Chapter 2

Deciphering signatures of mutational processes from mutational catalogues of cancer genomes

2.1 Introduction

The first chapter of this thesis defined *somatic mutations* as any change of DNA that is present in the genome of a somatic cell and has occurred after conception. Building upon this well-known definition, the chapter introduced several important concepts. A *somatic mutational process* was defined as a mixture of DNA damaging and repair mechanisms that act collectively and have the ability to cause mutations in somatic cells. A *mutational signature* was described as a characteristic pattern of somatic mutations exhibited by an operative mutational process in a genome of a cell. Lastly, a *mutational catalogue* of a cancer genome was defined as the conglomeration of all detected somatic mutations.

The main focus of the present chapter is to mathematically connect these biological terms and provide both the theoretical model and computational approach for examining and deciphering mutational signatures from sets of mutational catalogues of cancer genomes. The approach is evaluated extensively with simulated data, demonstrating that the developed computational framework is robust to a large range of different parameters and can be applied to both genome and exome sequences.

2.2 Theoretical model of mutational processes operative in cancer genomes

The mutational catalogue of a cancer genome is the cumulative result of all somatic mutational mechanisms, including DNA damage and repair processes, which have been operative during the cellular lineage of the cancer cell. Since the cellular lineage of the cancer cell can be traced back to the zygote, the mutational catalogue reflects the activity of all processes operative from the very first division of the fertilized egg (Stratton, 2011). The large majority of mutations in cancer genomes are believed to be passengers, and by definition their patterns are largely unmodified by selection (Rubin and Green, 2009). Thus, the mutational catalogue derived from a cancer cell may be treated as a representative archaeological record bearing the combined imprints (or signatures) of the mutational processes that have been operative.

2.2.1 Alphabets of mutation types

A mutational catalogue can include a diverse set of mutation classes including base substitutions, insertions/deletions, structural rearrangements and copy number changes. Each class of mutation can then be further subclassified. For example, base substitutions can be subclassified according to the six types of single base substitutions (using the pyrimidine of the Watson-Crick base pair as the reference, C>T, C>A, C>G, T>A, T>C, T>G) or the classification can be further elaborated to include a variety of mutational features such as the sequence context of the mutated base and the transcriptional strand on which the substitution has arisen.

For the purpose of mathematical modelling, a limited number of features of a mutational catalogue need to be selected. The choice of features may be influenced by prior biological knowledge. The choice is also often constrained by statistical considerations and the available data. Mathematically, a set of mutational features can be expressed as a finite alphabet Ξ with K letters, where each letter corresponds to a mutation feature. The simplest alphabet in this case, Ξ_6 contains $K = 6$ letters, and is based on the 6 types of single base substitution. The letters of this Ξ_6 alphabet are C>A, C>T, C>G, T>A, T>C, and T>G. It should be noted that this alphabet of mutation types could be easily extended by, for example, including other mutation types such as double substitutions.

In this thesis, mutational catalogues as well as the mutational signatures that contribute to these catalogues are examined predominantly using five distinct alphabets termed Ξ_6, Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536}. These five alphabets are discussed in further detail below as well as in Appendix I.

The Ξ_6 alphabet is perhaps the simplest possible alphabet as it considers only the six types of somatic substitutions. This alphabet will not be used in any analysis but, rather, its simplicity will be leveraged to provide examples and visual representations clarifying the developed mathematical model and computational approach.

The Ξ_{96} alphabet provides greater resolution for examining the six types of single nucleotide variants (*i.e.*, the Ξ_6 alphabet) by including the immediate sequence context of each mutated base. In this alphabet, a mutation type contains a somatic substitution and both the 5' and 3' base next to the somatic mutation. For example, a C>T mutation can be characterized as …Tp<u>C</u>pG…>…Tp<u>T</u>pG… (mutated base underlined and presented as the pyrimidine partner of the mutated base pair) generating 96 possible mutation types – (6 types of substitutions) * (4 types of 5' bases) * (4 types of 3' bases).

The Ξ_{1536} further extends Ξ_{96} by including two bases 5' and 3' to the mutated base resulting in 1,536 possible mutated pentanucleotides - (6 types of substitutions) * (16 types of the two immediate 5' bases) * (16 types of the two immediate 3' bases). For example, using the Ξ_{1536} alphabet, one of the 256 subclasses of a C>T mutation is …ApTp<u>C</u>pGpC… > …ApTp<u>T</u>pGpC…

The Ξ_{99} alphabet extends Ξ_{96} by including three additional mutation types, *viz.*, (i) double nucleotide substitutions, (ii) small insertions or deletions at short tandem repeats, and (iii) small insertions or deletions overlapping with microhomologies at breakpoints.

