Pacific Symposium on Biocomputing 25:262-273(2020)

DNA REPAIR FOOTPRINT UNCOVERS CONTRIBUTION OF DNA REPAIR MECHANISM TO MUTATIONAL SIGNATURES

Damian Wojtowicz[1]
National Center of Biotechnology Information, National Library of Medicine, NIH
Bethesda, MD 20894 USA
Email: damian.wojtowicz@nih.gov

Mark D.M. Leiserson
Department of Computer Science, University of Maryland
8125 Paint Branch Dr, College Park, MD 20740 USA
Email: mdml@cs.umd.edu

Roded Sharan[2]
Blavatnik School of Computer Science, Tel Aviv University
Tel Aviv, 69978 Israel
Email: roded@tau.ac.il

Teresa M. Przytycka[1,*]
National Center of Biotechnology Information, National Library of Medicine, NIH
Bethesda, MD 20894 USA
Email: przytyck@ncbi.nlm.nih.gov

Cancer genomes accumulate a large number of somatic mutations resulting from imperfection of DNA processing during normal cell cycle as well as from carcinogenic exposures or cancer related aberrations of DNA maintenance machinery. These processes often lead to distinctive patterns of mutations, called mutational signatures. Several computational methods have been developed to uncover such signatures from catalogs of somatic mutations. However, cancer mutational signatures are the end-effect of several interplaying factors including carcinogenic exposures and potential deficiencies of the DNA repair mechanism. To fully understand the nature of each signature, it is important to disambiguate the atomic components that contribute to the final signature. Here, we introduce a new descriptor of mutational signatures, DNA Repair FootPrint (*RePrint*), and show that it can capture common properties of deficiencies in repair mechanisms contributing to diverse signatures. We validate the method with published mutational signatures from cell lines targeted with CRISPR-Cas9-based knockouts of DNA repair genes.

Keywords: Somatic Mutations; Mutational Signatures; DNA Repair; Mismatch Repair Deficiency; Cancer.

[1]Work supported by Intramural Research Program of the National Library of Medicine, NIH.
[2]Work supported by Len Blavatnik and the Blavatnik Family Foundation.
[*]Corresponding author

1. Introduction

Over the life of an organism somatic cells are constantly exposed to DNA damage[1]. DNA lesions can emerge during normal cell cycle due to the stochastic nature of biochemical processes involved in DNA replication or can be triggered by exposure to carcinogenic agents such as tobacco smoking, ultraviolet light (UV), or chemicals. To counterbalance this threat, cells evolved complex DNA repair mechanisms that are able to correct various types of DNA damage[1,2] with astonishing accuracy. Yet these DNA repair mechanisms are not perfect, leading to the accumulation of mutations over the life of the organism. If the exposure to carcinogenic agents is increased, or if the DNA repair mechanism is in any way compromised, the accumulation of somatic mutations becomes more rapid often leading to cancer.

It has been appreciated for quite some time that various mutagenic causes leave characteristic mutation footprints on the genome[3] and in cancer[4,5]. Knowledge of elementary mutational processes underlying cancer cells is essential for understanding the etiology of cancer and its progression. The molecular mechanisms underlying specific signatures can suggest treatment strategies (see a review by Van Hoeck et al.[6]).

Focusing on single base substitutions, systematic analysis of mutation patterns typically start by considering each mutation in the context of its flanking base pairs, yielding 96 possible mutation types (since it is not possible to differentiate if a mutation occurred on forward or reverse strand, we have 6 possible base substitutions but 16 possible flanking combinations). Mutational signatures are typically defined by the relative frequencies of these 96 mutation types although other sequence contexts have also been considered[7,8]. Starting from the seminal work of Alexandrov et al.[9], several computational methods have been developed to identify distinct mutation signatures presented in the catalogue of cancer mutations[7,10–13] and to decompose patient's mutations into predefined mutation signatures[14,15]. As the number of sequenced cancer genomes increases, so does the number of observed signatures. For example, the initial, studies based on a small number of breast cancer samples, reported 22 distinct signatures[10,13]. A more comprehensive study based on 7,042 cancers proposed 30 substitution signatures[9] while the most recent study yielded more than 50 substitution signatures[16]. Some of these signatures have been associated with potential causes, while the cause of many others is still unknown.

Cancer mutational signatures can be seen as the end-effect of several interplaying factors: the nature of DNA damage including specific properties of the lesion (e.g. DNA break, covalent modification, bulky adducts), the properties and distribution of sites that are vulnerable to the damage (*mutation opportunity*) and the potential deficiencies of the repair mechanism responsible for repairing the primary damage (Figure 1). The distribution of mutation opportunities is typically assumed to be the same in different cancer genomes. Specifically, because of the usage of the same

