Identification of Mutational Signatures Active in Individual Tumors

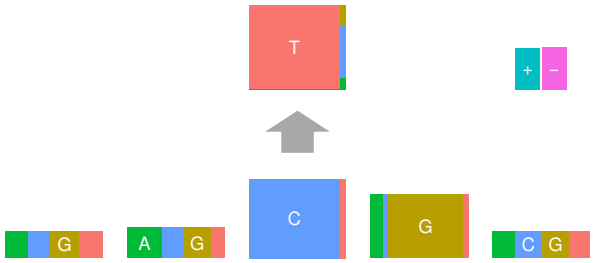
Sandra Krüger[1] and Rosario M. Piro[1,2,3,*]

[1]Institute of Computer Science and Institute of Bioinformatics, Freie Universität Berlin, Berlin, Germany
[2]Institute of Medical Genetics and Human Genetics, Charité Universitätsmedizin Berlin, Berlin, Germany
[3]German Cancer Consortium (DKTK)
[*]E-mail: r.piro@fu-berlin.de

Introduction

The mutational processes responsible for the somatic mutations observed in tumor samples can significantly vary not only between tumor types but also among the individual cancers within a tumor class. Mutational processes can be represented by so called "mutational signatures" [1-3] which reflect the occurrences of base changes within their sequence contexts (i.e., in dependence on their flanking bases). The age-related mutations[1] initiated by spontaneous deamination of 5-methylcytosine, for example, regard C>T transitions in the context of CpGs (because those are methylated; see Fig. 1). Other characteristic mutation patterns are known for exogenous mutagenic factors such as UV light and cigarette smoke (see [4] for a review).

There are two conceptualizations of mutational signatures. The model first described by Alexandrov et al. [1,2] separately counts all possible nucleotide triplets whose central base is mutated, describing, for example, the most frequent mutations caused by spontaneous deamination of 5-methylcytosine as A[C>T]G, C[C>T]G, G[C>T]G, and T[C>T]G. Given that there are a total of 96 possible triplet mutations (when removing the redundancy due to the reverse complement strand and considering only single nucleotide variants), each "Alexandrov signature", as we call it, consists of 96 mutation probabilities that indicate which of the changes are occurring most frequently due to the mutational process it describes.



Nucleotide change (central base)

C>A	C>G	C>T	T>A	T>C	T>G
0.004	0.006	0.928	0.009	0.038	0.015

Flanking bases

Position	A	C	G	T
-2	0.237	0.228	0.293	0.242
-1	0.362	0.220	0.279	0.139
+1	0.131	0.053	0.764	0.052
+2	0.232	0.277	0.277	0.214

Transcription strand

plus strand	minus strand
0.493	0.507

Figure 1: The probabilistic Shiraishi model of mutational signatures. The example uses two flanking bases in each direction and was obtained from 435 tumor genomes (see below). It likely represents mutations caused by spontaneous deamination of 5-methylcytosine. Left: graphical representation obtained from `pmsignature` [3]. Right: the $6 + 4 * 4 + 2 = 24$ parameters of the signature. Each parameter is the fraction of mutations caused by the mutational process which exhibit the respective characteristic. E.g., 92.8% of the changes are C>T and there is next to no transcriptional bias.

[1]Here and in the entire manuscript, we will consider only single nucleotide variants although we generically speak of mutations.

The second model [3] describes probabilistic mutational signatures in analogy to the way we model transcription factor binding motifs, i.e., considering the single bases of the motif as independent. Using this approach, the most frequent changes due to spontaneous deamination of 5-methylcytosine can be described as C>T transitions followed by a base that with a very high probability is a G, and preceded by any of the four bases with approximately equal probability (see Fig. 1). Instead of $6*4*4 = 96$ parameters, a "Shiraishi signature", as we call it, requires only $6+4+4 = 14$ parameters when considering nucleotide triplets [3]. It can therefore be more easily extended to more flanking bases (e.g., two flanking bases in each direction as shown in Fig. 1), and the incorporation of the transcription-strand direction requires only two additional parameters instead of doubling the number of parameters.

As an example, the table in Fig. 1 shows the parameters (equivalent to probabilities) of the Shiraishi signature for spontaneous deamination of 5-methylcytosine using a sequence context of five bases (with the altered base in the center) and taking transcription strand into consideration. Since the possible base changes, the individual flanking bases, and the transcription strand are treated as independent, the probabilities for each of them sums to 1. For more details, we refer to the original paper [3].

To summarize, while the Alexandrov model takes the full dependence between mutated nucleotides and their directly neighboring flanking bases into account, the simplified Shiraishi model treats mutated nucleotides and flanking bases as independent features of the signature.

The initial discovery and construction of mutational signatures requires a large amount of tumor samples, such that regular patterns can be identified [1-3]. This is impractical in a clinical setting, where each cancer patient is diagnosed individually. However, once accurate signatures have been defined, they can be used to evaluate their contribution to the mutational load in individual tumor samples. This helps to assess which mutational processes were likely involved in the development of the tumor, as has been demonstrated for the Alexandrov signatures [5,6], and may be of clinical relevance, e.g., when they hint at a DNA repair deficiency, because DNA repair mechanisms significantly affect the response to cytotoxic treatments [7].

