Impact of mutational signatures on microRNA and their response elements

Eirini Stamoulakatou[†], Pietro Pinoli[†] and Stefano Ceri[†]

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano,
Milano, Italy
[†]*E-mail: first.last@polimi.it*

Rosario M. Piro

Institut für Informatik, Institut für Bioinformatik, Freie Universität Berlin
Berlin, Germany
E-mail: r.piro@fu-berlin.de

MicroRNAs are a class of small non-coding RNA molecules with great importance for regulating a large number of diverse biological processes in health and disease, mostly by binding to complementary microRNA response elements (MREs) on protein-coding messenger RNAs and other non-coding RNAs and subsequently inducing their degradation. A growing body of evidence indicates that the dysregulation of certain microRNAs may either drive or suppress oncogenesis.

The seed region of a microRNA is of crucial importance for its target recognition. Mutations in these seed regions may disrupt the binding of microRNAs to their target genes. In this study, we investigate the theoretical impact of cancer-associated mutagenic processes and their mutational signatures on microRNA seeds and their MREs. To our knowledge, this is the first study which provides a probabilistic framework for microRNA and MRE sequence alteration analysis based on mutational signatures and computationally assessing the disruptive impact of mutational signatures on human microRNA–target interactions.

Keywords: microRNA seed, microRNA response element, mutational signatures, somatic mutations, tumor genomes.

1. Introduction

MicroRNAs are small non-coding RNA molecules which play significant roles in a variety of essential biological processes, including cell cycle regulation, differentiation, neural patterning, metabolism, aging and so on.[1] The regulatory role of microRNAs is mostly exerted by binding to complementary target sites—frequently called microRNA response elements (MREs)—on RNA molecules and subsequently inducing their degradation.[1] MicroRNAs can bind to both coding and non-coding RNAs and thus regulate their stability and expression. It has been observed that microRNA-related regulation is evolutionarily conserved.[2,3] Consequently, it is not a surprise that the dysregulation of microRNAs is associated with a range of human diseases, such as cancer, neurological disorders, cardiovascular disorders and so on.[4]

The first evidence for microRNA involvement in human cancer came in 2002, when a study about chronic lymphocytic leukemia (CLL) found that the smallest common genomic region of recurrent deletions on chromosome 13q14 harbors two microRNAs, miR-15a and miR-16-1.[5] In the past few years, an increasing number of microRNA–disease associations have been identified, but in many cases the mechanisms underlying the dysregulation of microRNAs have not yet been fully understood.

Here, we want to contribute to the study of microRNA-related dysregulation in cancer by specifically analyzing how somatic mutations may alter the microRNA seed sequences or their complementary response elements and therefore impact microRNA-target recognition.

The somatic mutations found in tumor genomes have in many cases been caused by multiple mutational processes. Both intrinsic processes such as the spontaneous deamination of 5-methylcytosine and external carcinogens like cigarette smoke or UV light have been implicated in the generation of somatic DNA changes.[6]

The spectrum of single nucleotide variants—only these mutations will be considered in the present work—associated with a particular mutational process can be mathematically represented as a so-called “mutational signature”. The most frequently used signature model, published by Alexandrov et al.,[7] conceptualizes mutational processes as vectors of 96 mutation probabilities for all possible single nucleotide variant mutation types within their context of adjacent bases. That is, each mutation type represents a nucleotide triplet whose central base is mutated, e.g., ACG>ATG or A[C>T]G. Since the two strands of double-stranded DNA are reverse complementary, mutation types are grouped if they are equivalent; A[C>T]G on one strand, for example, entails C[G>A]T on the opposite strand and *vice versa*. The catalog of mutations ultimately observed in a tumor depends on both the set of active mutational processes and the strength (“exposure”) with which these signatures contributed to the mutational load of the tumor.

In this study, we build a link between mutational signatures and their impact on microRNA activity. We specifically try to answer the question of how a particular mutational signature can potentially disrupt microRNA-mediated gene regulation. We therefore compute the theoretical impact of mutational signatures on both seed regions of mature microRNAs and their MREs on target genes. Mutating the seed region of a microRNA gene itself would of course affect its interaction with most if not all of its target genes, but also the alteration of an individual MRE may have important phenotypic effects if the disrupted microRNA–target interaction, for example, leads to the upregulation of an otherwise suppressed oncogene.

Based on the computed theoretical impact, we can predict the effective impact for different cancer types by taking into account the actual exposures of tumors to the corresponding mutational processes. However, since our main goal is to study the potential impact of mutational signatures alone, our current model does intentionally ignore effects such as selective pressure and variations in local mutation rates.

2. Methods

2.1. *Data sources*

We used the sequences of 2656 mature human microRNAs from miRBase.[8] The microRNA target sites (MREs) were obtained from TargetScan,[9] extracting only MREs of conserved microRNA families. In order to avoid including too many false-positive predictions, we filtered the data, keeping only MREs with a context++ score below -0.3 (negative scores indicate repression). For target sites, conservation was not required, since we're interested in all possible human microRNA–target interactions. We identified target sites for 2010 microRNAs.

