BMC Bioinformatics

SOFTWARE

Open Access

gcMECM: graph clustering of mutual exclusivity of cancer mutations

Ying Hu*, Chunhua Yan*, Qingrong Chen and Daoud Meerzaman

*Correspondence:
yhu@mail.nih.gov;
yanch@mail.nih.gov
Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD, USA

Abstract

Background: Next-generation sequencing platforms allow us to sequence millions of small fragments of DNA simultaneously, revolutionizing cancer research. Sequence analysis has revealed that cancer driver genes operate across multiple intricate pathways and networks with mutations often occurring in a mutually exclusive pattern. Currently, low-frequency mutations are understudied as cancer-relevant genes, especially in the context of networks.

Results: Here we describe a tool, gcMECM, that enables us to visualize the functionality of mutually exclusive genes in the subnetworks derived from mutation associations, gene–gene interactions, and graph clustering. These subnetworks have revealed crucial biological components in the canonical pathway, especially those mutated at low frequency. Examining the subnetwork, and not just the impact of a single gene, significantly increases the statistical power of clinical analysis and enables us to build models to better predict how and why cancer develops.

Conclusions: gcMECM uses a computationally efficient and scalable algorithm to identify subnetworks in a canonical pathway with mutually exclusive mutation patterns and distinct biological functions.

Keywords: Cancer driver genes, Mutually exclusive mutations, Network analysis, R package

Background

Next-generation sequencing technology has transformed the study of the cancer genome, enabling us to sequence whole-genome or whole-exome and measure somatic mutations in millions of cancer genomes. The Cancer Genome Atlas (TCGA), a publicly-funded genomics project, houses a collection of mutation profiles from thousands of patients and more than 30 different types of cancers [1]. Today's comprehensive mutation landscape affirms the importance of identifying genes and their associated networks in the hunt for cancer driver genes. By detecting highly recurrent mutations, called "significantly mutated genes", we can more reliably predict the development and trajectory of cancer. Finding these cancer-driver genes is difficult, and many simply escape identification using existing data sets and methods. For example, in breast cancer, only three genes (*TP53*, *PIK3CA* and *GATA3*) occur at > 10% incidence; for most tumor types, the current sample size is too small to reliably detect genes mutated at 5% or less above background mutation intensity [2]. Thus, we are not able to capture a complete representation of all genes and subsets of genes that drive the development and progression of cancers. Cancer genes tend to be altered in a finite number of pathways, typically related to differentiation, cell division, survival, and genomic maintenance [3]. Therefore, it is critical to identify pathway-level implications of genes, even those mutated at very low frequencies.

One approach to finding these drivers is to search for mutual exclusivity of altered genes since mutually exclusive pairs of genes often share the same pathways. For example, we know that a set of mutated genes rarely co-occurs in the same tumor and driver mutations are typically observed in exactly one gene in the pathway per patient [4]. The phenomenon could arise from functional redundancy or synthetic lethality in cancer pathways [5]. Classical examples of mutually exclusive driver mutations include EGFR and KRAS mutations in lung cancer [6] and TP53 and MDM2 mutations in glioblastoma [7]. Based on this rationale, MEMo (Mutual Exclusivity Modules in Cancer) draws on correlation analysis and statistical tests to identify network modules exhibiting patterns of mutually exclusive genetic alterations across multiple patients [8]. A more recent method Mutex uses a large, aggregated pathway model of human signaling processes to search groups of mutually exclusively altered genes, all of which share a common downstream event [9].

The drawback with the current methods is that they require extensive filtering of mutation data, which are limited to the most significantly mutated genes, focus on predefined network modules, and do not readily scale to reasonably sized datasets [10]. The mutual exclusivity signal can be biased toward identifying gene sets where the majority of the coverage comes from highly mutated genes [11, 12]. Although cancer genes have been shown to participate in multiple pathways, few existing methods identify polymorphic gene sets where a gene has different mutual exclusivity to other genes in different pathways at various mutation frequencies. To detect the mutually exclusive mutation pattern more broadly and capture the full range of mutations more comprehensively, we developed a novel graph-based unsupervised clustering approach to identify gene sets with mutually exclusive mutations. The Graph Clustering of Mutual Exclusivity of Cancer Mutations (gcMECM) is able to detect modules of different sizes with varying cutoffs and significance. The gene sets in a module can be mapped onto one or more canonical pathways to uncover functional subnetworks that can be associated with clinical features such as survival and tumor subtypes. The algorithm uses both high and low frequency mutations and is able to analyze a large set of genes.