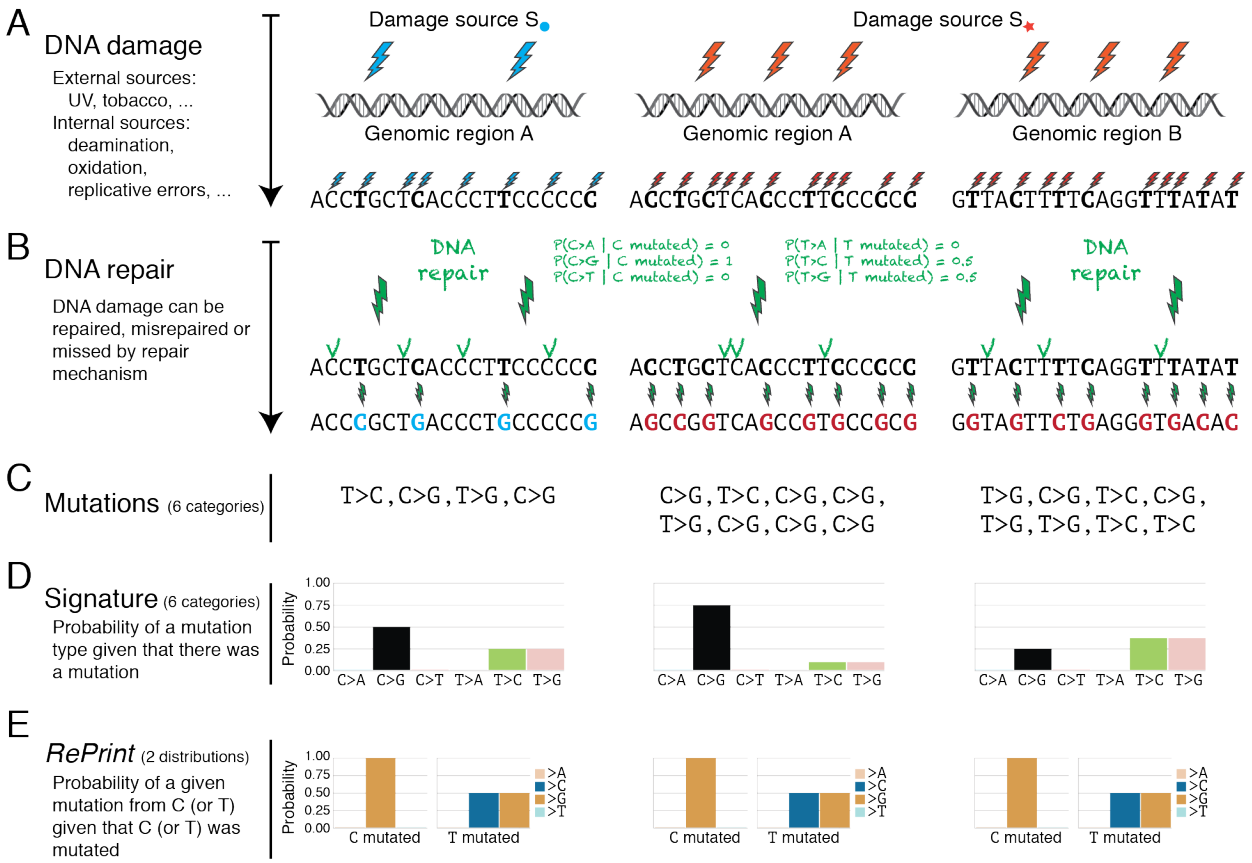

Fig. 1. Conceptualization of DNA Repair FootPrint (*RePrint*). (**A**) DNA is exposed to external and internal sources of DNA damage. DNA primary damage occurs with probability depending on the level of exposure to DNA damage source and the DNA sequence context and state (see three two genomic regions A and B subject to two damage sources). (**B**) DNA repair mechanisms serve to counteract the primary damage, but it is not a perfect mechanism and it has some deficiencies. As a result, some of primary damage becomes misrepaired or even missed by the repair mechanism. (**C**) Unrepaired DNA damage leads to mutations. For simplicity of signature and *RePrint* presentation, we used here the simplified notation of mutations with 6 mutation types instead of the commonly used 96 mutation types. (**D**) Three mutational signatures for the presented unrepaired damage are shown; note that they are all distinctive. Signature is a probability vector over 6 types of mutation. (**E**) *RePrint*s of the above signatures are identical, as *RePrint* may correct for differences between primary damage sources and nucleotide composition of affected genomic regions. Assuming that a given nucleotide is mutated, *RePrint* provides a probability of a particular mutation over 3 possible mutation types; such probability distribution is given for each possible reference nucleotide – here for C or T, but in general for 32 reference trinucleotides when 96 mutation categories are considered.

reference genome, in the absence of structural variants, the distribution of triplets is the same. As for other properties that contribute to DNA vulnerability, such as phosphorylation status, single versus double stranded DNA, open versus closed chromatin, etc., the situation is more complex. It is quite possible that DNA stress, double strand breaks (DSBs) or other cancer related changes create mutation opportunities that could lead to unique patterns of DNA primary damage.