The most recent version of triplet-based mutational signatures[10] was obtained through COSMIC.[a] We used only the 47 single base substitution (SBS) signatures which have not been identified only in exome sequencing samples (SBS 23 and 42) or characterized as possible sequencing artifacts (SBS 27, 43 and 45–60).

Somatic mutations were obtained for six whole genome sequencing (WGS) datasets with a total of 1270 primary tumor samples from the International Cancer Genome Consortium (ICGC).[11] These datasets include prostate adenocarcinoma (PRAD-CA, n=290), ER+ and HER2- breast cancer (BRCA-EU, n=569), colorectal cancer (COCA-CN, n=30), liver cancer (LIRI-JP, n=258), lung cancer (LUSC-KR, n=30), and ovarian cancer (OV-AU, n=93).

We additionally used somatic mutations for six whole exome sequencing (WES) datasets with a total of 1789 primary tumors from The Cancer Genome Atlas (TCGA) Research Network:[b] uterine corpus endometrial carcinoma (UCEC-US, n= 283), breast cancer (BRCA-US, n=248), colon adenocarcinoma (COAD-US, n=341), ovarian cancer (OV-US, n=178), stomach adenocarcinoma (STAD-US, n=320), and skin cancer/melanoma (SKCM-US, n=419).

Since we are particularly interested in mutations of MRE which are located on exonic regions of mRNAs or non-coding genes, for both WGS and WES data we took into consideration only samples with at least 100 somatic mutations falling into annotated exonic regions as defined by GENCODE release 31 for the human reference genome GRCh37.[12]

For evaluating our results we used the Human MicroRNA Disease Database (HMDD) version 3.0[13] which reports microRNA–disease associations of six categories according to different supporting evidence, including genetics research (e.g., knockdown or overexpression experiments), epigenetics research, circulating biomarker microRNAs, microRNA–target interactions (e.g, therapeutic targets), tissue expression, and other known microRNA–disease associations from the biomedical literature.

2.2. *Signature and mutational process*

Let the DNA nucleotide alphabet be represented by the set $A = \{a, c, g, t\}$ such that any genomic region r of length n corresponds to a sequence in the set A^n. Here, we will focus on nucleotide triplets $t = \langle a_1, a_2, a_3 \rangle \in A^3$ and single nucleotide variants described by 96 possible somatic mutation types $m = \langle a_1, [a_2 \rightarrow a_4], a_3 \rangle$ which mutate a triplet's central base.

We define $P_g(t)$ as the empirical probability (i.e., frequency) of observing the triplet t in the given genome g (e.g., the human assembly), such that:

$$\sum_{t \in A^3} P_g(t) = 1.$$

The actual frequency distribution of triplets within the human genome, shown in Fig. 1

[a] https://cancer.sanger.ac.uk/cosmic/signatures/SBS/
[b] https://www.cancer.gov/tcga

(blue line), evidences that the probability of occurrence is not uniform with some triplets being much more frequent than others.

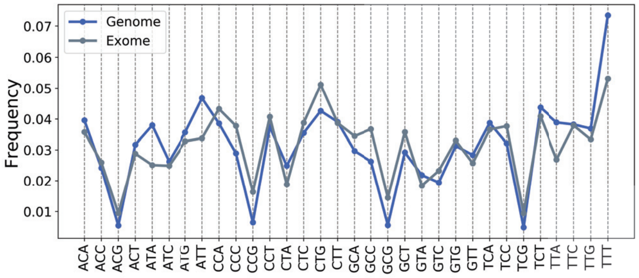

Fig. 1. Distribution of triplets in the whole human reference genome hg19 (blue) and in exonic regions alone (gray). The two curves show that (i) the distribution of triplets is not uniform; (ii) genome and exome distributions differ slightly.

Let s be the mutational signature associated with a specific mutational process (of possibly still unknown etiology). For each of the mutations generated by the process, $P_s(m)$ represents the probability that it will be of mutation type m. Consequently, if of all the N mutations present in a tumor sample N_s mutations were generated by the process associated with s, then the expected number of mutations of a certain type $m = \langle a_1, [a_2 \to a_4], a_3\rangle$ generated by s is:

$$\mathbf{E}\left[\langle a_1, [a_2 \to a_4], a_3\rangle, s\right] = N_s \times P_s(\langle a_1, [a_2 \to a_4], a_3\rangle).$$

For convenience, we define $P'_s(t)$ as the probability that a mutation induced by the mutational process of s affects a triplet $t = \langle a_1, a_2, a_3\rangle$, regardless of the precise nucleotide change:

$$P'_s(\langle a_1, a_2, a_3\rangle) = \sum_{a_4 \in A \setminus \{a_2\}} P_s(\langle a_1, [a_2 \to a_4], a_3\rangle).$$