Implementation

Mutation and pathway data

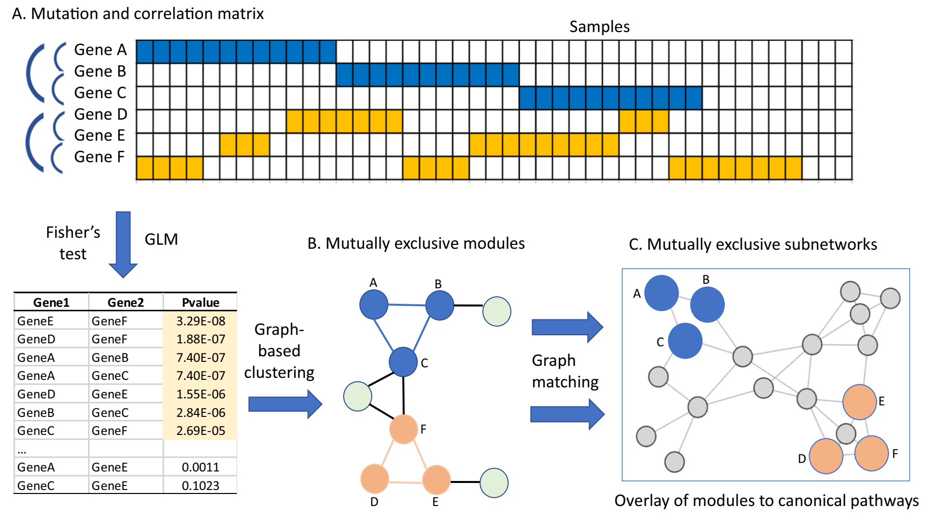
The mutation and clinical outcome datasets for TCGA-LUAD (Lung Adenocarcinoma) and TCGA-BRCA (Breast Invasive Carcinoma) in The Cancer Genome Atlas (TCGA) were downloaded from The NCI Genomic Data Commons (https://portal.gdc.cancer.gov, version 6). The missense, start_lost, stop_gained, and stop_lost mutations were included in the analysis due to the fact that single-nucleotide variants (SNVs) are the most reliable somatic mutation calls [13]. RAS pathway v2.0 is obtained from NCI Ras

Initiative (ras-pathway-v2). The pathway structure and gene coordinates were created manually for visualization. KEGG pathway images and gene relationships were from KEGG database (https://www.genome.jp/kegg).

Detection of modules with mutually exclusive mutations

gcMECM was implemented in R (http://www.R-project.org) and available on GitHub at https://github.com/CBIIT-CGBB/gcMECM. Its workflow consists of three steps (Fig. 1). First, it generates the association matrix or gene–gene adjacency distance matrix using Fisher's exact test and generalized linear models (GLM), which are performed for every pair of genes in the mutation matrix across all tumor samples. As shown in the schematic diagram (Fig. 1A), genes A-B-C and genes D-E-F have strong negative associations while the association between genes A and D is weak.

Secondly, gcMECM constructs network graphs from a set of negatively correlated genes selected from those with both negative correlations in the GLM and Fisher's exact test p value < 0.001 (Fig. 1B). This p-value is used as a distance between genes and can be set with a less stringent value to build an inclusive graph. As long as the p-value in this step is relaxed enough to retain those potential mutually exclusive mutation gene pairs, slightly different p-value selection will not affect the downstream analysis. The graph is subsequently clustered into modules by Louvain algorithm in the R package igraph (https://igraph.org). The Louvain algorithm can quickly detect modules in a large graph based on the modularity measure and a heuristic approach [14] and is one of the most popular algorithms in the biological network analysis [15]. By applying the stringent cutoff of edge values less than 1.00E−08 and 1.00E−12 in TCGA-BRCA and TCGA-LUAD respectively, genes in each module are closely related to each other with negative relationships due to the mutually exclusive mutations. The edge cutoff p-values were