Different combinations of the events leading to primary DNA damage and deficiencies of repair mechanisms can lead to drastically different primary mutation patterns. Indeed, the current catalog of cancer mutation signatures contains 8 mutation signatures associated with MMR deficiency. Mismatch repair is a process that corrects mismatched nucleotides in the otherwise complementary paired DNA strands. Such mismatches can arise during DNA replication and recombination, and repair of some forms of DNA damage. Previous results suggested that this large number of MMR deficiency signatures might be a consequence of combining different causes for the initial DNA damage with MMR repair deficiency. For example, a recent study linked two of the signatures associated with MMR deficiency (Signature 14 and 20), with simultaneous MMR deficiency and mutations in polymerase POLE/POLD1 respectively[8].

In general, it is possible that a lesion in the same triple can be repaired differently depending on larger genomic context. For example, Signature 2 and 13 are both assumed to emerge as the result of DNA modifications introduced by APOBEC family enzyme but are then assumed to be repaired by different mechanisms[17]. Similarly, in the case of AID-mediated cytosine deamination, there is evidence that repair differs with regards to the immediate flanking sequence and whether the mutation is in an immunoglobulin gene[1]. The reverse is also true. Complex interplay between exogenous and endogenous factors can lead to very different distributions of primary lesions that are then repaired by the same repair mechanism. For example, DNA stress can promote formation of ssDNA. The regions particularly prone to such stress induced destabilization (SIDDs) are typically AT-rich[18,19]. Thus, DNA stress can result in a shift in the distribution of mutations in ssDNA regions towards mutations of T/A. Yet, despite of this shift of mutation context, the repair mechanism could remain the same.

Motivated by this reasoning we asked if the mismatch repair deficiency leaves a common imprint on different signatures that are associated with this deficiency. To answer this question, we propose a simple probabilistic model, DNA Repair FootPrint (*RePrint*), which can be estimated given a mutational signature. We show that mismatch repair signatures have similar *RePrints*. Similar *RePrints* were also obtained for the signatures that are related to oxidation stress. In contrast, the two signatures (Signature 2 and 13) that are both assumed to emerge as the result of DNA modifications introduced by APOBEC family enzyme but are then assumed to be repaired by different mechanisms[17] have very different *RePrint*s.

2. *RePrint* – the DNA repair deficiency footprint of a signature

Different combinations of mutagenic factors can lead to different patterns of primary mutations, yet these diverse primary damage patterns might undergo the same DNA damage repair process that serve to counteract damage (Figure 1A,B). The observed mutations are an outcome of

an incorrect repair of the primary lesions due to deficiencies in MMR mechanisms (Figure 1B,C). The primary lesions have been created by different mutagenic processes and/or their interactions with different cellular processes such as differential methylation, replication stress, etc., leading to different distribution of the primary lesions. The end-effect of a given DNA damage process together with the interplaying factors and potential deficiencies of the DNA repair mechanism is manifested in mutational signature representing probabilities of all possible mutation (Figure 1D). However, assuming that there is a common repair deficiency, then we can assume that the probability of one of three possible mutations of a given reference nucleotide (given that it is mutated) depends only on that nucleotide (Figure 1B,E) and possibly its local context, and it is independent of the underlying primary DNA mutagenic process.

To discover cases for which a common repair deficiency may exists we analyzed existing mutational signatures. For every nucleotide triple, we ask what the probability of a specific mutation is assuming that this triple is mutated. We have 32 possible triples and for each triple we compute such conditional probabilities. The probabilities are represented as a vector called DNA Repair FootPrint (*RePrint*). Specifically, $RePrint(N_L[X > Y]N_R)$ is the probability that X is mutated to Y under the assumption that X is in the central position of the triple N_LXN_R and that X was mutated. In particular, $\sum_{Y \neq X} RePrint(N_L[X > Y]N_R) = 1$ (see Materials and Methods for details). Note that *RePrint* of a signature can be directly computed from the signature definition. If mutational signatures are very similar, their *RePrint*s are very similar as well; however, the reverse is not true – it is possible for distinct signatures to have similar *RePrint*s. Figure 2 shows an example of two dissimilar mutational Signatures 6 and 26, both associated with DNA mismatch repair deficiency, which despite their dissimilarity have similar *RePrint*s.