Lastly, Ξ_{192} elaborates Ξ_{96} by considering the transcriptional strand on which a substitution resides. In contrast to all previously discussed alphabets, Ξ_{192} is defined only in the regions of the genome where transcription occurs, which in these analyses has been limited to the genomic footprints of protein coding genes. Thus, the previously defined 96 substitution types are extended to 192 mutation types. For example, the C>T mutations at TpCpA are split into two categories: the C>T mutations at TpCpA occurring on the untranscribed strand of a gene and the C>T mutations at TpCpA occurring on the transcribed strand. In general, one would expect that these two numbers are approximately the same unless the mutational processes are influenced by the activity of the transcriptional machinery. This could happen, for example, due to recruitment of the transcription-coupled component of nucleotide excision repair (NER). For example, if a mutational process has a higher number of C>A substitutions on the transcribed strand compared to C>A substitutions on the untranscribed strand (note that a C>A mutation on the untranscribed strand is the same as a G>T mutation on the transcribed strand), this could indicate that the mutations caused by this process are being repaired by NER, although other explanations are not excluded. A known example of such strand-bias due to interplay between a mutational process and a repair mechanism is the formation of photodimers due to ultraviolet light exposure that are repaired by NER resulting in a higher number of C>T mutations on the untranscribed strand (van Zeeland et al., 2005).

2.2.2 Mathematical definition of a signature of a mutational process

A signature of a mutational process is mathematically defined in the context of a pre-selected mutational alphabet. A mutational signature is defined as a discrete probability density function with a domain of mutation features based on a pre-selected alphabet Ξ, $P\colon \Xi \to \mathbb{R}_{+}^{K}$. Thus, by definition, a mutational signature P is a lexicographically ordered k-tuple; $P = [p^1, p^2, ..., p^K]^{\mathrm{T}}$, where p^i is the probability of process P to cause the mutation feature corresponding to the i-th letter of the pre-selected alphabet Ξ, and since p^i are probabilities:

$$\sum_{i=1}^{K} p^i = 1 \text{ and } p^i \geq 0, i = 1 \dots K \qquad (2.1)$$

Examples of four mutational signatures defined over Ξ_6 and two mutational signatures defined over Ξ_{96} are given respectively in panels A and B of Figure 2.1. In the four examples of mutational signatures defined over Ξ_6, the mutational probability for each alphabet letter is displayed. For example, it can be seen that 35% of the mutations attributed to Signature 1 are C>G while only 3% of the mutations are T>G.

Further, while Signatures 1 through 4 are defined over Ξ_6, Signature 2 does not generate any C>T, T>A, and T>C mutations as the probability for each of these mutation types is equal to zero. Signatures 1 and 4 are defined both over Ξ_6 and Ξ_{96} to illustrate that, while a mutation type based on a given alphabet can be similar in two signatures (*e.g.*, C>A mutations are respectively 12% and 14% in these two signatures). Extending this alphabet may reveal an intrinsic internal structure making these mutation types significantly different.

Geometrically, a mutational signature can be examined as a vector in a *K* dimensional space. Since a mutational signature is modelled as a discrete probability density function defined over a given alphabet (see equation 2.1), all its components