A. Mutation and correlation matrix
Samples
Gene A
Gene B
Gene C
Gene D
Gene E
Gene F
Fisher's test
GLM
B. Mutually exclusive modules
Graph-based clustering
A
B
C
F
D
E
Graph matching
C. Mutually exclusive subnetworks
A
B
C
E
D
F
Overlay of modules to canonical pathways

Gene1	Gene2	Pvalue
GeneE	GeneF	3.29E-08
GeneD	GeneF	1.88E-07
GeneA	GeneB	7.40E-07
GeneA	GeneC	7.40E-07
GeneD	GeneE	1.55E-06
GeneB	GeneC	2.84E-06
GeneC	GeneF	2.69E-05
…		
GeneA	GeneE	0.0011
GeneC	GeneE	0.1023

Fig. 1 Schematic diagram of gcMECM. **A** The mutation matrix, displaying a landscape of mutation status across genes at vertical axis and samples at horizontal axis, is used to calculate gene–gene association of mutations with Fisher's exact test and generalized linear models (GLM). Arches are used to highlight subnetworks. **B** Identification of modules with graph-based clustering of negatively correlated genes. **C** Overlay of mutually exclusive subnetworks in the context of canonical pathways using the graph-matching algorithm

chosen to generate 5–10 moderate size modules, which are computationally efficient while maintain the biological function integrity. The number of genes in a moderate size module is expected to be around hundreds, similar to that in RAS and KEGG pathways. Modules with less than three genes were removed from further analysis.

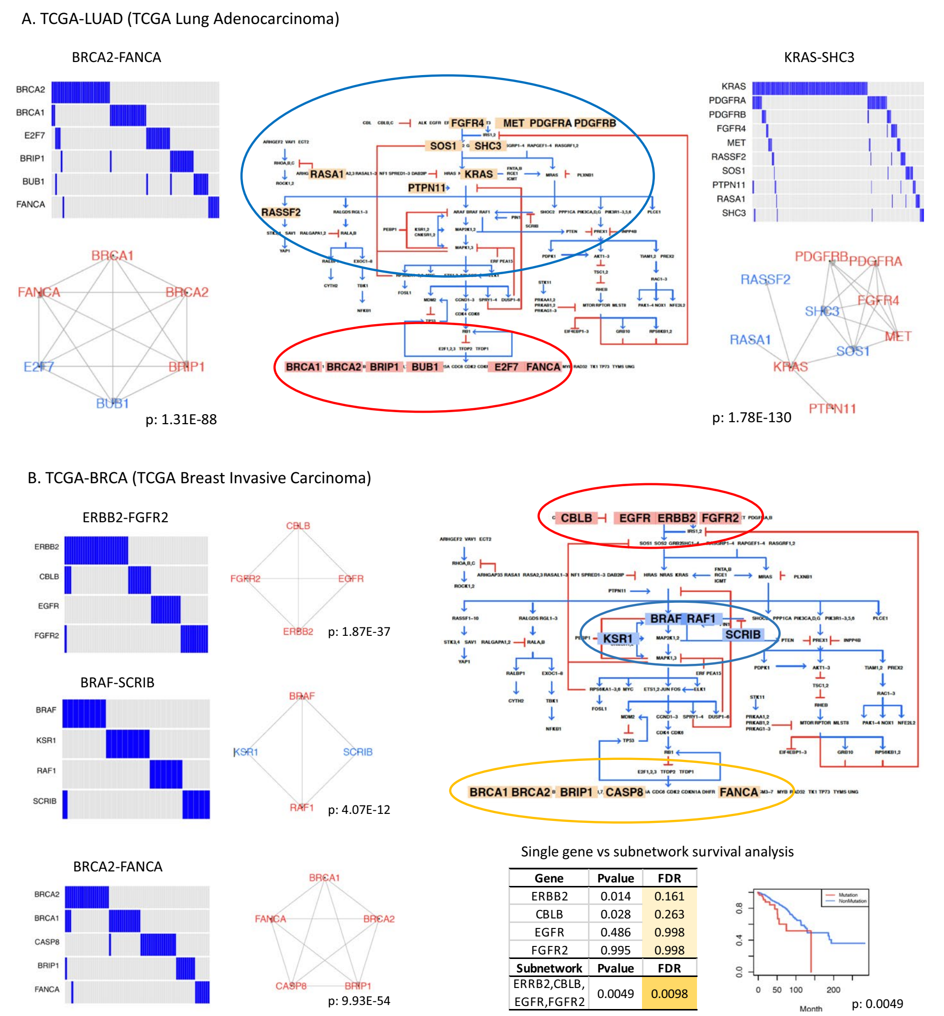
Finally, each module is compared to canonical pathways from NCI Ras or KEGG database to identify subnetworks via a graph matching algorithm in the R package igraph [16]: 1) determine subnetworks in a module consisting of the common genes between a module and a canonical pathway; 2) genes in a subnetwork must have a minimum of three genes with direct connections in the canonical pathway. This is to ensure the subnetwork from a module is localized and does not spread across the pathway. A combined p-value for a subnetwork is calculated from gene-pair Fisher's exact test p-values using combine.test function in the R package survcomp. All subnetworks in the context of pathways are visualized with the R package igraph and further examined for the enrichment with g:profiler (https://biit.cs.ut.ee/gprofiler/gost) and OmicPath (https://github.com/CBIIT-CGBB/OmicPath) against Gene Ontology and KEGG. The survival analysis is carried out with the R survival package.