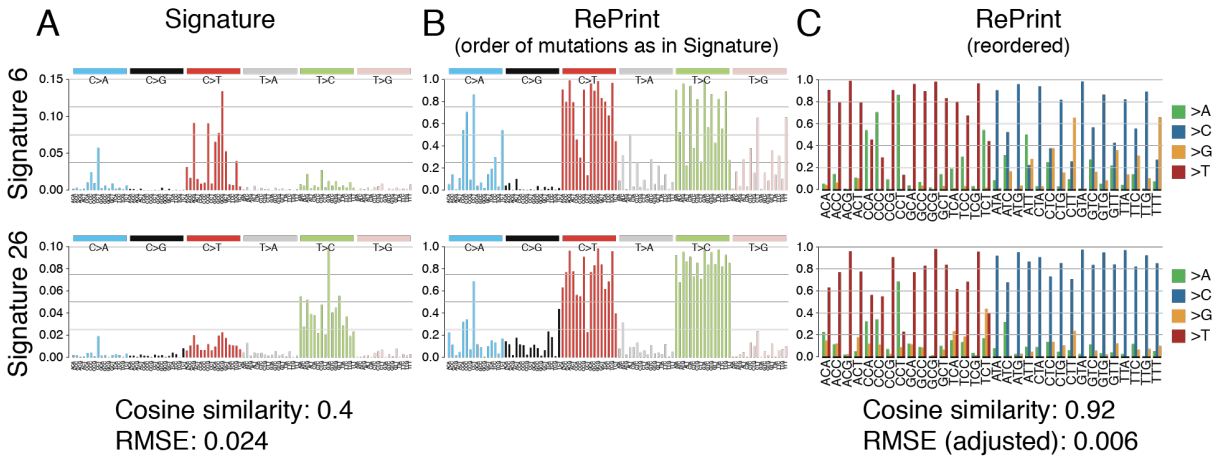

Fig. 2. Signatures 6 and 26 are both associated with DNA mismatch repair deficiency. While their signatures (**A**) show substantial differences (cosine similarity: 0.4), the *RePrints* of these signatures (**B**,**C**) are similar (cosine similarity: 0.92). This is consistent with common deficiency in the repair mechanism.

Consistent with the idea behind our model, one of the properties that distinguishes Signatures 6 and 26, is that mutations attributed to Signature 26 are highly enriched in late replicating regions – a property not observed for Signature 6.[17,20] Early and late replicating regions have very different nucleotide composition with early regions being G/C rich while late regions being A/T rich. Consistent with this observation Signature 26 has higher proportion of mutations from T than from C (Figure 2A) leading to big differences between these two signatures. However, *RePrint* is insensitive to the nucleotide composition of the region where the primary mutations preferentially occur. The similarity of *RePrint*s suggests that the differences between these two signatures are due to differences in the distribution of the primary mutations rather than differences in the repair deficiency step. Thus, *RePrint*s can provide information on similarities of repair mechanism deficiencies that cannot be detected from mutation signatures alone. As a potential confounding factor, note that it is possible that some mutagens do not produce lesions in the context of some triples and then the corresponding *RePrint* conditional probability for this triple will be uniform (see Materials and Methods). We discuss possible ways of mitigating this issue in the Discussion section.

3. Validating the ability of *RePrint* to capture common DNA repair mechanism deficiencies

To test the hypothesis that *RePrint*s can capture common repair mechanism deficiencies we clustered *RePrint*s of the 30 COSMIC signatures (version 2). We also included in the clustering analysis three signatures from cell line screens with targeted CRISPR-Cas9-based knockouts of DNA repair genes: *MSH6*, *FANCC*, and *EXO1*, from Zou et al.[21] These knockouts have been performed to recreate DNA repair related mutational signatures under highly experimentally-controlled conditions.

Strikingly, *RePrint*s of all COSMIC signatures associated with MMR deficiency clustered together (Figure 3B; Fisher test p-value = 6.5e-07). This cluster included also the MSH6 knockout signature. This is consistent with the fact that *MSH6* knockout leads to DNA mismatch repair (MMR) deficiency. Interestingly, the *RePrints* of the signatures associated with concurrent perturbation of MMR and DNA proofreading (Signatures 14 and 20)[8] cluster together with the MMR deficiency signatures. In contrast, when clustering these signatures directly (as opposed to their *RePrints*), the MMR signatures do not form a coherent cluster but instead partition into two clearly separated clusters (Figure 3A; Fisher test p-values > 1.0e-3). This suggests that *RePrint* is able to capture hidden similarities between these signatures.

