patients in each subtype. **a** Gender distributions. Our gender analysis shows that 52.3% of the patients in SC7 are female while 76.9% and 60.9% of patients in SC5 and SC3 are male, respectively. **b** Geographical distributions. Our geographical distribution analysis shows that 21 patients in SC7 are from China and 68.2% of patients in SC1 are from USA (e.g. COAD project)

that samples in SC5 are older than those in SC2. This observation is shown by the histogram plot for each subtype based on the age in Additional file 2: Figure S13. We also observed that most of the samples in SC4 had cancer, at the proximal and distal of the colon, especially in the rectum and cecum and ascending colon sites.

Considering the molecular data for a subset of CRC samples, we also performed a survival analysis using "survival" [64] and "survminer" [65] R packages. In many cases, we had the data of how many days patients were alive after joining this study for sample collection. We created survival curves for all subtypes using the Kaplan–Meier method [66] and compared the survival curves using log-rank test and recorded those with *P*-value less than 0.0001. Our survival analysis illustrates that SC3 patients have more chance to survive and SC7 has a short survival duration compared to the other subtypes. The Kaplan–Meier plot for survival and the *P*-values of a log rank test has been shown in Fig. 7 and Additional file 1: Table S12.

Discussion

Classification of cancers into subtypes is essential because it informs personalized medicine therapeutic and/or prevention strategies. The CRC Subtyping Consortium introduced four consensus molecular subtypes (CMS1–4) for CRC, based on gene expression profiles of tumors. Recently, a study by Lochlan et al. [18] investigated genome-scale DNA methylation in a large cohort of CRC patients and revealed five distinct CRC subtypes of colorectal adenocarcinomas with different therapeutic biomarkers for CRC treatment. This highlights the need for a better and more inclusive subtype classification approach to enable biomarker discovery and thus inform drug development for CRC.

Mutation is the hallmark of cancer genome, and many studies have reported cancer subtyping based on the type of frequently mutated driver genes [12, 21, 67], or the

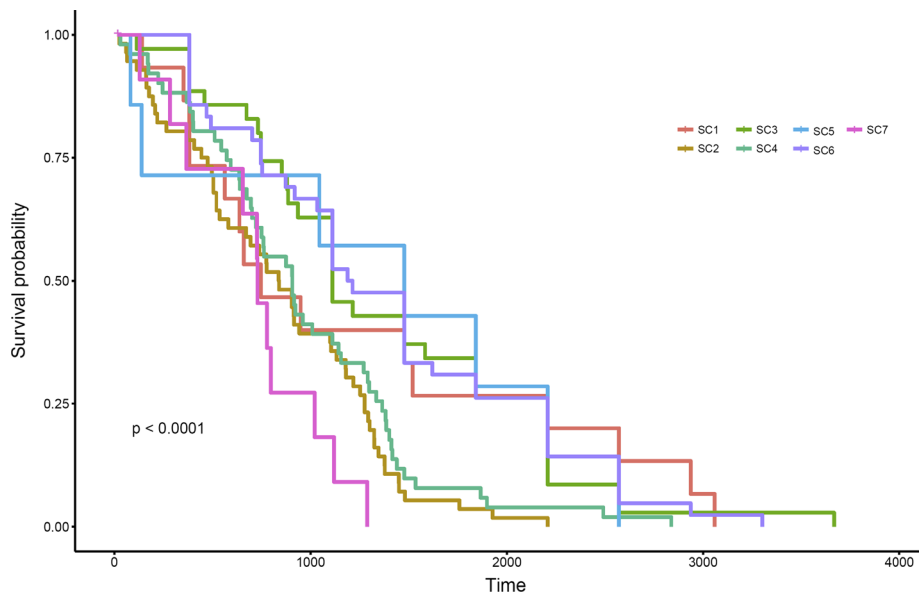

Fig. 7 Survival analysis. Using the Kaplan–Meier method we compared survival curves of the CRC subtypes. The analysis illustrated that SC3 patients have more chance to survive and SC7 have short survival duration compared to the other subtypes