Here, we present a user-friendly R package, called `decompTumor2Sig`, that can be used to evaluate the contribution of Shiraishi signatures to the somatic mutations found in an individual tumor, allowing larger sequence contexts to be taken into consideration than with Alexandrov signatures. (In addition, the package can also be used for Alexandrov signatures, but we will discuss only Shiraishi signatures here.)

`decompTumor2Sig` is freely available at: `http://rmpiro.net/index.html#downloads`

Methods

Contribution of signatures to individual tumor samples:

To derive the influence of a given set of Shiraishi signatures on the generation of the mutational catalog (i.e., the set of somatic mutations) of an individual tumor sample, the `decompTumor2Sig` package takes the same quadratic programming approach used by Lynch for Alexandrov signatures [6].

Let g be the tumor genome, described in terms of fractions of somatic mutations that have specific nucleotide changes, flanking bases and transcription strands. The representation of the tumor genome is thus identical to the representation of the single signatures as exemplified in the table in Fig. 1. Let further $\mathbf{S}$ be a $P \times K$ matrix, with each column being one of K signatures composed of P parameters (here: $P = 24$). The goal is to determine a vector w of weights w_s (Alexandrov et al. [1,2] called them "exposures") which indicates how strongly each signature $s \in (1, k)$ contributed to the mutation load of the tumor, i.e., what fraction of the somatic mutations in g were caused by s. Of course, we would like to have $\mathbf{S}w \approx g$ with as little error as possible. We therefore can solve the following problem:

$$\begin{aligned} \text{minimize} \quad & (g - \mathbf{S}w)^T(g - \mathbf{S}w) = g^T g - g^T \mathbf{S}w - (\mathbf{S}w)^T g + (\mathbf{S}w)^T \mathbf{S}w \\ & \text{subject to } \textstyle\sum_{s=1}^{k} w_s = 1, w_s \geq 0 \end{aligned}$$

Since $g^T g$ is constant and $(\mathbf{S}w)^T g = g^T \mathbf{S}w$, we can simplify the problem as:

$$\text{minimize} \quad -g^T \mathbf{S}w + \tfrac{1}{2} w^T \mathbf{S}^T \mathbf{S}w \qquad \text{subject to} \quad \sum_{s=1}^{k} w_s = 1, w_s \geq 0$$

We solve this classical quadratic programming problem using the R package `quadprog` [8] and thus compute the contributions of the given Shiraishi signatures to the overall mutation load of the tumor.

Estimation of accuracy:

To estimate the accuracy with which we can determine the contribution of mutational signatures to the mutation load of an individual tumor, we proceed as follows:

1. For a given set of T tumors, using the R package `pmsignature` [3] we collectively derive a set of Shiraishi signatures $\mathbf{S}$ and their corresponding contributions/weights $w^{(t)}$, $t \in (1, T)$ to the tumor genomes. We take these computed weight vectors as "truth" set.

2. We take either each single tumor out of the set (leave-one-out test) or, for larger sets, a randomly chosen subset (test set). The remaining tumors are used to collectively recompute signatures $\mathbf{S}'$ as in 1.

3. For each individual tumor t^* to be tested, we estimate the contributions/weights $w'^{(t^*)}$ using our tool (`decompTumor2Sig`) on the test signatures $\mathbf{S}'$, i.e., on the signatures that were derived without mutation data from t^*. Note, this constitutes a realistic application where one has a given set of signatures and wants to apply them to a novel tumor sample.

4. We determine a 1:1 mapping between the test signatures $\mathbf{S}'$ and the original signatures $\mathbf{S}$ (minimum Frobenius distance between the signature matrices), so that we can compare the contributions $w'^{(t^*)}$, estimated by `decompTumor2Sig`, to the "truth" contributions originally computed using `pmsignature` in step 1.

Results

We performed two tests to show that our tool can effectively decompose the mutation catalog of an individual tumor and determine the relative contributions of a given set of Shiraishi signatures: 1) We performed a leave-one-out cross validation using somatic mutations from a set of 21 breast cancer genomes [9]. Given the limited size of the dataset, we decomposed the tumors into four Shiraishi signatures. 2) We performed a second test using the somatic mutations of 435 tumor genomes with at least 100 somatic mutations from ten different tumor entities [2] (ALL, AML, CLL, breast cancer, liver cancer, lung cancer, pancreas cancer, lymphoma, medulloblastoma, and pilocytic astrocytoma). We randomly selected 44 tumors (≈10%) as a test set. Due to the larger cohort, we could decompose the tumors into 15 signatures.