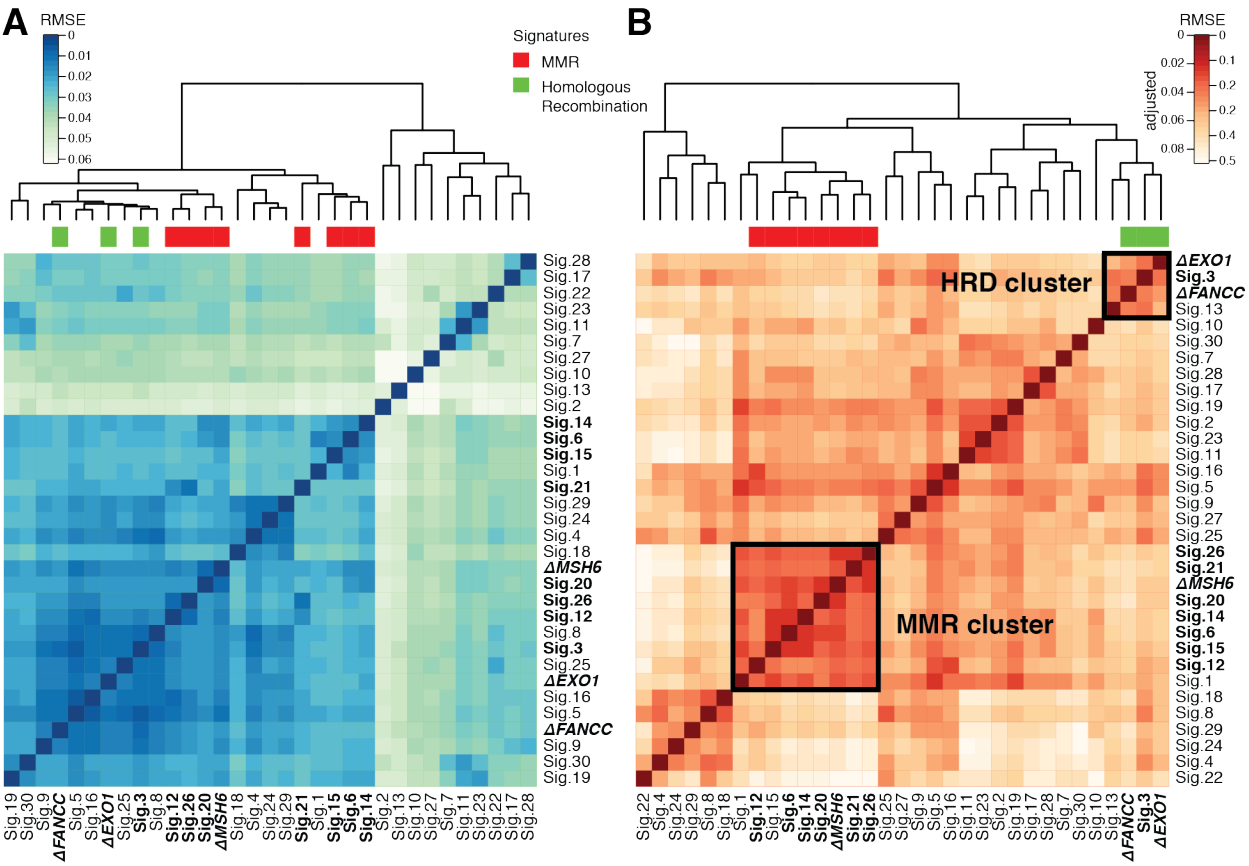

Fig. 3. Clustering of mutational signatures from COSMIC version 2 and three knockouts experiments of DNA repair genes: *MSH6*, *FANCC*, and *EXO1* (**A**) and their *RePrints* (**B**). The signatures known to be related to MMR are marked by red boxes. Green boxes mark signatures putatively associated with Homologous Recombination. RMSE scale legends for signatures and *RePrint*s are shown on the left and right, respectively; the latter contains also adjusted RMSE scale (left axis) for comparison between signature and *RePrint* RMSE values.

The second knockout, *FANCC*, is a component of the Fanconi anemia DNA repair system related to homologous recombination (HR). Homologous recombination is a mechanism used by cells to accurately repair breaks that occur on both strands of DNA, known as double-strand breaks (DSB). Therefore, *FANCC* knockout is expected to create an effect similar to the effect of Homologous Recombination Deficiency (HRD). Among COSMIC signatures, Signature 3 is known to be associated with HRD. Consistent with these expectations, *RePrint*s of Signature 3 and the *FANCC* knockout signature clustered together.

The last of the knockouts, *EXO1*, is assumed to be involved in both MMR and DSB repair. Indeed, its *RePrint* clustered with the *RePrint* of Signature 3 consistent with its role in DSB repair. Encouraged by these results, we considered a larger, recently completed compendium of mutational signatures from Alexandrov et al.[16] – COSMIC version 3. In addition to characterizing