While we are mostly interested in the latent probabilities that such mutations occur *at all* due to a given mutational process, the probability distributions P_s and P'_s indicate only of what type a mutation will likely be and what triplet will likely be affected *if* a mutation is caused. To derive the latent probabilities of interest, we can envision the generative mutational process associated with s as follows:

(1) let g be a genome composed of a multiset T of triplets with an observable triplet frequency distribution P_g;

(2) let further s be the mutational signature associated with a mutational process that mutates ("mut") a triplet $t \in T$ with latent probability P_L

$$P_L(\text{mut}|t, s) = 1 - P_L(\neg\text{mut}|t, s);$$

(note that we devise this probability as independent of the actual strength with which the mutational process contributes to a specific tumor genome, i.e., of the tumor's "exposure" to the mutational signature);

(3) the outcome generated by the process (2) on the genome's triplets (1) is a multiset of mutated triplets with an observable distribution P'_s.

The expected number of mutations caused by s which affect a given triplet t can be written as:

$$\begin{aligned}\mathbf{E}\left[t,s\right] &= N_s \times P'_s(t)\\ &= N_s \times P_g(t) \times |T| \times P_L(\text{mut}|t,s)\end{aligned}$$

where $|T| \approx |g|$ is the total number of triplets in genome g and $P_g(t) \times |T|$ the number of triplets of type t. Both expressions of $\mathbf{E}\left[t,s\right]$ embody the rationale that the observed number of t mutated by s is the product of the process's inherent probability to alter a triplet of type t and the overall strength or activity of the mutational process in the tumor sample (here represented by N_s). We hence can derive the latent probabilities of the model conveniently as:

$$P_L(\text{mut}|t,s) = \frac{P'_s(t)}{P_g(t)}\frac{1}{|T|}.$$

Here, like the distribution P'_s, also P_L is independent of the number N_s of mutations generated by s, i.e., the strength with which the mutational process contributed to the overall mutational load in the tumor sample. In contrast to P'_s, however, it explicitly accounts for the frequency distribution of triplets in the genome. For a specific mutation type $\langle a_1, \left[a_2 \rightarrow a_4\right], a_3\rangle$ we can now compute the associated latent probability as:

$$\begin{aligned}&P_L(\langle a_1, \left[a_2 \rightarrow a_4\right], a_3\rangle|s) =\\ &P_L(\text{mut}|\langle a_1, a_2, a_3\rangle, s)\frac{P_s(\langle a_1, \left[a_2 \rightarrow a_4\right], a_3\rangle)}{P'_s(\langle a_1, a_2, a_3\rangle)}\end{aligned}$$

2.3. *Impact on microRNA seeds and MREs*

Let R be the set of mature microRNA sequences. For a specific microRNA $r \in R$ we can determine possible seed sequences after mutation by artificially applying each mutation type $m_j = \langle a_1, \left[a_2 \rightarrow a_4\right], a_3\rangle$ to each corresponding sequence triplet $\langle a_1, a_2, a_3\rangle$ present in positions 2-7 of r. Assuming that every mutation in this minimum 6-mer seed region—on which the majority of functional MREs are based[14,15]—will severely affect the microRNA's target recognition, i.e., that every such mutation is deleterious, we can define a disruption score, or *impact score*, to measure the potential impact of a mutational signature s on a microRNA r:

$$I(r,s) = \frac{\sum_{i=2}^{7}\sum_{a\in A\setminus r_i} P_L(\langle r_{i-1}, \left[r_i \rightarrow a\right], r_{i+1}\rangle|s)}{6}$$

That is, for every nucleotide of the minimum 6-mer seed region, we sum the latent mutation probabilities of the three possible base changes, and take the average over all seed nucleotides as an indication of how likely the signature might disrupt the seed. Although this score is not a true disruption probability, it can be used for ranking: a signature s_1 is more likely to have a negative impact on r than a signature s_2 if $I(r,s_1) > I(r,s_2)$, and the seed of a microRNA r_1 is more likely to be disrupted by s than the seed of r_2 if $I(r_1,s) > I(r_2,s)$.

Canonical MREs contain sequences complementary to the microRNA seed nucleotides at positions 2–7. Thus, given the set of MREs Z_r of the microRNA r, we can analogously define a deleteriousness score $I(z,s)$ for an individual MRE $z \in Z_r$. By taking the reverse complement z' of the MRE z and due to the equivalence of reverse complementary mutation types, this score can be computed with exactly the same formula used for the microRNA r itself (see above). We now can define a deleteriousness score for the entire set of MREs of r:

$$I(Z_r, s) = \frac{1}{|Z_r|} \times \sum_{z \in Z_r} I(z', s)$$

where the reverse complements z' of the MREs can differ mostly in the nucleotides z'_1 and z'_8 which are adjacent to the seed region $z'_2 \dots z'_7$, i.e., seed regions of MREs can differ in their first $(\langle z'_1, z'_2, z'_3 \rangle)$ and last triplet $(\langle z'_6, z'_7, z'_8 \rangle)$, respectively.

