CRUK

Jorge Lemos Portela

June 30, 2021

1 Introduction

Throughout our lives our genome experiences many mutations. Besides ageing, there are many other factors that lead to mutations in the genome like smoking tobacco or excessive exposure to ultraviolet light. Some of these mutations will not have a significant effect in our health but some others may lead to serious diseases, including cancer. Thus, understanding the biological processes generating these mutations is crucial.

With the technologies that we have nowadays we can compare the reference genome for a patient with the sequenced genome of a cell of interest. This allows us to see all the aberrations present in that cell. However, there is no direct information about when these mutations took place.

Given the information of the mutations from a group of cancer patients, the signature analysis tries to find patterns in the mutations with the intention of clustering the mutations of interest and potentially understanding the cause behind them.

Signature analysis consists mainly in three steps. The first step is feature extraction. In this step, we need to define the features of the genome that will be studied and extract the data about these features for each sample. The second step is signature identification. With the help of mathematical methods, the data about the mutations collected in the previous step is clustered in signatures. Moreover, each sample will be defined as a combination of these signatures. The last step is signature assignment, where the signatures will be analysed and we will look for possible relations between these signatures and known biological processes.

In general, the first step of signature analysis is crucial, since different features will lead to different results. Once we have made a final decision for the features, we also need to decide on a clustering method for the signatures which, as we will see, may make a difference in the final results. Finally, the last step is the most challenging, as we need to find possible associations and there is no clear guidance for this and also sometimes not much data from the patients.

In 2013, one of the first papers in this area [1] was published. In this article, they analysed the Single Nucleotide Variants (SNVs) in a collection of samples from different types of cancer. By following the steps described above, they created a catalogue of 30 distinguishable signatures that were later associated with biological processes.

For this paper, the collected information about the SNVs was the substitution at the variant position (6 possibilities) and the nucleotides at the previous and posterior positions (16 possibilities), so 96 possible categories for each

SNV. Thus, the first step of feature extraction consisted in counting, for each sample, the number of mutations belonging to each of the categories. For the second step, as they explain in the methods section, all mutational signatures were clustered using unsupervised agglomerative hierarchical clustering and a threshold was selected to identify the set of consensus mutational signatures. Mis-clustering was avoided by manual examination. Once the mutational signatures were defined and each sample defined as a combination of signatures, Generalised Linear models (GLMs) and Kolmogorov–Smirnov tests were used to relate the signature to causes. This causes include ageing, use of tobacco, specific mutations of interest...

The promising results from this paper increased the general interest in the study of mutational signatures. As we have seen, [1] gave a possible method to define and extract the signatures. But there were many other possibilities to explore. Moreover this paper studied 30 different cancer types and tried to extract general signatures for all of them.

There have been numerous papers about this topic in the recent years. The different data sets and choices during the signature analysis give rise to different results. For the purpose this work we are particularly interested in two of these papers [2,3].

On the one hand [2] focused in ovarian carcinoma. This is a copy number driven type of cancer so the idea was to adapt the steps for SNVs that were mentioned above to the case of copy number aberrations in the genome. As we will see later this gives rise to a whole new range of choices to explore. For the step of signature identification they used Non-negative Matrix Factorization (NMF) which had been previously used for the case of SNVs [1]. Since there is no guarantee that NMF is the best way to cluster the features and extract the signatures, [3] presented a new method for signature identification in the case of SNVs. This method was based on Expectation Maximization (EM) and had some advantages with respect to NMF such as improved stability of statistical estimation, and easier implementation of additional components for the features.

The goal of this paper will be to combine the ideas of [2] and [3]. Using the same data that was used in [2] to analyse ovarian cancer samples, we will adapt the methods of [3] to the case of copy number aberrations in order to re-derive the signatures and then compare them with the current signatures that were identified in [2].

In the first section we will explain in detail the methods used in [2] for signature analysis, as they will be the base of our work. After that, the methods from [3], which give an alternative to NMF, will be explained. In the third section we will explain how the methods from [3] were adapted from the case of SNVs to the case of copy number analysis. Then, the fourth section will present the main results once we use this method with the data of ovarian cancer samples from the feature extraction, to signature identification and assignment. The results will be compared to the previously defined signatures and linked to other clinical information. The fifth section explores other possibilities when we make different choices and combines some of the alternatives to create a more complex yet more complete scenario. Finally there will be discussion of the results and some comments on possible future work.