Results

To demonstrate the utility and performance of gcMECM, we analyzed the missense, start_lost, stop_gained, and stop_lost mutations from TCGA breast invasive carcinoma (TCGA-BRCA) and lung adenocarcinoma (TCGA-LUAD) to identify modules with mutually exclusive mutation patterns [17, 18]. Using TCGA-BRCA data, we identified 9451 genes mutated in 985 samples. Of these, 3784 genes have negative correlations (Fisher's exact test p value $<$ 0.001). A total of 6 modules were detected with the minimal module size of 155 genes and the maximal module size of 1106 genes. Similarly, using TCGA-LUAD data, we found 12,683 genes mutated in 565 samples, with 4440 genes having negative correlation (Fisher's exact test p value $<$ 0.001). A total of 7 modules were identified with the minimal module size of 85 genes and the maximal module size of 1114 genes.

We next mapped modules from TCGA-LUAD and TCGA-BRCA onto the Ras pathway to identify subnetworks, which reduced the complexity and could be used for the detection of biologically relevant patterns. This pathway is critical in carcinogenesis and includes genes involved in oncogenic signaling, cell cycle, DNA replication, and DNA repair. Those genes are frequently altered in different cancers, including AKT1, EGFR, KRAS, and STK11 in lung cancer and AKT1, BRCA2, ERBB2, and PIK3CA in Breast cancer [19].

Two subnetworks in TCGA-LUAD, KRAS-SHC3 and BRCA2-FANCA, have been shown to have mutually exclusive mutations and more than half of those genes are present in COSMIC cancer census genes [20] (Fig. 2A). Most genes have a low mutation rate; only KRAS and PDGFRA have a mutation frequency greater than 5%. The KRAS-SHC3 subnetwork captures the upstream signaling component in the RAS pathway, which involves the ERBB signaling pathway, VEGF-PDGFR signaling pathway, and MAPK signaling pathway, as demonstrated using the Gene Ontology and KEGG analyses with g:Profiler [21]. The BRCA2-FANCA subnetwork is related to meiotic cell cycle process, cell signal transduction by p53 class mediator, DNA replication, and



Gene	Pvalue	FDR
ERBB2	0.014	0.161
CBLB	0.028	0.263
EGFR	0.486	0.998
FGFR2	0.995	0.998
Subnetwork	**Pvalue**	**FDR**
ERRB2,CBLB, EGFR,FGFR2	0.0049	0.0098