2.4. *Signature refitting and activity estimation*

While the *impact scores* described above indicate how likely individual mutational signatures can in theory disrupt microRNA–target interactions, they do not take the actual activity or strength of the mutational process in a given tumor into account. Mutational signatures with a low to moderate impact on microRNA seeds which are highly active in a tumor may ultimately cause more mutations in seed regions than signatures with a theoretically high impact but only marginal activity in the tumor.

We therefore need to determine the strength of the different signatures in a given tumor by estimating the fraction of the mutations that have been caused by the corresponding mutational processes. These fractions are often termed "exposures",[7] sometimes also "weights" or "contributions", and the estimation task is frequently referred to as *signature refitting.*

Given a catalogue of mutational signatures S and a set of somatic mutations M found in a given tumor genome, an exposure e_k is computed for each signature $s_k \in S$, such that the exposure-weighted sum of signatures reflects the distribution of mutation types $\langle m_j \rangle_M$ observed in M:

$$\langle m_j \rangle_M \approx \sum_k e_k \times s_k \qquad \text{with} \quad \sum_k e_k = 1 \quad \text{and} \quad e_k \geq 0$$

Therefore, the exposures e_k predict what fractions of the tumor's mutations can be attributed to the signatures s_k and hence the activity or strength of the associated mutational processes.

Here, we performed signature refitting using the Bioconductor R package *decompTumor2Sig*[16] which implements a quadratic programming approach to determine the set of exposures e_k that minimizes the error between $\langle m_j \rangle_M$ and $\sum_k e_k \times s_k$.

The determined exposures or weights can be used to score the *tumor-specific impact* of mutational processes on microRNA seed regions:

$$I_{tumor}(r, s_k) = I(r, s_k) \times e_k \qquad \text{and} \qquad I_{tumor}(Z, s_k) = I(Z, s_k) \times e_k$$

Finally, we can predict which microRNA seeds are most likely to be affected in a given tumor by computing the *sum of impact scores* over all mutational signatures:

$$I_{tumor}(r) = \sum_{s_k \in S} I_{tumor}(r, s_k) \qquad \text{and} \qquad I_{tumor}(Z) = \sum_{s_k \in S} I_{tumor}(Z, s_k)$$

3. Results

3.1. *Signature refitting*

We ran *decompTumor2Sig*[16] on the somatic mutations of all individual tumor samples in order to determine the tumor-specific exposures of the single mutational signatures, i.e., the fraction of mutations with which they contributed to the mutational load of the individual tumors. Averaged results for some of the datasets are reported in Fig. 2.

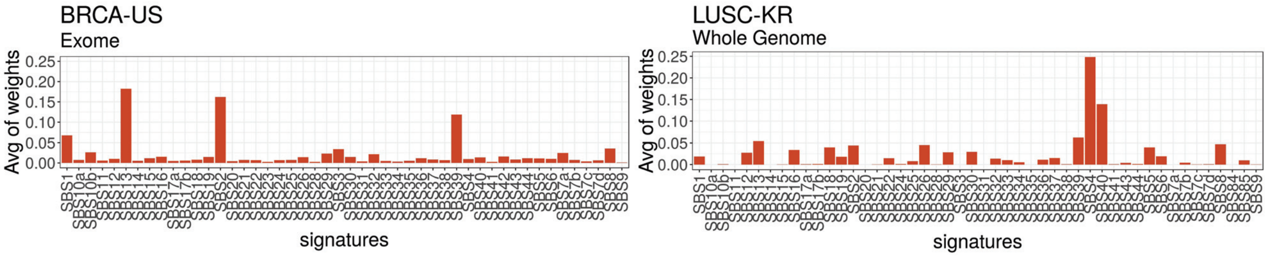

Fig. 2. The average contributions or weights (i.e., exposures) of mutational signatures for the BRCA-EU and LUSC-KR datasets. Signature refitting was done based on mutations from the whole genome but only cases with at least 100 mutations in exonic regions were considered.

Results are generally in line with previous studies.[10] Lung cancers from the LUSC-KR dataset, for example, tend to be strongly affected by signature SBS4 which is known to be associated with tobacco smoking.[6]

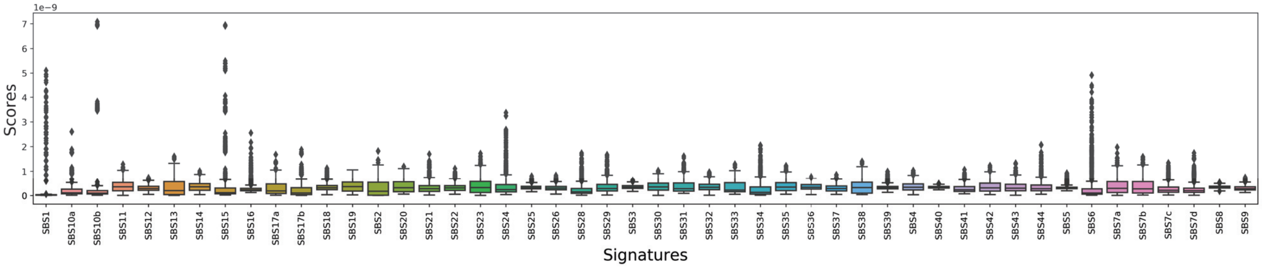

Fig. 3. Distribution of impact scores $I(r, s)$ on microRNA seeds for each mutational signature.

