The analysis performed in this study shows that the same mutational processes operate in the exome region including TSS and TTS. This is clearly observed from the mutational signatures extracted in every step of the analysis. However, these processes operate at different levels depending on the region. 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 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 suggesting that it could potentially be associated with replication processes, especially that it is correlated with the age of the individuals. Furthermore, it presents statistical association with NER deficiency and tobacco in bladder cancer (cite the cosmic webpage). APOBEC signatures do present an increased activity at TTS regions as well with a 136% increase for SBS13 and 126% increase for SBS2. The difference between the two signatures requires more investigations. 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 show no differences between the two regions in the inactive contexts, suggesting that the transcription mechanism may be associated somehow with this signature. Furthermore, it seems that APOBEC signature SBS13 is absent from R-loops occurring around TSS regions and SBS2 exhibit an increased activity of 76% in R-loops occurring around TTS, this may rise a potential direction to distinguish between SBS2 and SBS13. Additionally, there are only slight differences in the numbers between mutations occurring between convergent genes and mutations co-localizing with R-loops, confirming that R-loops are particularly enriched in TTS regions of convergent genes. Unfortunately, this data cannot tell if TRCs are responsible for the mutations in these areas.

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, SBS15 (defective MMR), 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 reconstructed one using the reference signatures, but their attribution is not random at all. In fact, they either replaced a similar signature that is associated with the same process (SBS15-SBS6, SBS4-SBS29), or they were shown to be present at a rare prevalence in breast cancer (Signal ref). This suggests that such signatures might be truly present at low levels and the reduced number of mutations allowed them to emerge. Nonetheless, such signatures should be carefully interpreted. In the same context, it is important to note that SBS1 and SBS5 are not a direct measure of the 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 solve the problem precisely in a reasonable amount of time. Hence, all the methods available are approximation methods that proved robust and biologically meaningful through experiments. This also means that the obtained results can only be interpreted on

context of a big picture, rather than a detailed perspective. Another point to consider is that bulk sequencing data hold different catalogs shaped by different signatures. Therefore, the signatures extracted from this dataset might be different from the results extracted from other 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. Furthermore, the lifestyle of the patient, their genetic backgrounds, and the treatments they had, are as well equally important. Furthermore, the mutation calling algorithms function differently and provide slightly different catalogs which might impact the results. Performing the analysis several times with different algorithms is important for consistency. Finally, 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 when they are considered, they could however impact the activity rates of the signatures. Therefore, performing the analysis on Whole Genome Sequencing (WGS) is planned.

The results provided by this analysis raise several questions that could be addressed to get a better view concerning

for future directions

replication intrinsically embedded in all signatures?