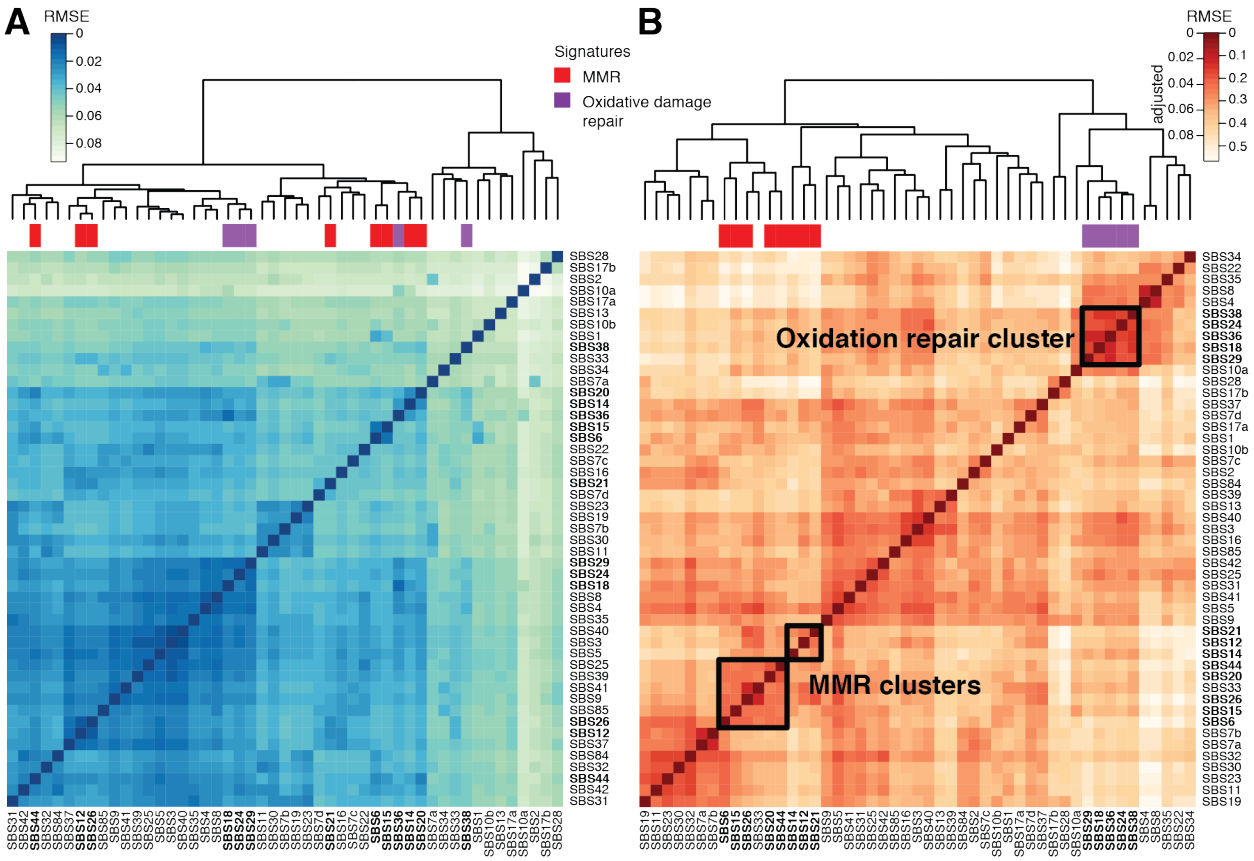

Fig. 4. Clustering of mutational signatures from COSMIC version 3 (**A**) and their *RePrints* (**B**). The signatures known to be related to MMR are marked by red boxes. Purple boxes mark signatures putatively associated with oxidative damage repair. RMSE scale legends for signatures and *RePrint*s are shown on the left and right, respectively; the latter contains also adjusted RMSE scale (left axis) for comparison between signature and *RePrint* RMSE values.

a larger set of signatures, many signatures, including MMR associated signatures, have been refined increasing separation between them. (The revised signatures are referred by SBSx where x is the signature number.) Strikingly, *RePrint*s of the refined MMR associated signatures still cluster together although in two clusters bringing together signatures with dissimilar profiles (Figure 4B; Fisher test p-values: 1.6e-4 and 3.0e-3 for the two cluster). The clustering of these signatures directly (not their *RePrints*) shows even less grouping than previously for signatures associated with MMR deficiency (Figure 4A and 3A).

These results clearly show that similarities of the *RePrint*s of signatures can capture common repair mechanism more accurately than the similarities between signatures themselves.

4. Novel biological insights

Further analysis of the *RePrints* of substitution signatures from COSMIC version 3 shows that Signature SBS33 clustered together with the signatures associated with MMR deficiency (Figure 4B). This is a very rare signature detected only in 4 patients. Currently no etiology has been proposed. The clustering of its *RePrint* with the *RePrint*s of MMR deficiency signatures suggests that this signature may be associated with defective DNA mismatch repair.

Another striking *RePrint* cluster is clearly related to oxidation damage repair (Figure 4B). It contains Signatures SBS18, SBS36, SBS24, and SBS38, and SBS29. As additional evidence shows, SBS18 is known to be related to oxidative stress[22]. Signature SBS36 (known as the MUTYH signature) is detected in patients with biallelic germline or somatic MUTYH mutations – a major player in oxidation damage repair pathway. SBS38 is found only in UV light associated melanomas. Ultraviolet A (UVA) light, UV associated with skin aging, is known to generate mutagenic oxidative stress[23]. Signature SBS29 is associated with tobacco chewing. This is consistent with previous studies demonstrating systemic oxidative stress in tobacco chewers[24]. Interestingly, this cluster contains also SBS24, a signature associated with exposure to aflatoxin, suggesting that SBS24 involves a repair pathway common with the oxidative stress repair pathway. Indeed, one of the causes for aflatoxin induced toxicity is oxidative stress[25]. In a separate analysis, when we analyzed all COSMIC signatures (version 3) including possible sequencing artefacts, we found that Signature SBS45 clusters together with the oxidative related signatures (data not shown). This is consistent with the fact that SBS45 is believed to be an artefact due to 8-oxo-guanines (one of the most common DNA lesions resulting from reactive oxygen species) introduced during sequencing. As a potential sequencing artifact, SBS45 was initially not included in the primary analysis (see Materials and Methods).