proportion of mutational processes [21], however, none of these existing cancer subtyping methods consider these features simultaneously. In other words, the sequence context of somatic point mutations in driver genes have not been taken into consideration in cancer subtyping and biomarker discovery. Here, we integrated these two features (frequently mutated genes and sequence context of mutated sites) and implemented a bioinformatics pipeline for CRC subtyping using the ICGC whole genome and whole exome somatic point mutations. Our analyses revealed 3131 driver gene-motifs, which were mutated at a significantly higher level in CRC samples compared to other cancer samples. Using 3131 gene-motifs as features, we identified seven unique CRC subtypes, which are associated with distinct symptoms and genotypes in CRC. As Fig. 2 shows some of the gene-motifs are subtype specific. For example, we found that 43% of SC7 samples had ACG-to-ATG mutations in *HDAC9*, while other subtypes had no mutation in this motif within *HDAC9*. 40% of SC5 samples had TCG-to-TTG mutations within *HDAC9*, but no other subtypes had a significantly elevated mutation in this gene-motif (i.e. maximum rate of mutation for other subtypes in this gene-motif was 5%). Another example is *BRAF* for which we found that 41% of samples in SC1 had GTG-to-GAG mutations. This is significantly higher than other CRC subtypes (1%, 11%, 2%, 0% 8%, and 0% for SC2 to SC7, respectively). Interestingly, none of the samples from SC5, which is a hypermutated subtype, had mutations in *BRAF* (Additional file 1: Table S3). Importantly, in some cases, most of the subtypes are significantly mutated in a gene, but these mutations are enriched in different trinucleotide contexts. For example, Grb2-associated binder 1 (*GAB1*) is a docking protein, which is strongly implicated in CRC. This gene is significantly mutated in SC1, SC3, SC5, SC6 and SC7. However, the mutations are enriched in CCG-to-CTG for SC1 (31%), ACT-to-ATT for SC3 (100%), TCT-to-TAT for SC5 (35%), TTG-to-TCG for SC6 (67%), and GCA-to-GTA for SC7 (100%). These data suggest that, the sequence context of mutations within driver genes (frequently mutated genes) contains invaluable information that can be used to accurately identify CRC subtypes.

Our investigation of coding and lncRNA genes also revealed that there is a different mutational load for coding and lncRNA genes in CRC subtypes (Additional file 2: Figures S8 and S9). Our analysis indicated that 74% of cancer mutations fall within lncRNAs, yet their role in driving tumorigenesis is not well understood. Our data showed that mutational load in some lncRNAs can be as high as protein coding genes (Additional file 2: Figure S10). Specifically, for the CRC subtype SC7, which includes tumors with less mutated genomes, the average rate of mutations for coding and lncRNA genes are very similar (Additional file 2: Figure S10). Similar to protein coding genes, somatic mutations in highly mutated lncRNAs are also sequence context specific. For example, we show that the well-known CRC-associated lncRNA *TTN-AS1* is highly mutated in all subtypes, except SC7. We found that mutations CTG->CGG, CTG->CAG, and CAT->CTT are enriched in SC1, SC3, and SC5, respectively (Additional file 1: Table S6).

Additionally, our analyses revealed that for 68.11% of candidate coding genes, somatic mutations were enriched in different transcripts of the genes (Fig. 3b; Additional file 2: Figure S12, and Additional file 1: Table S7). To our knowledge, this result represents the first account of transcript-specific mutations in cancer and may provide a new insight into cancer driver mechanisms.

We found that, on average, all CRC subtypes had small levels of mutational signatures 1 and 3 (Fig. 3a), both of which reported for CRC patients [56]. Our data indicate that only a small subset of tumors within each subtype show highly increased levels of these two types of mutations. It has been suggested that the mutational process underlying these signatures (a combination of these signatures represents COSMIC signature 10) is altered activity of the error-prone polymerase POLE [56]. We also found that mutational signature 2 in our study (Fig. 3b) is common among all CRC subtypes. This signature is reported to be associated with an endogenous mutational process derived by spontaneous deamination of 5-methylcytosine. In contrast to Signatures 1, 2, and 3, some of the other signatures showed significantly different levels in different CRC subtypes, suggesting different molecular mechanisms behind each subtype. For example, Signature 5 showed greater exposures in SC4 and SC5. There is no known etiology for this signature. Signature 6 exhibited greater exposures in SC3 and SC7. This signature is likely associated with defective DNA mismatch repair [56]. Altogether, some but not all the differences present in tumors in terms of the extent of mutational processes are reflected in our subtypes.

With genomic medicine emerging as a routine part of the health system, tumors mutational profiling will help to better understanding of the underlying genetic causes of cancers. The current treatments options are usually based on assessing single gene mutations. Our study and the proposed pipeline to identify CRC subtype associated genes and the context of the mutations within these genes (either by identifying gene-motifs or mutation signatures) could help more precise diagnoses by assigning patients to available therapeutic targets or ongoing clinical trials targeting specific mutations; and identifying subtype-specific pathways that might be useful treatment targets for therapeutic intervention.

Some of the pathways we identified in the CRC subtypes, including cell cycle, apoptosis and DNA damage response, are frequently seen and mutated in cancer. However, each CRC subtype may have still their specific targeted treatment options as subtype specific pathways also seen for each subtype. For example, we identified "Signaling by *FGFR3* point mutants in cancer" as potential targets for treatment of the SC4 subtype. *FGFR3* signaling has been previously associated with multiple cancers [62]. In addition, several cancer-related pathways for which targeted treatment options are available are specifically mutated in our identified subtypes detailed in Additional file 1: Table S9. By considering somatic mutations in all of the genes associated with our CRC subtype-specific mutation of signaling pathways, we might be able to find additional cancer patients who could benefit from targeted treatment options. One the one hand, the pathways identified in our analysis are mutated in a large number of CRC patients, and on the other hand, targeted treatment options are currently available for most of these pathways. We therefore believe that these treatment options are worthy for further investigation to develop better therapeutic targets.