Figure 2.1

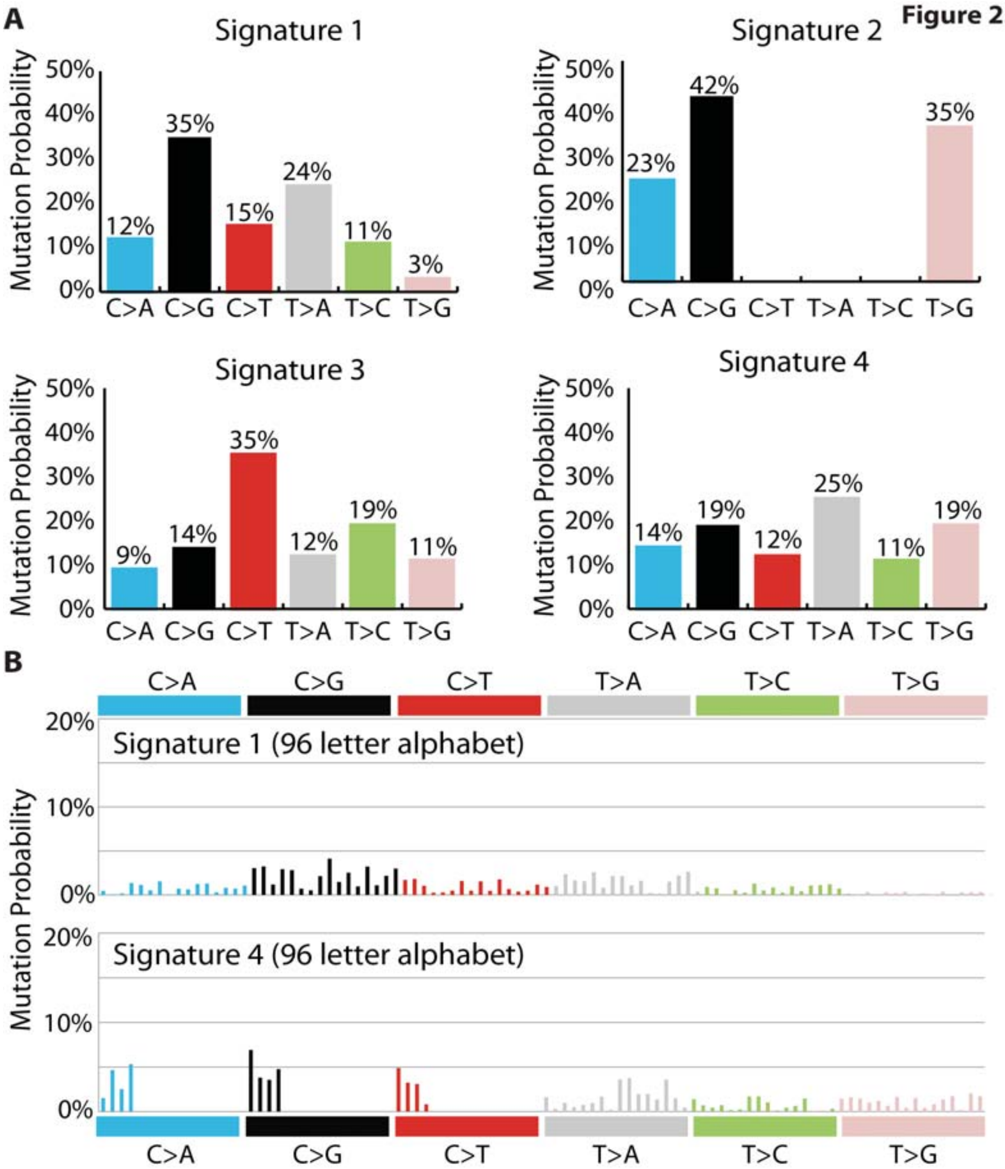
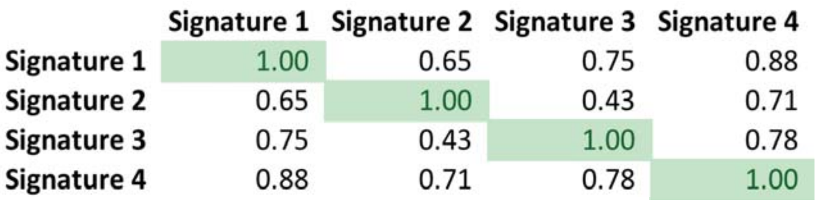


Figure 2.1: Simulated examples of mutational signatures defined over different mutational alphabets. (A) Four mutational signatures defined over Ξ_6 and **(B)** two mutational signatures defined over Ξ_{96}.

are nonnegative and this vector belongs to the first hyperoctant of this K dimensional space, $\mathbb{R}_{+}^{K}$. Further, as the sum of the vector components equals one, this vector is constrained by K-1 dimensional hyperplane. Examining two mutational signatures as vectors in a high dimensional space allows a convenient way for comparing these signatures based on the angle between the vectors. Thus, comparison between two mutational signatures $\vec{A}$ and $\vec{B}$, each defined over an alphabet Ξ with K mutation types, is done using a cosine similarity:

$$sim(\vec{A},\vec{B}) = \frac{\sum_{i=1}^{K} \vec{A}_i \, \vec{B}_i}{\sqrt{\sum_{q=1}^{K} (\vec{A}_q)^2} \sqrt{\sum_{j=1}^{K} (\vec{B}_j)^2}} \qquad (2.2)$$

Since the elements of $\vec{A}$ and $\vec{B}$ are nonnegative, the cosine similarity has a range between 0 and 1. When a cosine similarity between two signatures is 1, these signatures are exactly the same. In contrast, when the similarity is 0, the mutation types of these signatures are completely independent. The cosine similarity is a commutative function as $sim(\vec{A},\vec{B}) = sim(\vec{B},\vec{A})$. Two signatures should be compared only if they are defined over the same mutational alphabet. For example, one cannot compare a signature defined over Ξ_6 with a signature defined over Ξ_{192}. Lastly, one can also define a cosine distance between two mutational signatures as $dist(\vec{A},\vec{B}) = 1 - sim(\vec{A},\vec{B})$.