3.2. *Results for seed regions of microRNA genes*

Taking the WGS datasets, we first sought to evaluate which seed regions of microRNA genes might in principle be more likely to be negatively affected by the mutational signatures active in different tumor types. For this purpose we computed the sum of impact scores for each of the 2656 mature human microRNAs in each of the six WGS datasets and searched for the top scoring microRNAs.

Most notably, while we found no actual mutations in these microRNA genes, for each of the six cancer types many of the top scoring microRNA are actually known to be associated with the cancer type in question. Indeed, for colon cancer, ovarian cancer, liver cancer, breast cancer, prostate cancer and lung cancer, 32, 26, 32, 26, 25, and 28 out of the top 50 predicted microRNAs were confirmed by recent biomedical literature, respectively. Here, we report only the top 10 for two of the cancer types in Tables 1 to 2 with the PubMed IDs (PMIDs) of supporting literature as reported in HMDD.

Table 1. Top 10 scoring microRNA seeds (sum of impact scores) for breast cancer (BRCA-EU). Four microRNAs are supported by biomedical literature (max. four supporting PubMed IDs).

microRNA	Evidence (PMID)	HMDD category
miR-6869	-	-
miR-375	22400902;22952344;20978187;24746361	circulation biomarker, epigenetics
miR-1292	-	-
miR-937	-	-
miR-1307	26749252;29697201	circulation biomarker, target gene
miR-1908	-	-
miR-3178	30333478;27746365	target gene
miR-126	21249429;26261534;25844955;20801493	circulation biomarker,target gene
miR-598	-	-
miR-1306	-	-

Table 2. Top 10 scoring microRNA seeds (sum of impact scores) for liver cancer (LIRI-JP). Six microRNAs are supported by biomedical literature (max. four supporting PubMed IDs).

microRNA	Evidence (PMID)	HMDD category
miR-6869	-	-
miR-1292	-	-
miR-937	-	-
miR-375	25618599;29962816;25424171;22056881	epigenetics, circulation biomarker, target gene
miR-3178	26182877	regulation of tumorigenesis
miR-1908	-	-
miR-126	26756996;27774652;28639884;27499630	epigenetics, circulation biomarker, target gene
let-7e	17188425;28796071;21298008;23282077	genetics, circulation biomarker
let-7d	23682578;20347499;21903590	target gene
miR-1307	26646011	epigenetics

In addition, we investigated whether some signatures have a particularly high impact on microRNA seeds (Fig. 3). Signature SBS11 (related to alkylating agents), for example, has a much higher average impact score than many others. Also, signatures such as SBS1, SBS10b, SBS15 and SBS16 appear to have extremely high scores for individual microRNAs.

3.3. *Results for seed regions in MREs*

The fact that we did not directly identify mutations in the seed regions of the microRNA genes themselves is not surprising, considering that these constitute only an extremely small fraction of the human genome.

We therefore used six WES datasets and extended our search to the complementary seed regions of MREs, i.e., targets sites of the microRNAs. For each microRNA and dataset, we first computed average impact scores over all corresponding MRE seed regions, one average impact score for each mutational signature. We then computed the sum of these MRE-based impact scores over all signatures, weighting them according to the signatures' average exposures in the dataset.