Our survival analyses show that subtype 7 (SC7) is associated with a poor outcome (Fig. 7). There are 100 gene-motifs (from 13 different genes) and at least 50% of samples in SC7 showed significantly elevated mutations in these gene-motifs, while other subtypes had no mutation (or showed very mutations with low frequencies) in these gene-motifs. Interestingly, most of these 13 genes (*KMT2C, DPP6, PRIM2, SDK1, PTPRN2,*

CNTNAP2, CSMD1, CTNNA3, MAGI2, EYS, HLA-DRB1, DLGAP2, FANK1) are known to be associated with cancer. For example, somatic mutation in Cub and Sushi Multiple Domains 1 (*CSMD1*) gene is associated with an early age of presentation in CRC individuals [68]; or mutation in *PTPRN2* (Receptor-type tyrosine-protein phosphatase N2) can promote metastatic and cellular migration in breast cancer cells through lipid-dependent sequestration of an actin-remodeling factor [69]. Our analysis also showed that SC2 and SC4 had very similar survival curves. Importantly, the gene-motif profiles of these two subtypes are also very similar (Fig. 2). For example, all samples in both SC1 and SC4 subtypes had mutation in the processed transcript *XXbac-BPG254F23*. Interestingly, none of samples in other subtypes had mutation in *XXbac-BPG254F23* (Additional file 1: Table S6).

In-silico techniques have been recently used widely to investigate biomarkers including genomics biomarkers [10, 11, 17, 19, 70–80]. In this study, we developed an in-silico technique in integration with novel gene-motif concept to study mutational patterns in colorectal cancer. Both the proposed method and the gene-motif concept have the potential to study genomic biomarkers and potential therapeutic options in other cancers.

Conclusions

In conclusion, with genomic medicine emerging as a routine part of the healthcare system and cancer individuals now frequently sequenced for specific sets of mutations, a large amount of whole genome and whole exome sequencing samples have become available for analysis. Accurate classification of cancer individuals with similar mutational profiles may help clinicians to accurately identify individuals who could receive the same types of treatment. Here, by application of a unique statistical pipeline and a novel "gene-motif" concept, and integration of somatic mutations from the ICGC consortium, we have identified seven subtypes with strong biological characterization in CRC. Our list of subtype-specific genes provides a better understanding of the underlying genetic causes of CRC. More importantly, for the first time we provided a system-wide analysis of the enrichment of de novo mutations in a specific motif context of the genes in CRC. By knowing the genes and motif associated with the mutations, a personalized treatment plan can be developed that considers the specific motif context of mutations within responsible genes. Lastly, the pipeline developed in this study is freely available and will be useful in the analysis of cancers and in narrowing down the causative genes within each cancer.

Method and material

Overall design

The flowchart of our pipeline is shown in Fig. 1. At the first step, mutations are annotated with the FANTOMCAT robust gene list. In the next step, we find a distribution representing somatic mutations data and compute the *P*-values for each gene and gene-motif based on negative-binomial and beta-binomial distributions, respectively and identify significant genes and gene-motifs using a standard threshold of $P\text{-value} < 0.01$. Then, we use model-based clustering to identify subtypes in the CRC samples based on

significant gene-motifs. Finally, we identify biomarkers and interpret biological characterization in each subtype.

ICGC dataset

We used the ICGC dataset [24], which contains data from 19 types of cancers including CRC. It contains 536 patients with CRC across three projects (READ-US, COAD-US, COCA-CN) from China and USA where 45% are female and 55% are male. 96% of these data were captured by whole genome sequencing (WGS) technology and the rest of them by non-NGS technology. We have downloaded the mutational profiles of CRC samples in a "vcf" format from ICGC data portal. But in short, raw sequencing data were processed using bcl2fastq tool, and then mapped to the human reference genome hg19 using BWA-mem v0.7.Sa tool. GATK BQSR and Haplotype Caller v3.4.46 and Strelka v1.0.14 were also used to call somatic mutations.

FANTOMCAT coding and non-coding genes list

We used a robust gene list of the FANTOMCAT consortium which contains 21,069 protein-coding genes and 27,919 lncRNAs. FANTOMCAT used cap analysis of gene expression (CAGE) technology to conduct high-confidence coding and non-coding genes. These genes are sturdily approved by transcription initiation evidence. Therefore, we decided to annotate mutations based on this dataset instead of other consortiums.