	Signature 1	Signature 2	Signature 3	Signature 4
Signature 1	1.00	0.65	0.75	0.88
Signature 2	0.65	1.00	0.43	0.71
Signature 3	0.75	0.43	1.00	0.78
Signature 4	0.88	0.71	0.78	1.00

Table 2.1: Similarities between simulated mutational signatures. The values of the cosine similarities between the signatures displayed in panel A of Figure 2.1 are shown in this table.

Table 2.1 contains the similarities between the simulated mutational processes displayed in Figure 2.1A. The two signatures that are most similar are Signatures 1 and Signature 4 with a cosine similarity of 0.88 while the signatures that are most different are Signatures 2 and 3 with a similarity of only 0.43. As expected, the similarity of Signatures 1 and Signature 4 is not the same when the signatures are defined and compared over different mutational alphabets. While Signatures 1 and Signature 4 have a similarity of 0.88 when defined over Ξ_6, they have a similarity of only 0.53 when defined over Ξ_{96}. As previously mentioned, this is due to the existence of an internal structure. In this simulated example, all C>X mutations belonging to Signature 4 are in ApCpN sequence context while Signature 1 has no specific sequence context (Figure 2.1).

2.2.3 Mathematical definition of a mutational catalogue of a cancer genome

Quantitatively, a mutational catalogue of a cancer genome is a vector, m, containing the number of somatic mutations of a genome, g, defined over a finite alphabet of mutation types Ξ. Mathematically, a mutational catalog is a morphism between the pre-defined finite alphabet, Ξ, and a set of K nonnegative integers, $\mathbb{N}_0^K$, *i.e.*, $m: \Xi \to \mathbb{N}_0^K$. Thus, for a given genome, its mutational catalogue can be expressed as a K-tuple of natural numbers, $m = [m^1, m^2$, ..., $m^K]^T$. A simulated example of a cancer genome defined over the mutational alphabet Ξ_6 is provided in Figure 2.2. This cancer genome has a total of 3,315 somatic substitutions and does not have any specific mutational features.

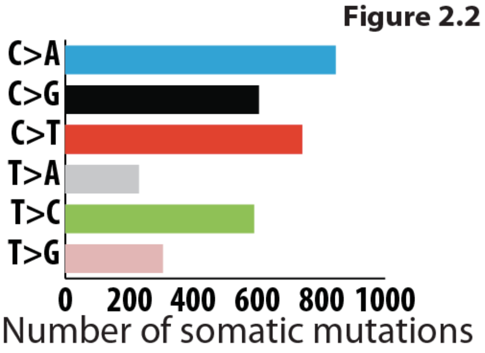


Figure 2.2: Simulated example of a mutational catalogue of cancer genome. The mutational catalogue is defined over the Ξ_6 alphabet.

Comparing the mutational catalogues of two cancer genomes, $m_1 = [m_1^1, m_1^2$, ... $m_1^K]^T$ and $m_2 = [m_2^1, m_2^2$, ... $m_2^K]^T$, requires that both mutational catalogues are defined over the same mutational alphabet Ξ. The similarity of two mutational catalogues can be evaluated in two distinct ways. The first comparison is based on Euclidean distance and examines whether mutational catalogues m_1 and m_2 are exactly the same:

$$dist(m_1, m_2) = \sqrt{\sum_{i=1}^{K} (m_1^i - m_2^K)^2} \qquad (2.3)$$

With this comparison, a distance of zero is equivalent to the two mutational catalogues being exactly the same. Further, the larger the distance the more different the mutational catalogues.

While two mutational catalogues can have different numbers of somatic mutations (and therefore a large Euclidean distance between them) they can have exactly the same patterns of somatic mutations. Thus, a correlation distance is used to compare whether the patterns of mutations of two mutational catalogues are similar. The simplest correlation distance is based on the Pearson product-moment correlation coefficient. However, this correlation coefficient is very sensitive to outliers and it might be misleading if a small subset of mutation types have significantly larger values when compared to the rest of the mutation types (Abdullah, 1990). More robust measurements of correlations are Spearman’s rank correlation coefficient and Kendal’s rank correlation coefficient (Croux and Dehon, 2010). These two rank correlations usually produce very similar results and rarely is there a reason to choose one over the other (Croux and Dehon, 2010). In this work, I make use of Spearman’s correlation coefficient to compare the patterns of mutations in two mutational catalogues. Spearman’s correlation is defined as the Pearson’s correlation coefficient between the ranked variables. Thus, the patterns of mutations in mutational catalogues m_1 and m_2 can be compared by the formula:

$$\rho = \frac{\sum_{i=1}^{K}(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{K}(x_i - \bar{x})^2 \sum_{i=1}^{K}(y_i - \bar{y}\,)^2}} \tag{2.4}$$

where x_i is the rank of the i-th letter of the pre-selected alphabet Ξ in m_1, y_i is the rank of the i-th letter in m_2, $\bar{x}$ is the mean of $x_i, i = 1 \ldots K$, and $\bar{y}$ is the mean of $y_i, i = 1 \ldots K$.