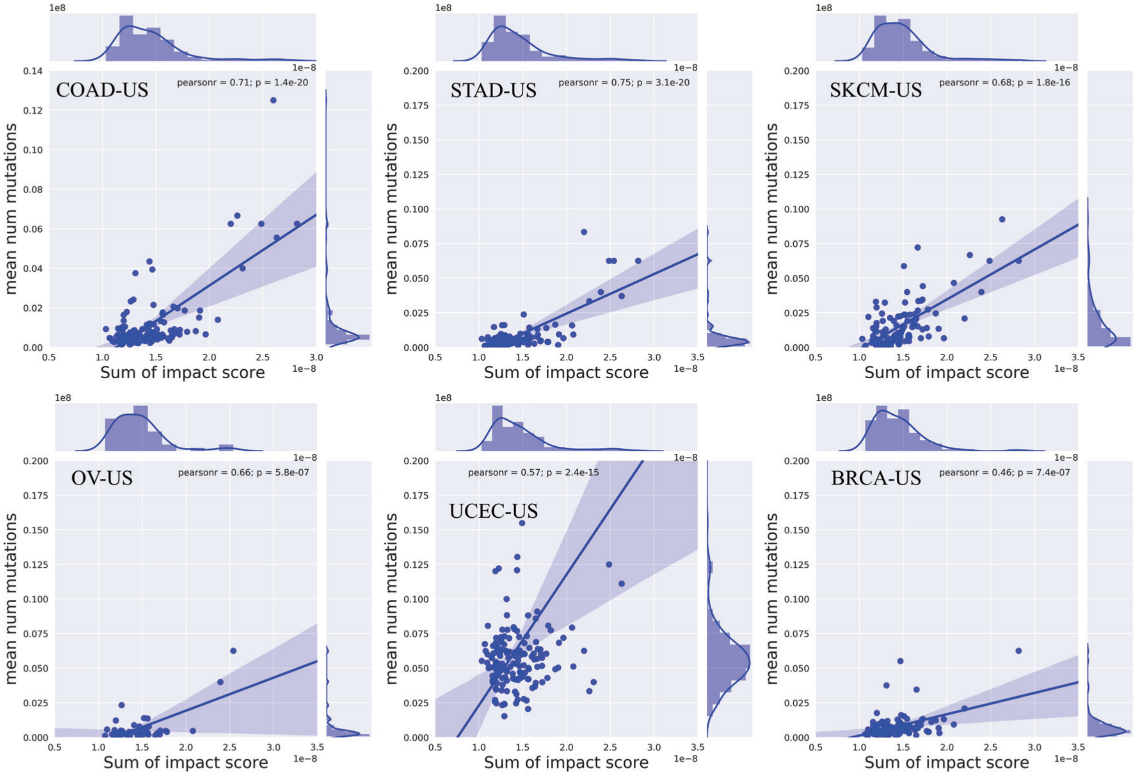

Fig. 4. The average number of mutations located in the MREs of a microRNA (number of mutations divided by number of MREs in the genome) plotted against the weighted sum of the impact scores for COAD-US, STAD-US, SKCM-US, OV-US, UCEC-US, BRCA-US.

As illustrated in Figure 4, we found a positive correlation between the sum of MRE-based impact scores of a microRNA and the average number of somatic mutations observed in its MREs, confirming that our framework indeed predicts the joined effects of the mutational signatures active in a cancer type on the seed sequences in MREs of a microRNA. We therefore can hypothesize that the higher the sum of impact scores for the MREs of a microRNA, the higher the number of mutated MRE seeds and consequently disrupted microRNA–target interactions.

3.4. *Exemplary case studies*

To further confirm that there is actually a relationship between particular mutational processes and mutations in seed regions of MREs, we took a closer look at the joined effect of signatures SBS2 and SBS13—both of which have been attributed to the activity of the AID/APOBEC family of cytidine deaminases[7,10]—which drive a significant subset of breast cancer samples of the BRCA-US dataset (see Fig. 5, right panel).

We first identified microRNAs which have a higher impact score for signatures SBS2 and SBS13 than the sum of the impact scores of SBS1, SBS3 and SBS39 which are also prominent in many breast tumor samples (see the upper left panel of Fig. 2). Then, for each tumor sample, we summed the total number of mutations in MREs of these microRNAs and plotted them against the joined exposure of signatures SBS2 and SBS13 (see Fig. 5, left panel). We observed a clear trend of higher mutation rates for higher exposures (Spearman rank correlation coefficient of 0.51).