Statistical pipeline to identify significant genes

We identify coding and non-coding genes that significantly mutated in the colorectal samples in the following manner. We first count the number of samples that has mutation in each gene. We next use a negative-binomial distribution as the best fit distribution for our data (Additional file 2: Figure S2) to identify significant genes. Finally, we calculate the *P*-value for each gene using the following formula:

$$P(x > k) = 1 - \sum_{i=r}^{k} P(x = i) = 1 - \sum_{i=r}^{k} \binom{i-1}{r-1} p^r q^{k-r} \tag{1}$$

Using the mentioned pipeline, we identify 382 coding genes (listed in the Additional file 1: Table S1) that significantly mutated in the colorectal samples. We also identify beta-binomial distribution as the best fit compared to other methods to identify gene-motifs that significantly mutated in the colorectal samples. Additional file 2: Figure S3 shows the fit diagrams of different distributions on the gene-motifs. We then compute the *P*-values for each gene-motif as follow:

$$P(x > k) = 1 - \sum_{i=r}^{k} P(x = i) = 1 - \sum_{i=r}^{k} \binom{n}{i} \frac{B(i+\alpha, n-i+\beta)}{B(\alpha, \beta)} \tag{2}$$

where $B(\alpha, \beta)$ is the beta function. Using the mentioned pipeline, we identify 3131 significantly mutated motifs within 382 significant coding genes (listed in Additional file 1: Table S2). We call this concept "gene-motif". Note that we use the R package "VGAM" [26] to perform the above analyses.

Clustering

Here we use model-based clustering to identify subtypes in colorectal samples. Model-based clustering is one of the density-based unsupervised machine learning methods which is non-parametric. Hence, it does not consider any assumption for the number of available clusters. This feature is suitable here, as the number of subtypes of CRC is unknown and it is better to do clustering without any assumptions. From a statistical perspective, each sample is independent with respect to its number of mutations. Hence, via the central limit theorem we expect that if subtypes are correctly identified, the mutational load in determined features come from a Gaussian distribution. Model-based clustering automatically fits the best Gaussian distribution to samples, and finds the optimal number of clusters [29]. Here, we use model-based clustering in a hierarchical manner. Using the clustering method, we identify two clusters with 450 and 83 patients at the first clustering level. These two clusters in turn are divided into smaller and still distinguishable subgroups in the second clustering level.

Biological characterization

Mutational signature analysis

We also use CANCERSIGN [57] to identify signatures that are represented in our CRC samples. CANCERSIGN uses a non-negative matrix factorization (NMF) algorithm to identify optimal number of signatures. As shown in Additional file 2: Figure S14, in overall, we identify seven signatures in CRC samples. The exposure matrix for the identified signatures are also provided in Additional file 1: Table S11.

Gene, lncRNA, and transcript rates analysis

We use Fisher's exact test to identify coding and lncRNA genes that significantly mutated in each subtype. To do this, we make a contingency table as follow: number of CRC samples with mutation in the gene; number of CRC samples that had no mutation in the gene; number of samples with other cancer with mutation in the gene; number of samples with other cancers that had no mutation in the gene. We do the same process for transcripts to identify transcripts that significantly mutated in each subtype.

Consequence type of mutations

The consequence type of mutations from the ICGC dataset are used to count the consequence type of each mutation. We then calculate the relative frequency of consequence types in each subtype.

Gene ontology analysis and gene pathway analysis on the significantly mutated coding genes

We use the gene ontology analysis tool WebGestalt to observe the over-representation of gene ontology and pathways associated with significant genes in each subtype,

separately. We use default values for WebGestalt parameters (FDR < 0.05, 2000 permutation, minimum gene = 5). We also use the top 50 genes associated with each subtype as the candidate genes.

Literature search for non-coding interacting genes

Our literature searches were focused on human studies and English language publications available in the PubMed, Scopus, and Web of Science. Both Medical Subject Headings (MeSH) terms and related free words were used in order to increase the sensitivity of the search. We also used data and text mining techniques to extract additional related studies [76, 79, 81–90]. A decision tree approach and a knowledge-based filtering system technique have been also used to categorize the texts from the literatures search [91, 92]. The search terms included "noncoding RNA" or "lncRNAs" or "genes name + cancer". "BC" or "breast carcinoma" and "breast neoplasm".

Clinical information

We downloaded clinical data for the CRC samples from ICGC (http://cancer.digitalslidearchive.net). Two metadata files "sample.tsv" and "donor.tsv" have been used for analysis of gender and age of patients.

Survival analysis

We use the Kaplan–Meier method to conduct survival curves for all subtypes. We use "survival" [64] and "survminer" [65] R packages to conduct Kaplan–Meier curves and obtain the significance of survival prediction for subtypes. Long-rank test was also applied to obtain the *P*-value for survival analysis.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04652-8.