In general, the Euclidean distance will be used to compare two mutational catalogues when one wants them to be as similar as possible. For example, a Euclidean distance will be used when extracting mutational signatures and evaluating the accuracy of the extraction. In contrast, the correlation distance will be used to compare the similarity of the patterns of somatic mutations between two mutational catalogues. For example, a correlation distance will be used when performing clustering of cancer genomes in order to identify distinct groups of mutational patterns.

2.2.4 Modelling mutational processes operative in a cancer genome

In the previous sections of this chapter, I provided mathematical definitions for mutation types, mutational signatures, and mutational catalogues. In this section, I make use of these definitions to provide a linear model of mutational processes operative in cancer genomes.

Different cancer genomes can be exposed to a particular mutational process at different intensities. For example, a mutational process could cause 1,000 mutations in one cancer genome while causing 20,000 in another. I will refer to this number of mutations as a *mutational exposure* (or simply *exposure*) of a signature of a mutational process in a cancer genome. Hence, one may say that a mutational process with a signature P has an exposure e, corresponding to the number of mutations caused by this process, in a mutational catalogue m of a given cancer genome.

Multiple mutational processes can be operative in a single cancer genome (Stratton, 2011) and each of these processes can have a distinct mutational exposure. In this section, I model a cancer somatic mutational catalogue as a linear combination of the signatures and intensities of the exposure of the mutational processes active at some point in the lineage of cells leading to the cancer cell, plus added noise vector accounting for non-systematic sequencing or analysis errors. Thus, the mutational catalogue of a cancer genome $m = [m^1, m^2, \ldots m^K]^{\mathrm{T}}$, defined over the mutation alphabet Ξ with K letters, is a superposition of the signatures of the N operative mutational processes $P_i = [p_i^1, p_i^2, \ldots\ p_i^K]^{\mathrm{T}}, i = 1 \ldots N$, each defined over the mutation alphabet Ξ, with their respective exposures $e^i, i = 1 \ldots N$, and non-systematic noise n. In particular, the number of the j-the mutation type in m is:

$$m^j = \sum_{i=1}^{N} P_i^j e^i\ + n^j \qquad (2.5)$$

Note that in this definition, m, e, and n are vectors, while P is expressed as a matrix. Indeed, a set of signatures of N mutational processes, can be represented by a nonnegative matrix $P = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$ with size $K \times N$, where K is the number of mutation types and N is the number of signatures. The subscript index indicates the signature, while the superscript index corresponds to the mutation type.

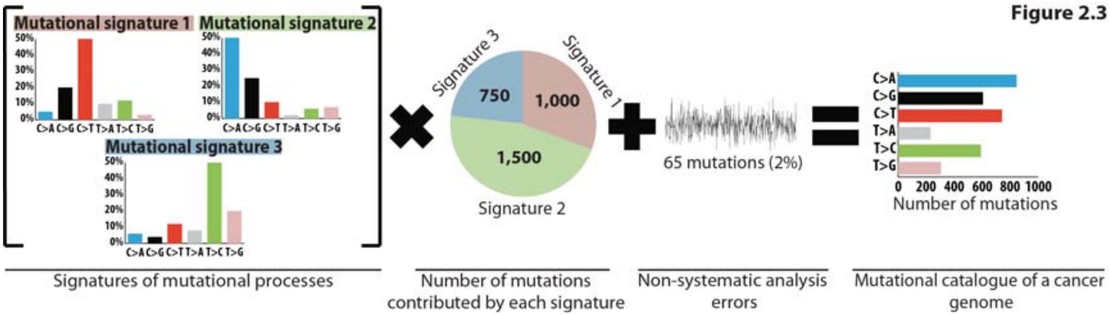


Figure 2.3: Simulated example of three mutational signatures active in a single cancer genome. The three mutational signatures were defined over the Ξ_6 alphabet. The mutational catalogue of the cancer genome is modelled as a linear superposition of the signatures of three processes and the respective number of mutations contributed by each signature, plus added non-

A simulated example illustrating this model is provided in Figure 2.3. Each of the signatures has a specific pattern over the six base substitutions. The first signature has a substantial proportion of C>T mutations and contributes, in total, 1,000 mutations to the cancer genome. The second process has a high proportion of C>A mutations while contributing 1,500 mutations. The third process generates substantial numbers of T>C mutations and contributes 750 mutations (Figure 2.3). The mutational catalogue of the cancer genome formed by these three processes, however, does not have any notable or specific features and does not obviously resemble any of the mutational signatures that are operative in it. This simulated mutational catalogue contains, in total, 3,315 mutations, 3,250 (~98%) contributed by the three mutational processes and the remaining 65 (~2%) by white noise corresponding to minor processes or experimental errors in generating the mutation catalogue of the genome.