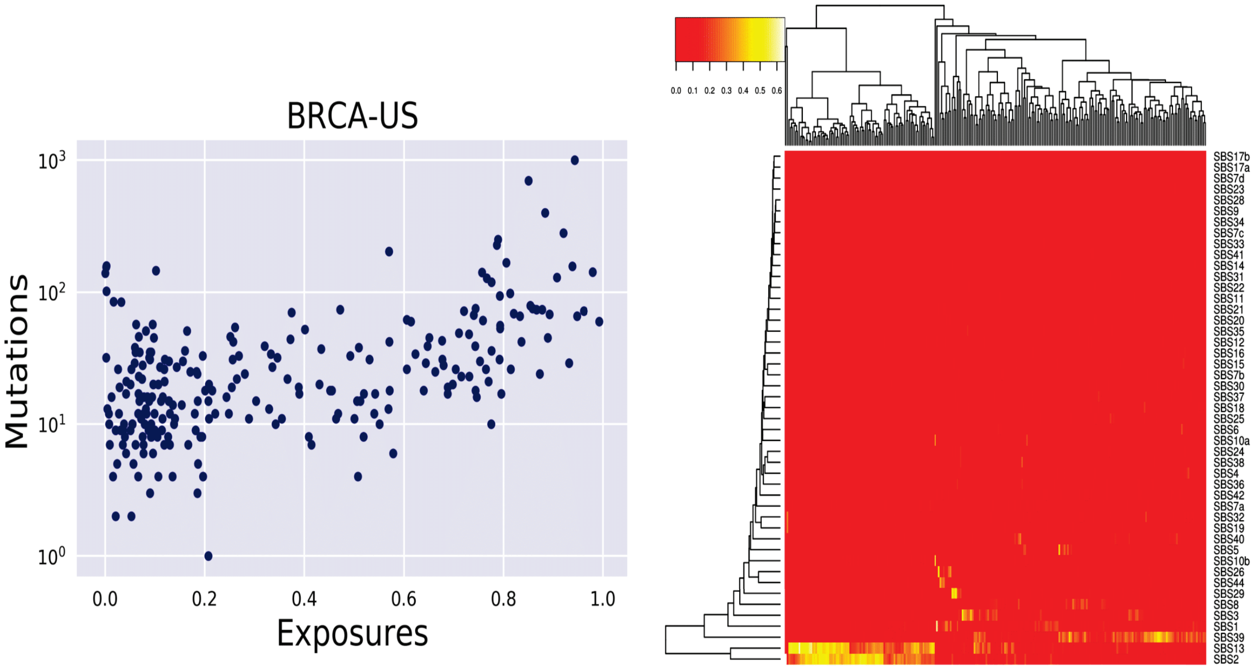

Fig. 5. Number of mutations in MREs of microRNAs with higher impact score for signatures SBS2 and SBS13 than the sum of impact scores for signatures SBS1, SBS3 and SBS39, plotted against the summed exposure of SBS2 and SBS13. Each data point represents a breast tumor from the BRCA-US dataset. Spearman rank correlation coefficient: 0.51. The heatmap on the right side shows that tumors strongly associated with signatures SBS2 and SBS13 constitute an own cluster.

As a second case study, we analyzed the impact of signatures SBS10a and SBS10b in uterine corpus endometrial carcinoma (UCEC-US). Both signatures are thought to be associated with polymerase epsilon (*POLE*) exonuclease domain mutations[7,10] and frequently occur together in the same tumor samples.

We first identified microRNAs which have a higher impact score for signatures SBS10a and SBS10b than the sum of the impact scores of SBS1, SBS26 and SBS44 which are also strongly contributing to the mutation load in many uterine corpus endometrial carcinomas. Then we compared the distributions of the number of mutations in their MREs in tumors with less than 10% contribution of (exposure to) SBS10a and SBS10b against the number of mutations in the same MREs in tumors with a joined exposure of 10% or more. As can be clearly observed in Fig. 6, most tumors with <10% contribution by the *POLE*-associated signatures have only few mutation in the MREs of the selected microRNAs while tumors with at least 10% contribution harbour many more mutations in these MREs ($P = 2 \times 10^{-27}$, Student's t-test).

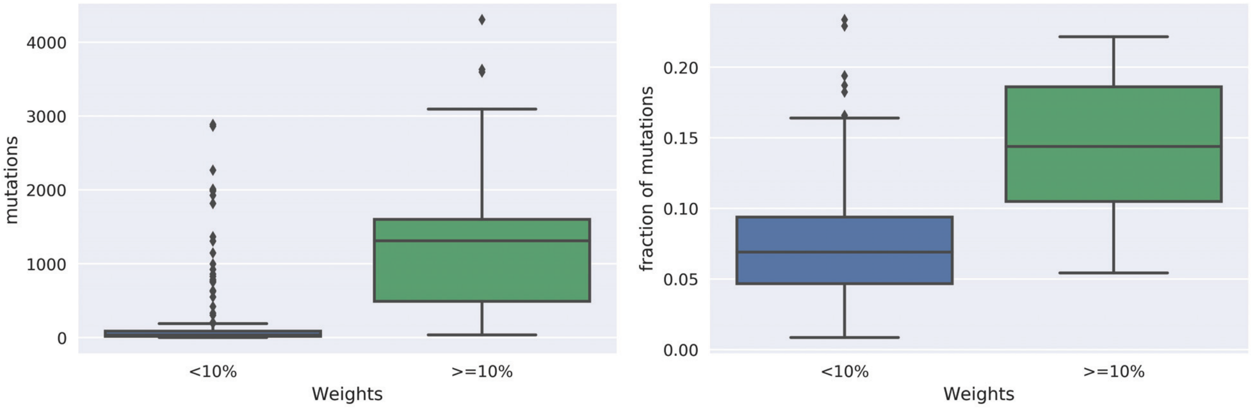

Fig. 6. Number of mutations (left panel) and fraction of mutations (right panel) in MREs of microRNAs with higher impact score for signatures SBS10a and SBS10b than the sum of impact scores for signatures SBS1, SBS26 and SBS44 in the UCEC-US dataset. Left group: tumors with a total exposure of less than 10% for signatures SBS10a and SBS10b; right group: tumors with a total exposure of at least 10%.

4. Conclusion and future perspectives

Taken together our results suggest that our framework can indeed be useful to study which microRNA–target interactions are more likely to be effected by mutations in seed regions due to the mutational processes identified in a tumor genome.