Additional file 1: Table S1. List of 382 significantly mutated genes that were identified by negative-binomial distribution. **Table S2.** List of 3131 significantly mutated gene-motifs that were identified by beta-binomial distribution. **Table S3.** List of CRC patients that were grouped in each subtype using a model-based clustering. **Table S4.** Frequency of mutation in 3131 significantly mutated gene-motifs in each subtype. **Table S5.** Subtype-specific genes that were identified by a Fisher exact test. Rare mutated genes in each subtype also highlight by green color. **Table S6.** List of protein-coding genes that become a significant gene in different subtypes but with enrichment of mutations in different motif context. **Table S7.** Fraction of mutations in different transcripts of candidate genes in each subtype. **Table S8.** Gene ontology analysis from WebGestalt tool [61] on 50 highly mutated coding genes in each subtype. We only considered those ontology terms with FDR < = 0.05. **Table S9.** Gene pathway analysis from WebGestalt tool on 50 highly mutated coding genes in each subtype. We only considered those terms with FDR < = 0.05. **Table S10.** Gender analysis on CRC samples in each subtype. **Table S11.** The exposure matrix for identified signatures in our CRC samples. **Table S12.** The data of survival probability with recurrence score.

Additional file 2: Figure S1. Mutation rates of the CRC patients. This plot shows number of mutations in each CRC sample. **Figure S2.** Identify best-fitted distribution to discover significant genes. This plot shows our comparison of different distribution techniques to fit the number of mutations in the genes and identify significantly mutated genes. **Figure S3.** Identify best-fitted distribution to discover significant gene-motifs. This plot shows our comparison of different distribution techniques to fit the number of mutations in the gene-motifs and identify significantly mutated gene-motifs. **Figure S4.** An overview of gene-motifs concept. We first identify 382 significantly mutated coding genes in colorectal cancers (candidate genes). We then used Fisher exact test to identify those motifs that significantly mutated within candidate genes. **Figure S5.** Selected 3131 features in two most significant PCAs before scaling. This plot shows two principal components (PCs) that demonstrates the potential discrimination that can be obtained from our identified features. **Figure S6.** Illustration of patients in two first PCAs of features. Distribution of CRC samples through 3131 gene-motif features by PCA analysis. **Figure S7.** Correlation between our identified signatures and Alexandrov's signatures in each CRC subtype separately. **Figure S8.** Mutational load of protein coding genes in each subtype separately. Each bar chart shows fraction of samples with mutation in a gene. **Figure S9.** Mutational load of long non-coding RNA genes in each subtype separately. Each bar chart shows fraction of samples

with mutation in a lncRNA. **Figure S10.** Mutation rates in coding and lncRNA genes in each subtype. Red color indicates average number of mutations in lncRNA genes and green color indicates average number of mutations in coding genes. **Figure S11.** Consequence type analysis. Figure shows fraction of mutations in different consequence types for each subtype. **Figure S12.** *Mutation rate in transcripts in genes TTN, PCDHA2, BRAF, APC*. Figure shows mutational rate in different transcripts of genes *TTN, PCDHA2, BRAF, and APC* across the CRC subtypes identified in this study. **Figure S13.** Analysis of age distribution of CRC samples in the identified subtypes. **Figure S14.** Evaluation plot for deciphering 3-mer mutational signatures in the CRC samples. We used the CANCERSIGN tool [57] to identify mutational signatures in CRC samples. The evaluation plot of deciphering 3-mer mutational signatures become optimized for seven signatures.

Acknowledgements
HAR is supported by a UNSW Scientia Program Fellowship. HAR was previously a research associate in Harry Perkins Institute of Medical Research under supervision of Prof Alistair Forrest. Analysis was made possible with computational resources provided by the BioMedical Machine Learning Lab and Telethon Kids Institute Bioinformatics Serves with funding from the UNSW Scientia Program Fellowship and the Government of Western Australia. HRR is supported by National Science Foundation (INSF) Grant No. 96006077.

Authors' contributions
HAR designed the study. HD, HAR, and ID developed hypothesis and experiments (the problem definition, candidate gene-motif identification, clustering, and hierarchical model. HAR, HD, and ID wrote the manuscript; HD, HAR, ID, HRR, JB, AB, NL edited the manuscript; HD carried out the analyses including the statistical analyses, candidate gene and gene-motif identification, text mining, gene prioritization, gene ontology, and survival analysis, HRR helped with the statistical analyses, MB performed mutational signatures analysis. HD generated all figures and tables. All authors have read and approved the final manuscript.

Funding
This work was supported by the UNSW Scientia Program Fellowship and the Australian Research Council Discovery Early Career Researcher Award (DECRA) under grant DE220101210 to HAR. This work was also supported by a grant to Prof Alistair Forrest from an Australian Research Council Discovery Project grant (DP160101960) and WA Department of Health Near-Miss Merit Award to HAR.

Availability of data and materials
The source code has been uploaded on https://github.com/hamidrokny/CRCcancer and the somatic point mutations can be downloaded from ICGC website.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing financial and non-financial interests.