2.3 Deciphering mutational signatures from a set of cancer genomes

In the previous section of this chapter, I described a mutational catalogue of a cancer genome as a linear combination of the signatures of the underlying mutational processes active in this cancer genome. A single mutational catalogue does not allow identification of the operative mutational signatures since there are many ways to decompose a single mutational catalogue into multiple mutational signatures. However, the availability of hundreds and even thousands of mutational catalogues of cancer genomes can address this limitation, as mutational signatures will have different exposures in different catalogues, constraining the number of solutions and thus allowing deconvolution of the signatures.

In summary, the approach developed here is used to identify the signatures of mutational processes from a large number of mutational catalogues. In order to do this, I will start by introducing matrix notations for mutational signatures, mutational catalogues, exposures of mutational signatures, and noise terms. These matrix notations are necessary to alleviate and shorten the description of the developed algorithm for deciphering mutational signatures.

2.3.1 Matrix notations for deciphering mutational signatures

The signature P_1 of a mutational process, defined over an alphabet Ξ with K letters, can be expressed as a nonnegative K-tuple, $P_1 = [p_1^1, p_1^2, \dots p_1^K]^\mathrm{T}$, where $\sum_{i=1}^{K} p_1^i = 1$ and p_1^i is the probability of the mutational processes P_1 to cause the mutation type corresponding to the i-th letter of the alphabet Ξ. As previously described, a set of N mutational signatures can be expressed as a nonnegative mutational signature matrix $P = \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$ with size $K \times N$, where K is the number of mutation types and N is the number of signatures. The subscript index indicates the signature, while the superscript index corresponds to the mutation type.

The mutational catalogue of a cancer genome, defined over the alphabet of mutation types Ξ, is represented by a morphism m, where $m: \Xi \to \mathbb{N}_0^K$. For a given genome, i, its mutational catalogue can be expressed as a nonnegative K-tuple, $m_i = [m_i^1, m_i^2, \dots m_i^K]^\mathrm{T}$. Hence, the mutational catalogues of G cancer genomes can be expressed as a nonnegative matrix of mutational catalogues $M = \begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix}$ of size $K \times G$. In this case, the mutational catalogues form the columns of the matrix, where K is the number of mutation types and G is the number of genomes. The subscript index indicates the mutational catalogue while the superscript index corresponds to the mutation type.

The exposure to a mutational process with a signature $P_1 = [p_1^1, p_1^2, \dots p_1^K]^{\mathrm{T}}$ is a scalar describing the number of mutations, $e^1 \in \mathbb{N}_0$, attributed to that signature in a given mutational catalogue. In this notation, the product $p_1^2 \times e_g^1$ is the number of mutations of type corresponding to the 2nd letter of alphabet Ξ caused by the mutational process P_1 in a cancer genome with number g. Hence, one can define a set of exposures of G genomes to a set of N processes as a nonnegative matrix $E = \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$ with size $N \times G$. Here, the subscript index indicates the genome while the superscript index corresponds to the signature.

In addition to the signatures of the operative mutational processes, the mutational catalogue of a cancer genome also reflects the effect of random error processes, which may occur due to the used experimental approach (*e.g.*, DNA sequencing) and/or bioinformatics methods (*e.g.*, algorithms for identifying somatic mutations from next-generation sequencing data). To reflect the existence of such errors, a random noise term is introduced in equation 2.5. This noise term n reflects an additive white Gaussian noise that occurs due to non-systematic errors. The noise term is specific to each mutational catalogue and it is defined over the alphabet Ξ of the mutational catalog, where $n: \Xi \rightarrow \mathbb{R}^K$. Hence, for a set of mutational catalogues of G cancer genomes, the noise term can be expressed as a matrix $N = \begin{bmatrix} n_1^1 & n_2^1 & \cdots & n_{G-1}^1 & n_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_1^K & n_2^K & \cdots & n_{G-1}^K & n_G^K \end{bmatrix}$ of real numbers with size $K \times G$. The subscript index indicates the noise term for the mutational catalogue while the superscript index corresponds to the mutation type. It should be noted that systematic sequencing and analysis errors are considered as "synthetic mutational processes" with specific profiles present in some (or all) genomes. A whole subsection in chapter 4 is devoted to examining such systematic sequencing and analysis errors across a large set of cancer genomes.