5. Conclusions and future work

Cancer mutational signatures are result of the interplay between DNA damage resulting from exposure to mutagenic processes and potential deficiencies of the DNA repair mechanism. While the effects of different environmental mutagenic processes can be assumed to be additive, the combined effect of DNA damage and repair deficiency are often non-additive[27]. Most methods for inferring mutational signatures do not account for non-linear effects and this is the most likely cause of the fact that the current catalog of cancer mutation signatures contains 8 mutation signatures associated with MMR deficiency. This study demonstrates that, using our simple method based on signature *RePrints,* it might be possible to identify signatures that share common DNA repair deficiency. In particular, our analysis revealed that despite the differences in signatures related to MMR deficiency, they all share common MMR deficiency footprint. *RePrint* is the first attempt to identify such common imprint of DNA repair deficiency, and still has some

limitations. In particular, if a triplet is never (or rarely) mutated the respective *RePrint* conditional probability will be set to uniform. This might create both false positive and false negative results. We hope to overcome these shortcomings in future improvements to the method. One possible direction to achieve this goal, is not to consider the signatures by themselves but additionally take into account all mutations in patients with the given signature. Since DNA repair mechanism acts on whole genome, this additional information might help to overcome the problems discussed above. Overall, we hope that the results of paper will inspire researchers to seek further methods that will allow for a rigorous decomposition of deficiencies in repair mechanism and primary mutation causes.

6. Materials and Methods

6.1. Mutation signatures

Mutation signatures inferred from human cancers were downloaded from the Catalogue of Somatic Mutations in Cancer (COSMIC)[26]: https://cancer.sanger.ac.uk/cosmic/signatures. We analyzed the 30 signatures from COSMIC version 2 and the 49 validated signatures from COSMIC version 3. Signatures from COSMIC version 3 that are considered to be sequencing artefacts were not included in the primary analysis (e.g. SBS45 that is a possible artefact due to 8-oxo-guanine introduced during sequencing). We downloaded and processed the mutation signatures discovered by Zou et al.[21] from cell lines targeted with CRISPR-Cas9-based knockouts of DNA repair genes deposited at https://github.com/xqzou/NatComms_KOSig. We analyzed all knockouts that produced mutational signatures, i.e. *ΔEXO1*, *ΔMSH6*, and *ΔFANCC* knockouts.

6.2. Definition of *RePrint*

We present a formal definition for the *RePrint* of a given signature. Let $S(N_L[X>Y]N_R)$ be the signature emission probability, i.e. the probability with which the signature generates mutation category $N_L[X>Y]N_R$, where each mutation category is defined by the 5' flanking base N_L, the reference base X (always C or T by convention), the variant base Y, and the 3' flanking base N_R. The *RePrint* probability is given by

$$RePrint(N_L[X>Y]N_R) = \frac{S(N_L[X>Y]N_R)}{\sum_{Z\neq X} S(N_L[X>Z]N_R)}.$$

Please note, that while the signature probability S is a single categorical probability distribution over 96 mutational types, i.e. $\sum_{N_L,X,Y,N_R} S(N_L[X>Y]N_R) = 1$, *RePrint* of a signature is a vector of 32 categorical probability distributions – one distribution over 3 possible mutations for each of 32 nucleotide triplets, i.e. $\sum_{Y\neq X} RePrint(N_L[X>Y]N_R) = 1$ for each $N_L X N_R$ triplet.

It is possible that a mutagenic process does not create a mutation in some context. This reduces the information carried by *RePrint* and some similarities in the DNA repair might be missed.

Moreover, if mutations in some triples are very rare, this might create a noisy *RePrint*. To reduce noise and missing values in *RePrints* due to rare or missing mutations, we we add a small pseudocount (set to $\varepsilon = 10^{-4}$ in this work) to all signature probabilities, i.e. we computed *RePrints* for modified signature probability vectors $S'(N_L[X > Y]N_R) = S(N_L[X > Y]N_R) + \varepsilon$. Thus, for a triple that is rarely or not mutated in a given signature, corresponding *RePrint* conditional probability distribution for this triplet is almost uniform (all mutations of this triplet equally probable). We confirmed that the results did not depend qualitatively on this pseudocount value.

The signatures and signature *RePrints* were clustered using unsupervised hierarchical clustering, as implemented in R function `heatmap` with default parameters (Euclidean distance and complete linkage method) that was applied to a matrix of the root-mean-square error (RMSE) values computed between each pair of signatures or *RePrints*. When needed for comparison with signature RMSEs, *RePrint* RMSEs were adjusted by scaling factor $\sqrt{1/32}$ to account for that fact that *RePrint* is a vector of 32 probability distributions not just one. We also confirmed that using COSINE similarities instead of RMSE has not qualitatively impacted the results.