Author details
[1]Bioinformatics and Computational Biology Lab, Department of Computer Engineering, Sharif University of Technology, 11365 Tehran, Iran. [2]Center for Computational and Integrative Biology (CCIB), Rutgers University, Camden, NJ 08102, USA. [3]South Australian Health and Medical Research Institute, Adelaide, SA 5000, Australia. [4]Robinson Research Institute, University of Adelaide, Adelaide, SA 5006, Australia. [5]Bioinformatics Hub, University of Adelaide, Adelaide, SA 5006, Australia. [6]Department of Computing, Macquarie University, Sydney, NSW 2109, Australia. [7]Tyree Institute of Health Engineering and The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia. [8]BioMedical Machine Learning Lab, The Graduate School of Biomedical Engineering, UNSW Sydney, Sydney, NSW 2052, Australia. [9]UNSW Data Science Hub, The University of New South Wales, Sydney, NSW 2052, Australia. [10]Health Data Analytics Program, AI-Enabled Processes (AIP) Research Centre, Macquarie University, Sydney 2109, Australia.

Received: 9 July 2021 Accepted: 24 March 2022
Published online: 19 April 2022

References

1. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330–7.
2. Felipe De Sousa EM, et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. Nature Med. 2013;19(5):614.
3. Sadanandam A, et al. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. Nat Med. 2013;19(5):619.

4. Marisa L, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. PLoS Med. 2013;10(5): e1001453.
5. Roepman P, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. Int J Cancer. 2014;134(3):552–62.
6. Guinney J, et al. The consensus molecular subtypes of colorectal cancer. Nat Med. 2015;21(11):1350.
7. Hon C-C, et al. An atlas of human long non-coding RNAs with accurate 5′ ends. Nature. 2017;543(7644):199.
8. Alinejad-Rokny H, Anwar F, Waters SA, Davenport MP, Ebrahimi D. Source of CpG depletion in the HIV-1 genome. Mol Biol Evol. 2016;33(12):3205–12.
9. Alinejad-Rokny H, Heng JI, Forrest AR. Brain-enriched coding and long non-coding RNA genes are overrepresented in recurrent neurodevelopmental disorder CNVs. Cell Rep. 2020;33(4): 108307.
10. Ebrahimi D, et al. Insights into the motif preference of APOBEC3 enzymes. PLoS ONE. 2014;9(1): e87679.
11. Gooneratne SL, Alinejad-Rokny H, Ebrahimi D, Bohn PS, Wiseman RW, O'Connor DH, Davenport MP, Kent SJ. Linking pig-tailed macaque major histocompatibility complex class I haplotypes and cytotoxic T lymphocyte escape mutations in simian immunodeficiency virus infection. J Virol. 2014;88(24):14310–25.
12. Kuijjer ML, et al. Cancer subtype identification using somatic mutation data. Br J Cancer. 2018;118(11):1492.
13. Olivier M, Hollstein M, Hainaut P. TP53 mutations in human cancers: origins, consequences, and clinical use. Cold Spring Harbor Perspect Biol. 2010;2(1): a001008.
14. Rajaei P, et al. VIRMOTIF: A user-friendly tool for viral sequence analysis. Genes. 2021;12(2):186.
15. Rowan A, et al. APC mutations in sporadic colorectal tumors: a mutational "hotspot" and interdependence of the "two hits." Proc Natl Acad Sci. 2000;97(7):3352–7.
16. Sanz-Garcia E, et al. BRAF mutant colorectal cancer: prognosis, treatment, and new perspectives. Ann Oncol. 2017;28(11):2648–57.
17. Dashti H, Dehzangi A, Bayati M, Breen J, Lovell N, Ebrahimi D, Alinejad-Rokny H. Integrative analysis of mutated genes and mutational processes reveals seven colorectal cancer subtypes. bioRxiv 2020.