2.3.2 Defining the mutational signatures deciphering problem

The signatures of N different mutational processes and their respective exposures need to be extracted from the mutational catalogues of M cancer genomes (Figure 2.4). Using the introduced matrix notation, this could be expressed as:

$$\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix} =$$

$$\begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix} \times \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix} + \begin{bmatrix} n_1^1 & n_2^1 & \cdots & n_{G-1}^1 & n_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ n_1^K & n_2^K & \cdots & n_{G-1}^K & n_G^K \end{bmatrix} \quad (2.6)$$

or one can simplify equation 2.6 in a matrix form as:

$$M = P \times E + N \quad (2.7)$$

In practice, one knows only the mutational catalogues in the matrix M and the goal is to identify P and E such that these matrices best describe the original matrix M without over-fitting the data. Figure 2.4 provides a graphic representation of the problem for deciphering signatures of mutational processes from a set of mutational catalogues.

2.3.3 Examining the problem as a blind source separation

The examined problem can be considered as a specific case of the classic "cocktail party" problem, where multiple people attending a party are speaking simultaneously while several microphones placed at different locations are recording

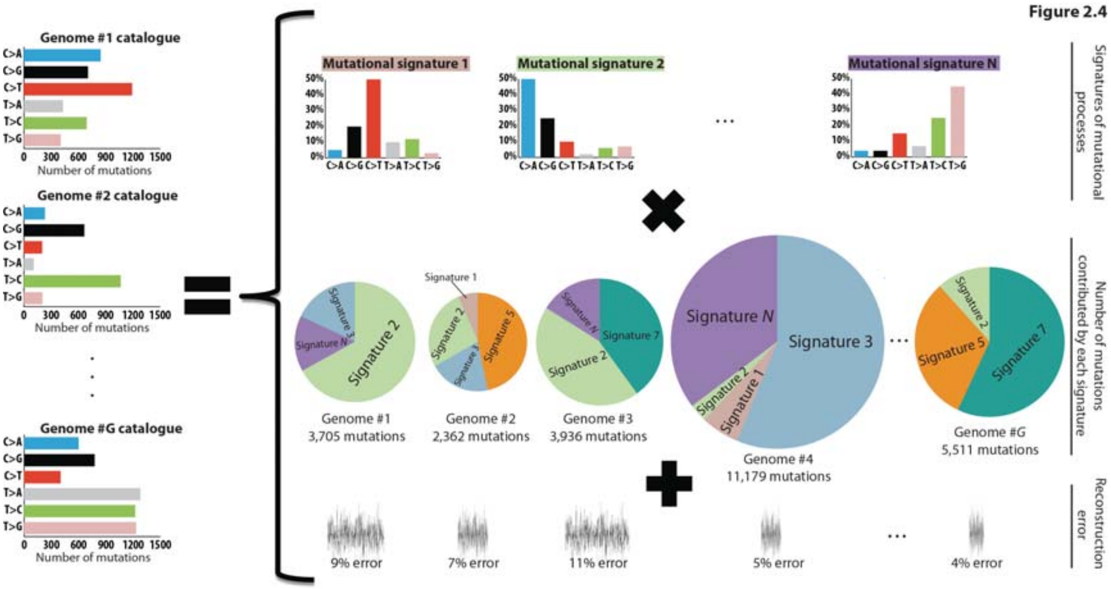


Figure 2.4: Simulated example of mutational signatures deciphered from a set of mutational catalogues. The mutational catalogues of *G* cancer genomes are used to decipher the signatures of *N* mutational processes as well as the number of mutations caused by each of the processes in each of the genomes. The extracted signatures and contributions do not allow an exact reconstruction of the original set, thus resulting in genome-specific reconstruction error.

the conversations. Each microphone captures a mixture of all sounds and the problem is to decipher the individual conversations from the recordings. This becomes possible because each microphone captures each conversation with a different intensity depending on the distance between the microphone and the conversation. Analogously, the provision of a catalogue of somatic mutations from a cancer genome only provides the final mixture of the signatures of all mutational processes operative in a cancer sample, and the goal is to decipher these signatures from a set of available mixtures (Figure 2.4). Thus, the mutational processes and their signatures are the "conversations," the exposure to a process is the "loudness of the conversation," the cancers themselves are the "microphones," and the final mutational catalogues are the "recordings."

The "cocktail party" problem is a type of blind source separation (BSS) problem that involves unscrambling latent (not observed) signals from a set of mixtures of these signals, without knowing anything about the mixing. To be able to "unmix" and reconstruct the original sources from the records, a BSS algorithm is needed for best possible extraction of the original signals from the mixtures. These BSS algorithms are capable of revealing hidden features and dependencies in large sets of observed data, and, based on these features, building a representation of the data that can contribute to understanding the biological mechanisms behind these data. The unmixing and reconstruction of the original signals is usually based on some constrained and/or regularized optimization procedure minimizing an objective (cost) function together with a few imposed constraints, such as: maximum variability, statistical independence, nonnegativity, smoothness, sparsity, simplicity, *etc*. The choice of the optimization constraints is usually based on *a priori* knowledge about the processed data, and hence the constraints could be different for every particular case.