Figure 2 compares the weights predicted for individual tumors (`decompTumor2Sig`) to those computed using the entire set of tumors (`pmsignature`). Interestingly, in the leave-one-out cross validation using 21 tumors, we obtained two extreme outliers with a highly discordant prediction (see Fig. 2, left). We found these to be predictions for a hypermutated sample (PD4120a; 70,690 somatic SNVs as opposed to a mean of 5661 SNVs for the remaining tumors). Removing the tumor for the leave-one-out test significantly affects the derived signatures and hence the contributions/weights predicted for the hypermutated tumor.

Both tests demonstrate a good prediction performance, yielding a generally high concordance between predicted and "true" weights. For the leave-one-out test on 21 breast cancers, the median deviation (difference) of predicted and expected weights is 0.018 (mean: 0.0317 including and 0.0197 excluding the outliers); for the 44 out of 435 tumors of different cancer types, the median deviation is 0.0187 (mean: 0.0262, max: 0.155).

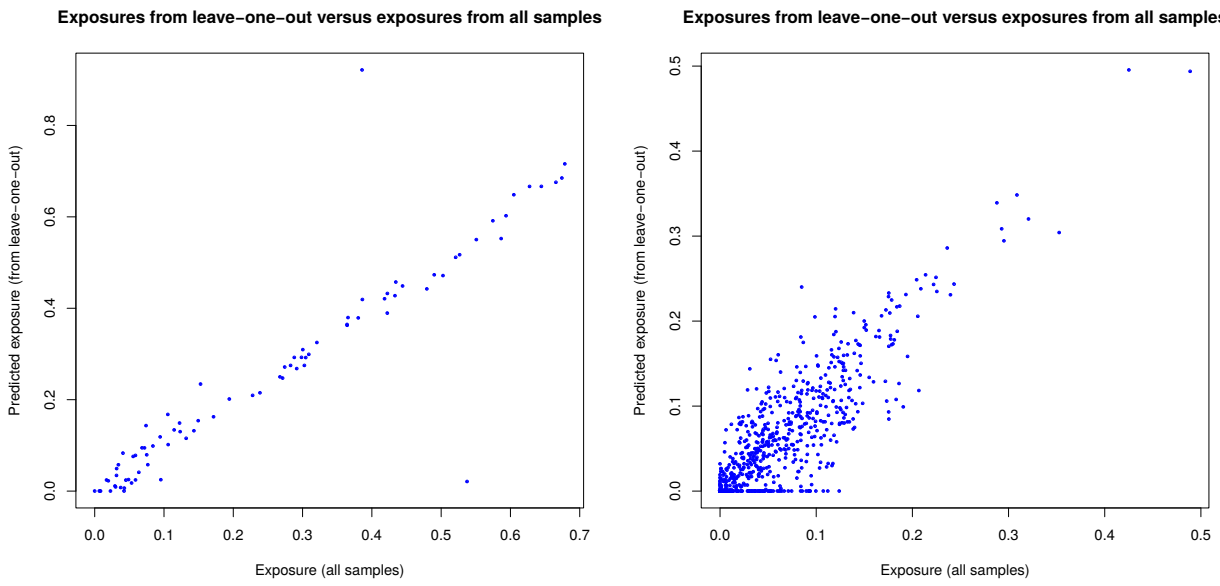


Figure 2: Comparison of contributions/weights ("exposures") predicted for individual tumors (`decompTumor2Sig`; y-axis) and collectively computed (`pmsignature`; x-axis). Left: leave-one-out test on 21 breast cancers ($r = 0.923$). Right: test set of 44 out of 435 tumors ($r = 0.807$).

Conclusions

We have developed a tool for dissecting mutational catalogs of individual tumor samples in terms of the simplified mutational signature model proposed by Shiraishi et al [3]. The tool is implemented in a fast, user-friendly R package, `decompTumor2Sig`, and shows a good performance.

References

1. Alexandrov LB, et al. Deciphering signatures of mutational processes operative in human cancer. Cell Reports. 2013. 3(1):246–259. doi:10.1016/j.celrep.2012.12.008

2. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013. 500(7463):415–421. doi:10.1038/nature12477

3. Shiraishi Y, et al. A simple model-based approach to inferring and visualizing cancer mutation signatures. PLoS Genetics. 2015. 11(12):e1005657. doi:10.1371/journal.pgen.1005657

4. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev. 2014. 24:52–60. doi:10.1016/j.gde.2013.11.014

5. Rosenthal R, et al. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology. 2016. 17:31. doi:10.1186/s13059-016-0893-4

6. Lynch AG, et al. Decomposition of mutational context signatures using quadratic programming methods. F1000Research. 2016. 5:1253. doi:10.12688/f1000research.8918.1

7. Kelley MR, et al. Targeting DNA repair pathways for cancer treatment: what's new? Future Oncology. 2014. 10(7):1215–1237. doi:10.2217/fon.14.60

8. Berwin A (original implementation for S), Weingessel A (port to R). quadprog: Functions to solve Quadratic Programming Problems (R package version 1.5-5). 2013. `https://cran.r-project.org/package=quadprog`

9. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012. 149(5):979–993. doi:10.1016/j.cell.2012.04.024