Until now, we have evaluated our approach considering only the seed regions of microRNAs and their targets. Of course, deleterious or disruptive mutations may also occur in other positions of the microRNAs themselves or their MREs. The future extension of our framework will therefore have to take the full mature microRNA sequences and their complementary target sites into account.

Acknowledgments

This research is funded by the ERC Advanced Grant project 693174 “GeCo” (Data-Driven Genomic Computing), 2016-2021. We acknowledge the contributions of Gaia Ceddia.

References

1. D. Bartel, MicroRNAs: target recognition and regulatory functions, *Cell* **136**, 215 (2009).
2. K. Chen and N. Rajewsky, Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes, *Cold Spring Harb. Symp. Quant. Biol.* **71**, 149 (2006).
3. J. Li, Y. Liu, X. Xin, T. S. Kim, E. A. Cabeza, J. Ren, R. Nielsen, J. L. Wrana and Z. Zhang, Evidence for positive selection on a number of microRNA regulatory interactions during recent human evolution, *PLoS Genetics* **8**, p. e1002578 (2012).
4. P. Paul, A. Chakraborty, D. Sarkar, M. Langthasa, M. Rahman, M. Bari, R. S. Singha, A. K. Malakar and S. Chakraborty, Interplay between miRNAs and human diseases, *Journal of Cellular Physiology* **233**, 2007 (2018).
5. G. Calin, C. Dumitru, M. Shimizu, R. Bichi, S. Zupo, E. Noch, H. Aldler, S. Rattan, M. Keating, K. Rai, L. Rassenti, T. Kipps, M. Negrini, F. Bullrich and C. Croce, Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia, *Proc Natl Acad Sci U S A* **99**, 15524 (2002).
6. L. B. Alexandrov and M. R. Stratton, Mutational signatures: the patterns of somatic mutations hidden in cancer genomes, *Current Opinion in Genetics & Development* **24**, 52 (2014).
7. L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, S. A. Aparicio, S. Behjati, A. V. Biankin, G. R. Bignell, N. Bolli, A. Borg, A.-L. Børresen-Dale *et al.*, Signatures of mutational processes in human cancer, *Nature* **500**, 415 (2013).
8. A. Kozomara, M. Birgaoanu and S. Griffiths-Jones, miRBase: from microRNA sequences to function, *Nucleic Acids Res.* **47**, D155 (2019).
9. V. Agarwal, G. Bell, J. Nam and D. Bartel, Predicting effective microRNA target sites in mammalian mRNAs, *eLife* **4**, p. e05005 (2015).
10. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. T. Ng, A. Boot, K. R. Covington, D. A. Gordenin, E. Bergstrom, N. Lopez-Bigas *et al.*, The repertoire of mutational signatures in human cancer, *bioRxiv* (2018).
11. J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, B. Whitty *et al.*, International Cancer Genome Consortium Data Portal–a one-stop shop for cancer genomics data, *Database (Oxford)* **2011**, p. bar026 (2011).
12. A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, I. Barnes, A. Berry, A. Bignell, S. Carbonell Sala, J. Chrast, F. Cunningham, T. Di Domenico, S. Donaldson, I. T. Fiddes, C. Garcáa Girón, J. M. Gonzalez, T. Grego, M. Hardy, T. Hourlier, T. Hunt, O. G. Izuogu, J. Lagarde, F. J. Martin, L. Martínez, S. Mohanan, P. Muir, F. C. P. Navarro, A. Parker, B. Pei, F. Pozo, M. Ruffier, B. M. Schmitt, E. Stapleton, M.-M. Suner, I. Sycheva, B. Uszczynska-Ratajczak, J. Xu, A. Yates, D. Zerbino, Y. Zhang, B. Aken, J. S. Choudhary, M. Gerstein, R. Guigó, T. J. P. Hubbard, M. Kellis, B. Paten, A. Reymond, M. L. Tress and P. Flicek, GENCODE reference annotation for the human and mouse genomes, *Nucleic Acids Research* **47**, D766 (2018).
13. Z. Huang, J. Shi, Y. Gao, C. Cui, S. Zhang, J. Li, Y. Zhou and C. Q, HMDD v3.0: a database for experimentally supported human microRNA-disease associations, *Nucleic Acids Res.* **47**, 1013 (2019).
14. D. C. Ellwanger, F. A. Büttner, H.-W. Mewes and V. Stümpflen, The sufficient minimal set of miRNA seed types, *Bioinformatics* **27**, 1346 (2011).
15. S. Werfel, S. Leierseder, B. Ruprecht, B. Kuster and S. Engelhardt, Preferential microRNA targeting revealed by in vivo competitive binding and differential Argonaute immunoprecipitation, *Nucleic Acids Research* **45**, 10218 (2017).
16. S. Krüger and R. M. Piro, decompTumor2Sig: identification of mutational signatures active in individual tumors, *BMC Bioinformatics* **20**, p. 152 (2019).