Fig. 2 Biological implication of subnetworks in the context of pathway. **A** Subnetworks from TCGA-LUAD (TCGA Lung Adenocarcinoma) are mapped to the RAS pathway. Genes colored in red are present in the COSMIC Cancer Gene Census. Blue oval: genes in KRAS-SHC3; Red oval: genes in BRCA2-FANCA. **B** Subnetworks from TCGA-BRCA (TCGA Breast Invasive Carcinoma) are mapped to the RAS pathway and associated with survival. Red oval: genes in ERBB2-FGFR2; Blue oval: genes in BRAF-SCRIB; Yellow oval: genes in BRCA2-FANCA. P-values are combined p-values from gene-pair fisher's tests. The KM-plot of survival analysis is obtained from the comparison of mutated vs. non-mutated samples for four genes in the subnetwork ERBB2-FGFR2. Coxph p-values adjusted by sex and age for both single gene and a group by subnetwork genes are shown in the table

homologous recombination. These two subnetworks with distinct biological functions suggest that the mutual exclusivity in genes with related functionality could be used to identify cancer-relevant genes, especially when the subnetworks also include well-established cancer genes.

Three subnetworks, ERBB2-FGFR2, BRAF-SCRIB, and BRCA2-FANCA, are identified in TCGA-BRCA after mapping to RAS pathway (Fig. 2B). The ERBB2-FGFR2 subnetwork is related to ERBB signaling pathway and BRCA2-FANCA subnetwork is linked to DNA repair and meiotic cell cycle process, which are similar to KRAS-SHC3 and BRCA2-FANCA subnetworks in TCGA-LUAD respectively. The BRAF-SCRIB subnetwork is found to regulate MAP kinase activity and ErbB signaling pathways that are linked to many cancers such as melanoma, lung, ovarian, breast, and prostate [22]. The mutation rate of genes in these three subnetworks is all less than 5%. Each subnetwork consists of genes with similar functions, which can be used to group samples into mutated and non-mutated categories for survival analysis. As seen in ERBB2-FGFR2 subnetwork, the survival difference is not statistically significant between two groups at the individual gene level. However, if samples are divided into two groups based on the mutation status of all genes in ERBB2-FGFR2 subnetwork, the mutation group exhibits a significantly lower survival (FDR < 0.05). These results demonstrate that integrating subnetworks of mutually exclusive mutations with pathways and clinical features can aid in interpreting the subnetwork's resulting biological functions.

Conclusions

gcMECM expands the current mutual exclusivity software capability into studying genes with low frequency mutation and data with large sample sizes through the integration of graph modules, canonical pathways, and clinical information. It uses a computationally efficient Louvain algorithm, which is readily scalable to handle a large number of genes, samples, and networks. The edge cutoff p values as parameters are subject to tuning since mutation types and mutation rates have an impact on the module and subnetwork structure. For example, the current use case consists of functional SNVs. If frameshift mutations were included, subnetwork structures would be partially altered due to the change of mutation frequency. As shown here, the mapped subnetworks are suitable for gene set enrichment analysis to identify functionally coherent gene groups which are particularly important in exploring functional redundancy or synthetic lethality in cancer pathways [5]. Depending on the canonical pathway, a single gene may be instrumental in multiple subnetworks in cancer, with very distinct biological functions. Thus, examining the subnetwork, and not just the impact of a single gene, should significantly increase the statistical power of clinical analysis.

Availability and requirements

Project name: gcMECM; Project home page: https://github.com/CBIIT-CGBB/gcMECM; Operating system(s): Platform independent; Programming language: R; Other requirements: R (>= 3.5), igraph, Rtsne; License: GNU GPL v2.0; Any restrictions to use by non-academics: No.

Abbreviations

gcMECM: Graph Clustering of Mutual Exclusivity of Cancer Mutations; TCGA: The Cancer Genome Atlas; BRCA: Breast invasive carcinoma; LUAD: Lung Adenocarcinoma; KEGG: Kyoto Encyclopedia of Genes and Genomes; GO: Gene Ontology; COSMIC: The Catalogue of Somatic Mutations in Cancer; FDR: False discovery rate; GLM: Generalized linear model.

