The analysis performed in this study shows that breast cancer is predominantly impacted by SBS1, SBS2, SBS5, SBS10b, and SBS13. These mutational signatures are found in the exome regions including TSS and TTS. This is clearly observed from the mutational signatures extracted in every step of the analysis. However, the processes behind these signatures operate at different levels depending on the regions. For instance, TTS regions are more mutated than TSS regions where all signatures produce roughly the same number of mutations (apart from SBS5). Whereas, in TTS regions, signatures exhibit different activity rates suggesting either an increased level of deleterious events at TTS or better repair mechanisms around TSS or both. Strikingly, the signature of unknown etiology SBS5 exhibit a 151% increased activity around TTS compared to TSS. This increase might be potentially associated with the mutational process behind this signature. The mutational process behind SBS5 could be related to replication processes considering that SBS5 is correlated with aging. Comparably, APOBEC signatures also present an increased activity in TTS regions with a 136% increase for SBS13 and a 126% increase for SBS2. The difference between the two signatures requires more investigation. HRD signature SBS3 and the spontaneous deamination of cytidine (SBS1) are also presenting increased rates of 109% and 94,5%, respectively. Contrastingly, the POLE signature presents a slight decrease (11%) of activity around TTS regions. These observations hold true for active regions. However, SBS1 shows no differences between the two regions in the inactive contexts, suggesting that the transcription mechanism may be associated somehow with this signature. Additionally, there are only slight differences in the numbers of mutations and the activities of signatures occurring between convergent regions and R-loops around TTS, considering that R-loops are particularly enriched in TTS regions of convergent genes. This raises an important question on whether the mutations observed in the regions between convergent genes are associated with R-loops.

During this analysis, some mutational signatures that are not supposed to be associated with breast cancer arose and it is not clear whether they were truly present or just artifacts. Namely, SBS42 (Haloalkane exposure), SBS7a, SBS7b (Ultraviolet light exposure), SBS35 (Platinum chemotherapy), SBS29 (Tobacco chewing), and SBS4 (Tobacco smoking). While it is true that their attribution increased the level of error and reduced the similarity between the extracted signature and the signature reconstructed with COSMIC reference signatures, their attribution is not random at all. Mostly, these signatures either replaced a similar signature associated with the same process (e.g, SBS4 replaced SBS29), or they were shown to be present at a rare prevalence in breast cancer (e.g, SBS7b) (Degasperi et al., 2022). This suggests that such signatures might be truly present at low levels in a few samples and the reduced number of mutations allowed them to emerge. However, such signatures should be carefully interpreted. In the same context, it is important to note that SBS1 and SBS5 are not direct measures of age. In fact, they are described as clock-like signatures because they accumulate at a constant rate. Whether this rate is associated with time or cell divisions is unclear.

It is important when interpreting the results to keep in mind that deciphering mutational signatures from mutational catalogs is an NP-hard problem. This means that it is impossible to accurately solve the problem in a reasonable amount of time. Hence, all the methods available are mathematical approximation methods that proved robust and biologically meaningful through experiments. This also means that the obtained results should be carefully interpreted. Another point to consider is that bulk sequencing data hold different catalogs shaped by different signatures. Therefore, the signatures extracted from such datasets might vary between different cohorts. In fact, the cancer heterogeneity between cancer subtypes and between patients plays a role in shaping the results. The age of the patient and the stage at which cancer has developed are also impactful. Additionally, these catalogs are also impacted by different technical factors. The mutation calling algorithms function differently and provide slightly different mutational profiles, which might impact the results. Hence, it is important to perform the analysis several times with different calling algorithms to ensure consistency. Moreover, the sequencing technology itself is crucial. Every technology has its advantages and inconveniences. This study was performed using WES data, which reduced the number of mutations necessary for the algorithm to function but also discarded important parts from the TSS and TTS regions, because of this, the numbers of mutations in these regions are probably under-estimated. While it is unlikely to see differences in the mutational signatures, it is possible that the activity rates of the signatures change. Therefore, performing the analysis in the future on Whole Genome Sequencing (WGS) is important. In addition, flat signatures such as SBS5 and SBS3 could be underestimated as well due to WES limitations (Abbasi et al., 2021).

The results provided by this analysis raise several questions that could be addressed to get a better view concerning the mutational signatures associated with replication in cancer.

For future directions, it would be interesting to extend the study to doublet-base and INDEL signatures. This will allow for a comprehensive characterization of the signatures in the cancers of interest. It could also help identify signatures associated with replication stress. In fact, it was shown that topoisomerase 1 depletion increases the risk of double-strand breaks around TTS containing R-loops (Promonet et al., 2020). The repair of these DSBs could leave characteristic signatures. Another important direction is to investigate the increased levels of mutations around TTS compared to TSS and to look for any associations with TRCs. It is equally important to investigate the association between the mutations happening between convergent genes and those co-localizing with R-loops around TTS regions as they display striking similarities in terms of signatures’ activities. Such analysis could benefit from multi-omics approaches and will provide new insights into the understanding of cancer genomics and cancer development. Ultimately, it would be valuable if signatures related to replication stresses get identified. This will provide a better understanding of the consequences of such phenomena and may also provide opportunities for clinical applications, such as stratification of patients based on these molecular footprints or the development of new targeted therapies. Finally, the field might profit from new deciphering methods that do not require a huge amount of data. Despite the fact that mutational catalogs are getting more and more available, studies limited to specific regions of the genome could be challenging for the current approaches. Deep learning proved to be useful and overcame the classical methods in different fields. Hence, it could be interesting to consider developing a deep-learning-based approach and even better to consider an explainable model. This could allow associating signatures to specific mutations without the need to limit the data to regions of interest.

Abbasi, A., & Alexandrov, L. B. (2021). Significance and limitations of the use of next-generation sequencing technologies for detecting mutational signatures. *DNA repair*, *107*, 103200. <https://doi.org/10.1016/j.dnarep.2021.103200>