References

1. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. *Cell* **168**, 644–656 (2017).
2. D'Andrea, A. D. DNA Repair Pathways and Human Cancer. in *Molecular Basis of Cancer* (ed. John Mendelsohn Peter Howley Mark Israel Joe Gray Craig Thompson) 47–64 (Elsevier Health Sciences, 2014).
3. Miller, J. H. Mutagenic specificity of ultraviolet light. *J. Mol. Biol.* **182**, 45–65 (1985).
4. Giglia-Mari, G. & Sarasin, A. TP53 mutations in human skin cancers. *Hum. Mutat.* **21**, 217–228 (2003).
5. Hollstein, M., Sidransky, D., Vogelstein, B. & Harris, C. C. p53 mutations in human cancers. *Science* **253**, 49–53 (1991).
6. Van Hoeck, A., Tjoonk, N. H., van Boxtel, R. & Cuppen, E. Portrait of a cancer: mutational signature analyses for cancer diagnostics. *BMC Cancer* **19**, 457 (2019).
7. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLoS Genet.* **11**, e1005657 (2015).
8. Haradhvala, N. J. *et al.* Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat. Commun.* **9**, 1746 (2018).
9. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415 (2013).
10. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
11. Van Allen, E. M. *et al.* Somatic ERCC2 mutations correlate with cisplatin sensitivity in muscle-invasive urothelial carcinoma. *Cancer Discov.* **4**, 1140–1153 (2014).
12. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16 (2017).
13. Fischer, A., Illingworth, C. J. R., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, R39 (2013).
14. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs:

delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
15. Huang, X., Wojtowicz, D. & Przytycka, T. M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* (2017). doi:10.1093/bioinformatics/btx604
16. Alexandrov, L., Kim, J., Haradhvala, N. J. & Huang, M. N. The repertoire of mutational signatures in human cancer. *bioRxiv* **322859***,* (2018).
17. Morganella, S. *et al.* The topography of mutational processes in breast cancer genomes. *Nat. Commun.* **7**, 11383 (2016).
18. Benham, C. J. Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *J. Mol. Biol.* **255**, 425–434 (1996).
19. Kouzine, F. *et al.* Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. *Cell Syst* **4**, 344–356.e7 (2017).
20. Wojtowicz, D. *et al.* Hidden Markov models lead to higher resolution maps of mutation signature activity in cancer. *Genome Medicine* **11**, (2019).
21. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, 1744 (2018).
22. Viel, A. *et al.* A Specific Mutational Signature Associated with DNA 8-Oxoguanine Persistence in MUTYH-defective Colorectal Cancer. *EBioMedicine* **20**, 39-49 (2017).
23. O'Donovan, P. *et al.* Azathioprine and UVA light generate mutagenic oxidative DNA damage. *Science* **309**, 1871–1874 (2005).
24. Samal, I. R., Maneesh, M. & Chakrabarti, A. Evidence for systemic oxidative stress in tobacco chewers. *Scand. J. Clin. Lab. Invest.* **66**, 517–522 (2006).
25. Shen, H. M. *et al.* Detection of elevated reactive oxygen species level in cultured rat hepatocytes treated with aflatoxin B1. *Free Radic. Biol. Med.* **21**, 139–146 (1996).
26. Tate, J. G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2018).
27. Volkova, N.V. *et al.* Mutational signatures are jointly shaped by DNA damage and repair. *bioRxiv* **686295**, (2019).

Pacific Symposium on Biocomputing 25:262-273(2020)

Genome Gerrymandering: optimal division of the genome into regions with cancer type specific differences in mutation rates

Adamo Young[1,†,‡,⋆], Jacob Chmura[†], Yoonsik Park[†], Quaid Morris[†,‡,⋆], Gurnit Atwal[‡,⋆]

[†]*Department of Computer Science, University of Toronto,*
40 St. George Street, Room 7224
Toronto, ON M5S 2E4, Canada

[‡]*Donnelly Centre for Cellular and Biomolecular Research,*
160 College Street, Room 230
Toronto, ON M5S 3E1, Canada

[⋆]*Vector Institute for Artificial Intelligence,*
661 University Ave, Suite 710
Toronto, ON M5G 1M1, Canada

[1]*E-mail: adamo.young@mail.utoronto.ca*

The activity of mutational processes differs across the genome, and is influenced by chromatin state and spatial genome organization. At the scale of one megabase-pair (Mb), regional mutation density correlate strongly with chromatin features and mutation density at this scale can be used to accurately identify cancer type. Here, we explore the relationship between genomic region and mutation rate by developing an information theory driven, dynamic programming algorithm for dividing the genome into regions with differing relative mutation rates between cancer types. Our algorithm improves mutual information when compared to the naive approach, effectively reducing the average number of mutations required to identify cancer type. Our approach provides an efficient method for associating regional mutation density with mutation labels, and has future applications in exploring the role of somatic mutations in a number of diseases.

Keywords: Genome Segmentation, Tumour Classification, Dynamic Programming, Information Theory

1. Introduction

Somatic cells are exposed to multiple mutational events throughout their lifetime. The phenotypic effect of these mutations varies, and the aggregate effect of all somatic mutations has been implicated in the development of a number of neurodegenerative diseases and cancer [1], [2]. Somatic mutations are generated by multiple mutational processes ranging from exogenous mutagens to endogenous DNA repair mechanisms. Mutational processes are mechanisms for generating different types of mutations, and their signal in the genome is manifested through different mutational signatures. Single base substitution signatures (SBS) summarize muta-