Inclusion of opportunities and large nucleotide contexts give robust and accurate mutational signatures

Mutational signatures describe the pattern of mutations over the different mutation types. Usually, mutation types are defined by the base mutation and the flanking nucleotides at the right and left of the base mutation. We extend this definition and include flanking nucleotides further away from the base substitution.

The importance of larger contexts lies in the way they affect mutation rates. Furthermore, when looking at more flanking nucleotides, it becomes relevant to consider mutational opportunities for different mutation types. Indeed, sets of nucleotide sequences occur with various rates along the genome, and this results in very different opportunities for the mutation types to occur.

Mutational signatures are usually derived using non-negative matrix factorization (NMF) in tri-nucleotide contexts. We show that using parametrized signatures combined with opportunities provides more robust and interpretable signatures. Using cross-validation, we also show that including opportunities increases the predictive power of the model on new data. By analyzing data where mutation types are defined considering three, five, and seven nucleotides, we also show that the advantages of including opportunities in the model become even more pronounced with more flanking nucleotides. Overall, the use of mutational opportunities and larger contexts provides more robust signatures and a more accurate representation of the underlying mutational processes.