Acknowledgements

Not applicable.

Authors' contributions

YH and CY implemented the gcMECM package. YH, CY, QC, and DM have contributed to interpreting the results and writing of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding provided by the National Institutes of Health (NIH) The authors are employees of the National Cancer Institute (NCI), National Institutes of Health. This work used the computational resources of the NIH High-Performance Computing Biowulf Cluster. The funding agencies did not play any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript. The content is the sole opinion of the authors, not of NCI.

Availability of data and materials

gcMECM R package is available at https://github.com/CBIIT-CGBB/gcMECM; Ras pathway was downloaded from https://www.cancer.gov/research/key-initiatives/ras/ras-central/blog/2015/ras-pathway-v2; TCGA breast cancer and lung cancer mutation data were obtained from https://portal.gdc.cancer.gov.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 1 February 2021 Accepted: 30 November 2021
Published online: 14 December 2021

References

1. Jensen MA, Ferretti V, Grossman RL, Staudt LM. The NCI Genomic Data Commons as an engine for precision medicine. Blood. 2017;130(4):453–9.
2. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, Meyerson M, Gabriel SB, Lander ES, Getz G. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495–501.
3. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546–58.
4. Cisowski J, Bergo MO. What makes oncogenes mutually exclusive? Small GTPases. 2017;8(3):187–92.
5. Deng Y, Luo S, Deng C, Luo T, Yin W, Zhang H, Zhang Y, Zhang X, Lan Y, Ping Y, et al. Identifying mutual exclusivity across cancer genomes: computational approaches to discover genetic interaction and reveal tumor vulnerability. Brief Bioinform. 2019;20(1):254–66.
6. Gazdar AF, Shigematsu H, Herz J, Minna JD. Mutations and addiction to EGFR: the Achilles "heal" of lung cancers? Trends Mol Med. 2004;10(10):481–6.
7. Cancer Genome Atlas Research N: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455(7216):1061–8.
8. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22(2):398–406.
9. Babur O, Gonen M, Aksoy BA, Schultz N, Ciriello G, Sander C, Demir E. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations. Genome Biol. 2015;16:45.
10. Nguyen H, Shrestha S, Tran D, Shafi A, Draghici S, Nguyen T. A Comprehensive survey of tools and software for active subnetwork identification. Front Genet. 2019;10:155.
11. Zhang J, Zhang S. The discovery of mutated driver pathways in cancer: models and algorithms. IEEE/ACM Trans Comput Biol Bioinform. 2018;15(3):988–98.
12. Dimitrakopoulos CM, Beerenwinkel N. Computational approaches for the identification of cancer genes and pathways. Wiley Interdiscip Rev Syst Biol Med. 2017;9(1):66.
13. Kroigard AB, Thomassen M, Laenkholm AV, Kruse TA, Larsen MJ. Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS ONE. 2016;11(3):e0151664.
14. Blondel VD, Guillaume J, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp. 2008;66:P10008.
15. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, Lin J, Hescott B, Hu X, Mercer J, et al. Assessment of network module identification across complex diseases. Nat Methods. 2019;16(9):843–52.
16. Lyzinski V, Fishkind DE, Fiori M, Vogelstein JT, Priebe CE, Sapiro G. Graph matching: relax at your own risk. IEEE Trans Pattern Anal Mach Intell. 2016;38(1):60–73.
17. Cancer Genome Atlas N: Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.
18. Cancer Genome Atlas Research N: Comprehensive molecular profiling of lung adenocarcinoma. Nature.2014;511(7511):543–50.
19. Li S, Balmain A, Counter CM. A model for RAS mutation patterns in cancers: finding the sweet spot. Nat Rev Cancer. 2018;18(12):767–77.
20. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019;47(D1):D941–7.
21. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, Vilo J. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res. 2019;47(W1):W191–8.
22. Croce L, Coperchini F, Magri F, Chiovato L, Rotondi M. The multifaceted anti-cancer effects of BRAF-inhibitors. Oncotarget. 2019;10(61):6623–40.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.