signeR: Signature finder for mutational processes in cancer genomes

Rodrigo Drummond (AC Camargo Cancer Center, Brazil), Renan Valieris (AC Camargo Cancer Center, Brazil), Rafael Rosales (Universidade de São Paulo, Brazil), Israel Tojal (AC Camargo Cancer Center, Brazil)

BACKGROUND: Somatic mutations found in cancer genomes are the result of DNA damage and repair processes present in tumor cells. A considerable effort has been made in order to characterize somatic mutational patterns in cancer. Recently, a powerful feature extraction method based on Non-negative Matrix Factorisation (NMF) techniques has shown to be quite effective in this context. Briefly, an input matrix of somatic mutation (SNV) counts is decomposed as the product of two smaller matrices: a matrix of mutation signatures, each corresponding to a particular mutational process, and a exposure matrix containing the relative contribution of each mutational process to the observed counts. This decomposition may however be not unique and the rank of the factorisation, i.e. the number of mutational processes, also has to be estimated. **HYPOTHESIS:** The identification of mutational signature patterns may unravel the processes responsible for cancer development providing a better understanding of its causes. **METHODS**: We considered an empirical Bayesian approach to NMF. This allows for the estimation of the underlying mutational processes and the NMF model dimension. The method's first step is to obtain a matrix of SNV counts, in which the mutations observed in each analysed sample are divided according to the mutated base and its flanking basis. Then it is assumed that the (i,j)-entry of this matrix is a Poisson random variable with rate $\lambda=(PE)_{ij}W_{ij}$, where P is the matrix of mutation processes (signatures), E is the matrix of exposures of samples to processes, and W is a fixed matrix of mutational opportunities (the counts of the triplets where each kind of mutation can occur in each sample). Both the P and E matrices together with the hyperparameters of a conjugate hierarchical model are estimated by using Markov chain Monte Carlo (MCMC) methods. These methods produce consistent estimates for P and E and also for the uncertainty related to these. The MCMC/Bayesian analysis developed here also provides the means to asses the NMF model rank via the computation of the marginal likelihood. **RESULTS**: We demonstrate the application of our method in a data set formed by 21 breast cancer genomes. Remarkably, a mutational pattern associated with the APOBEC genes was found. APOBECs are a family of DNA/RNA editing enzymes, which catalyze cytosine to uracil deamination on single-stranded DNA. Our results highlight APOBEC as having a key role in the destabilization of the genome, suggesting that this mechanism is a source of somatic mutation associated with breast cancer. We conclude that our method is a valuable tool to improve our understanding of cancer etiology, being therefore useful for clinical applications. The design of the MCMC algorithm developed here presents a combination of speed and low memory overhead that makes it feasible to be run on a standard computer. It was written in the R programming environment and low level functions were implemented in C++ to improve performance.