2 A first method for the analysis of copy number signatures

In this section the methods for copy number signature analysis from [2] will be explained. This paper was written (among others) by Geoff Macintyre. Thus, for ease of understanding, we will refer to the procedures of [2] as '*Geoff's method*'. The article studies copy number signatures from shallow whole-genome sequencing of 117 high-grade serous ovarian cancer (HGSOC) cases. The motivation to study copy number aberrations instead of SNVs is that ovarian cancer is characterized by changes in copy number.

The basis of the procedure for the analysis of copy number aberration (CNA) signatures will be similar to the one for SNVs, but there will be some additional complications. For starters, in the case of SNVs the features were clear, the interesting position was the variant baseposition, and then there was information about inmediate 3' and 5' bases to that position. On the other hand, CNAs can affect pieces of the genome of different sizes, there are multiple possibilities for the number of copies and also many other characteristics about the neighbor regions to explore. This implies that, for CN signatures, there will be some challenging decisions to be made even for the step of feature selection.

Let us explain the main steps of Geoff's method for CN signature analysis, which can be seen in Figure **??**. First, the absolute copy number is derived from the shallow whole genome sequencing data. Next, the features to study need to be chosen, as was mentioned above this is a crucial step. Once the features are selected, the genome-wide distributions of these features are computed. Each of these distributions can be seen as a mixture of either Poisson or Normal distributions. Thus, using a method of mixture models, the distributions for the features are divided into combinations of Poisson or Normal distributions, which will be refered to as the components of the feature. We are interested in having a tumor by component matrix that would reflect, for each sample, the presence of each component in that tumor. For each CNA event, there is a posterior probability that this event belongs to each of the components. This posterior probabilities are summed for all the CNA of each sample, creating a tumor by component matrix, where the rows are the samples and the columns are the intensities of each component in that sample.

Once we have the tumor by component matrix, we need to decide the number of signatures that we are going to be clustering the samples in. This is a delicate decision. Besides the interesting mutations that we are interested in, there is always some noise in the experiments, either from mutations that are unrelated to the cancer or from experimental noise. A higher number of signatures will focus on small details of the samples and this will not be ideal when the noise in the samples is relatively high. Also, we need to keep the final goal in mind, which is to relate the signature to possible causes. This will be more complicated if we have a high number of signatures or if they are highly correlated. Thus, there will not be a clear ideal number of signatures and we will just need to find a sufficiently good number of signatures to collect the most important information from the samples, that does not give much importance to the noise and which can be correlated later with possible causes.

Taking all of this into account the Non-negative Matrix Factorization (NMF) is used to decompose the tumor by component matrix into a tumor by signature matrix and a signature by component matrix. The NMF package in R [4] allows us to do this decomposition and it also calculates the cophenetic, dispersion, silhouette and sparseness coefficients for the resulting matrices. A range of possible signatures from 3 to 12 was selected and the NMF was used for each of this possibilities. The resulting coefficients were used to decide that the optimal number of signatures that yielded both high stability and spareseness was 7 signatures.

3 Summary working flow

Started with Geoff's data, the information is stored as QDNA sequencing format. We use the first part of the 'CN_signature_identification' document to get a clean dataset called 'all_CN'. This data set contains, by columns: The sample ID, the

References

[1] Alexandrov, L., Nik-Zainal, S., Wedge, D. *et al.*; Signatures of mutational processes in human cancer. Nature **500**, 415–421 (2013).

[2] Macintyre, G., Goranova, T.E., De Silva, D. *et al.*: Copy number signatures and mutational processes in ovarian carcinoma. Nat Genet **50**, 1262–1270 (2018).

[3] Shiraishi Y, Tremmel G, Miyano S, Stephens M: A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. PLOS Genetics 11(12): e1005657, (2015) .

[4] NMF package in *R*.