18. Fennell L, et al. Integrative genome-scale DNA methylation analysis of a large and unselected cohort reveals 5 distinct subtypes of colorectal adenocarcinomas. Cell Mol Gastroenterol Hepatol. 2019;8(2):269–90.
19. Ghareyazi A, et al. Whole-genome analysis of de novo somatic point mutations reveals novel mutational biomarkers in pancreatic cancer. Cancers. 2021;13(17):4376.
20. Heidari R, et al. A systematic review of long non-coding RNAs with a potential role in Breast Cancer. Mutat Res Rev Mutat Res. 2021;787: 108375.
21. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415.
22. Scrucca L, et al. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. R J. 2016;8(1):289.
23. Kan Z, et al. Multi-omics profiling of younger Asian breast cancers reveals distinctive molecular signatures. Nature Commun. 2018;9:1725.
24. Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, Stein LD, Ferretti VT. The international cancer genome consortium data portal. Nature Biotechnol. 2019;37(4):367–9.
25. Cullen AC, Frey HC, Frey CH. Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs. Berlin: Springer; 1999.
26. Yee TW. Vector generalized linear and additive models: with an implementation in R. Berlin: Springer; 2015.
27. Ester M, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;96:226–31.
28. McInnes L, Healy J, Astels S. hdbscan: hierarchical density based clustering. J Open Source Softw. 2017;2(11):205.
29. Chang F, et al. clues: an R package for nonparametric clustering based on local shrinking. J Stat Softw. 2010;33(4):1–16.
30. Hamada T, Nowak JA, Ogino S. PIK3CA mutation and colorectal cancer precision medicine. Oncotarget. 2017;8(14):22305.
31. Iacopetta B. TP53 mutation in colorectal cancer. Hum Mutat. 2003;21(3):271–6.
32. Tan C, Du X. KRAS mutation testing in metastatic colorectal cancer. World J Gastroenterol. 2012;18(37):5171.
33. Zhang X, et al. Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T. Proc Natl Acad Sci. 2007;104(10):4060–4.
34. Wu Q, Maniatis T. A striking organization of a large family of human neural cadherin-like cell adhesion genes. Cell. 1999;97(6):779–90.
35. Wang KH, et al. Global methylation silencing of clustered proto-cadherin genes in cervical cancer: serving as diagnostic markers comparable to HPV. Cancer Med. 2015;4(1):43–55.
36. Chauveau C, Rowell J, Ferreiro A. A rising titan: TTN review and mutation update. Hum Mutat. 2014;35(9):1046–59.
37. Fodde R, Smits R, Clevers H. APC, signal transduction and genetic instability in colorectal cancer. Nat Rev Cancer. 2001;1(1):55–67.
38. Peterson LE, Kovyrshina T. Progression inference for somatic mutations in cancer. Heliyon. 2017;3(4): e00277.
39. Ohno H. Clathrin-associated adaptor protein complexes. J Cell Sci. 2006;119(18):3719–21.
40. Ting SB, et al. Asymmetric segregation and self-renewal of hematopoietic stem and progenitor cells with endocytic Ap2a2. Blood. 2012;119(11):2510–22.
41. Kaur S, et al. Mucins in pancreatic cancer and its microenvironment. Nat Rev Gastroenterol Hepatol. 2013;10(10):607.
42. Joshi-Tope G, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33(suppl_1):D428–32.
43. Kashiwa A, et al. Isolation and characterization of novel presenilin binding protein. J Neurochem. 2000;75(1):109–16.
44. Chen Q, et al. Loss of modifier of cell adhesion reveals a pathway leading to axonal degeneration. J Neurosci. 2009;29(1):118–30.
45. Sanz-Moreno V, et al. Rac activation and inactivation control plasticity of tumor cell movement. Cell. 2008;135(3):510–23.
46. Phan NN, et al. Voltage-gated calcium channels: Novel targets for cancer therapy. Oncol Lett. 2017;14(2):2059–74.

47. de Koning PJ, et al. Intracellular serine protease inhibitor SERPINB4 inhibits granzyme M-induced cell death. PLoS ONE. 2011;6(8): e22645.
48. Izuhara K, et al. Squamous cell carcinoma antigen 2 (SCCA2, SERPINB4): an emerging biomarker for skin inflammatory diseases. Int J Mol Sci. 2018;19(4):1102.
49. Ali MA, et al. Transcriptional modulator ZBED6 affects cell cycle and growth of human colorectal cancer cells. Proc Natl Acad Sci. 2015;112(25):7743–8.
50. Guda K, et al. Novel recurrently mutated genes in African American colon cancers. Proc Natl Acad Sci. 2015;112(4):1149–54.
51. Kinzfogl J, Hangoc G, Broxmeyer HE. Neurexophilin 1 suppresses the proliferation of hematopoietic progenitor cells. Blood. 2011;118(3):565–75.
52. De Rie D, et al. An integrated expression atlas of miRNAs and their promoters in human and mouse. Nat Biotechnol. 2017;35(9):872.
53. Mao X, et al. NKAIN2 functions as a novel tumor suppressor in prostate cancer. Oncotarget. 2016;7(39):63793.
54. Yu G, et al. The proliferation of colorectal cancer cells is suppressed by silencing of EIF3H. Biosci Biotechnol Biochem. 2018;82(10):1694–701.
55. Hamidi H, Alinejad-Rokny H, Coorens T, Sanghvi R, Lindsay SJ, Rahbari R, Ebrahimi D. Signatures of mutational processes in human DNA evolution. bioRxiv. 2021.
56. Helleday T, Eshtad S, Nik-Zainal S. Mechanisms underlying mutational signatures in human cancers. Nat Rev Genet. 2014;15(9):585.
57. Bayati M, et al. CANCERSIGN: a user-friendly and robust tool for identification and classification of mutational signatures and patterns in cancer genomes. Sci Rep. 2020;10(1):1–11.
58. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
59. Kohler S, et al. The human phenotype ontology in 2017. Nucleic Acids Res. 2017;45(D1):D865-d876.
60. Pinero J, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833-d839.
61. Wang J, et al. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017;45(W1):W130–7.
62. Touat M, et al. Targeting FGFR signaling in cancer. Clin Cancer Res. 2015;21(12):2684–94.
63. Slattery ML, et al. Interferon-signaling pathway: associations with colon and rectal cancer risk and subsequent survival. Carcinogenesis. 2011;32(11):1660–7.
64. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. Berlin: Springer; 2013.
65. Kassambara A et al. Package 'survminer'. 2017.
66. Bland JM, Altman DG. Survival probabilities (the Kaplan–Meier method). BMJ. 1998;317(7172):1572–80.
67. Dietlein F, Weghorn D, Taylor-Weiner A, Richters A, Reardon B, Liu D, Lander ES, Van Allen EM, Sunyaev SR. Identification of cancer driver genes based on nucleotide context. Nat Genet. 2020;52(2):208–18.
68. Shull AY, et al. Somatic mutations, allele loss, and DNA methylation of the Cub and Sushi Multiple Domains 1 (CSMD1) gene reveals association with early age of diagnosis in colorectal cancer patients. PLoS ONE. 2013;8(3): e58731.
69. Sengelaub CA, et al. PTPRN2 and PLCβ1 promote metastatic breast cancer cell migration through PI (4, 5) P2-dependent actin remodeling. EMBO J. 2016;35(1):62–76.
70. Sharma J, et al. An in-silico evaluation of different bioactive molecules of tea for their inhibition potency against non structural protein-15 of SARS-CoV-2. Food Chem. 2021;346: 128933.
71. Bhardwaj VK, et al. Evaluation of acridinedione analogs as potential SARS-CoV-2 main protease inhibitors and their comparison with repurposed anti-viral drugs. Comput Biol Med. 2021;128: 104117.
72. Bhardwaj VK, et al. Identification of bioactive molecules from tea plant as SARS-CoV-2 main protease inhibitors. J Biomol Struct Dyn. 2021;30(10):3449–58.
73. Singh R, et al. A computational approach for rational discovery of inhibitors for non-structural protein 1 of SARS-CoV-2. Comput Biol Med. 2021;135: 104555.
74. Singh R, et al. Identification of potential plant bioactive as SARS-CoV-2 Spike protein and human ACE2 fusion inhibitors. Comput Biol Med. 2021;2021(136): 104631.
75. Deif MA, Solyman AA, Kamarposhti MA, Band SS, Hammam RE. A deep bidirectional recurrent neural network for identification of SARS-CoV-2 from viral genome sequences. Math Biosci Eng. 2021;18(6):8933–50.
76. Alinejad-Rokny H, Sadroddiny E, Scaria V. Machine learning and data mining techniques for medical complex data analysis. Neurocomputing. 2018;276(1).
77. Alinejad-Rokny H, Ghavami R, Rabiee HR, Rezaei N, Tam KT, Forrest AR. MaxHiC: robust estimation of chromatin interaction frequency in Hi-C and capture Hi-C experiments. bioRxiv. 2020.
78. Javanmard R, et al. Proposed a new method for rules extraction using artificial neural network and artificial immune system in cancer diagnosis. J Bionanosci. 2013;7(6):665–72.
79. Shamshirband S, et al. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. J Biomed Inform. 2021;113: 103627.
80. Singh R, et al. In-silico evaluation of bioactive compounds from tea as potential SARS-CoV-2 nonstructural protein 16 inhibitors. J Tradit Complement Med. 2021;35(2):235.
81. Esmaeili L, et al. Hybrid recommender system for joining virtual communities. Res J Appl Sci Eng Technol. 2012;4(5):500–9.
82. Hasanzadeh E, et al. Text clustering on latent semantic indexing with particle swarm optimization (PSO) algorithm. Int J Phys Sci. 2012;7(1):16–120.
83. Hosseinpoor M, et al. Proposing a novel community detection approach to identify cointeracting genomic regions. Math Biosci Eng. 2020;17(3):2193–217.
84. Niu H, et al. An ensemble of locally reliable cluster solutions. Appl Sci. 2020;10(5):1891.

85. Parvin H, et al. A heuristic scalable classifier ensemble of binary classifier ensembles. J Bioinform Intell Control. 2012;1(2):163–70.
86. Parvin H, et al. A classifier ensemble of binary classifier ensembles. Int J Learn Manag Syst. 2013;1(2):37–47.
87. Parvin H, et al. Using clustering for generating diversity in classifier ensemble. JDCTA. 2011;3(1):51–7.
88. Parvin H, et al. An innovative combination of particle swarm optimization, learning automaton and great deluge algorithms for dynamic environments. Int J Phys Sci. 2011;6(22):5121–7.
89. Parvin H, et al. A new imbalanced learning and decision tree method for breast cancer diagnosis. J Bionanosci. 2013;7(6):673–8.
90. Sharifrazi D et al. CNN-KCL: automatic myocarditis diagnosis using convolutional neural network combined with k-means clustering. Preprints 2020
91. Mahmoudi MR, et al. Consensus function based on cluster-wise two level clustering. Artif Intell Rev. 2021;54(1):639–65.
92. Parvin H, et al. Proposing a classifier ensemble framework based on classifier selection and decision tree. Eng Appl Artif Intell. 2015;37:34–42.

Publisher's Note