The main difficulty in solving a BSS problem is that it is usually an under-determined (ill-posed) problem. There are two main/widely-used methods for resolving the under-determination of BSS: Independent Component Analysis (ICA) and Nonnegative Matrix factorization (NMF) (Comon, 2010; Roberts and Everson, 2001). Below, I briefly describe the basic principles of ICA and NMF.

ICA estimates the source and the mixing matrices by maximizing the statistical independence of the retrieved source signals (*i.e.*, the matrix columns are expected to be statistically independent). Typically, the source independence is achieved by maximizing some high-order statistics for each source signal, such as the kurtosis or negentropy (negative entropy). The main idea behind ICA is that while the probability distribution of a linear mixture of sources is expected to be close to a Gaussian (according to the Central Limit Theorem), the probability distribution of the original independent sources is expected to be non-Gaussian. As a result, ICA aims to maximize the non-Gaussian characteristics in the estimated sources with the goal of finding statistically independent non-Gaussian sources that reproduce the experimental data.

In contrast to ICA, NMF does not seek statistical independence or constrain any other statistical property. Thus, nonnegative matrix factorization allows the estimated sources to be partially or entirely correlated. Instead, NMF enforces a nonnegativity constraint on the original sources and their mixing components (*i.e.*, all the estimated matrix elements are greater than or equal to zero).

The differences between NMF and ICA have important implications for choosing one method over another. In general, ICA is used when one is looking for statistically independent signals. However, in practice, there are many cases where the ICA assumption of statistical independence contradicts the biological reality. For example, two distinct mutational processes may be reliant on the same components of the cellular machinery making them (at least partially) statistically dependent and, as such, these signals cannot be deciphered with an algorithm whose basis is to seek statistical independence. In contrast to ICA, NMF focuses purely on part-based decomposition (Lee and Seung, 1999). The part-based decomposition is particularly useful as it allows describing the original data only by additive signals that cannot cancel one another. This part-based decomposition results in “natural sparseness” of the underlying processes and it has been shown to extract meaningful components from complex datasets (Lee and Seung, 1999).

The nonnegative nature of the developed model in equation 2.7 requires a method that assumes (at the very least) nonnegativity of the original sources. The elegance, simplicity, and ability to extract meaningful processes make NMF the method of choice in this thesis. It should be noted that there are different algorithms that can be used for nonnegative matrix factorization. The results presented in this study are exclusively based on the multiplicative update algorithm (Lee and Seung, 1999).

2.3.4 Approach for deciphering signatures of mutational processes

For a given mutational catalogue M that contains G cancer genomes defined over an alphabet Ξ with K letters corresponding to mutation types (*i.e.*, M has a size $K \times G$), the algorithm extracts N mutational signatures defined over the same alphabet Ξ. The algorithm has the following steps:

STEP 1 (Dimension Reduction): Reduce the dimensions of the original matrix M by removing any mutation types that together account for $\leq 1\%$ of the mutations in all genomes, *i.e.* remove the maximum set of rows R in M for which:

$$\sum_{r \in R} \sum_{g=1}^{G} m_g^r \leq 0.01 \times \sum_{i=1}^{K} \sum_{j=1}^{G} m_j^i$$

and the cardinality of the set R, $|R|$, is maximized. The matrix M is transformed into a new matrix $\dot{M}$ with dimensions $\dot{K} \times G$, where $\dot{K} = K - |R|$.

STEP 2 (Bootstrap): Apply Monte Carlo bootstrap resampling to avoid over-fitting the extracted mutational signatures. The dimensionally reduced matrix $\dot{M}$ resulting in a new matrix $\breve{M}$, where the probability for getting a mutation of type corresponding to the q^{th} letter in the alphabet Ξ in a genome g is $Pr\left(\breve{m}_g^q\right) = \frac{\dot{m}_g^q}{\sum_{i=1}^{K} \dot{m}_g^i}$ while the total number of mutations in each genome g remains unaffected, *i.e.*, $\sum_{i=1}^{K} \breve{m}_g^i = \sum_{j=1}^{K} m_g^j$.

STEP 3 (NMF): Apply the multiplicative update algorithm (Lee and Seung, 1999) for nonnegative matrix factorization to the bootstrapped data by finding the solution to $\min_{P \in \mathbb{M}_{\mathbb{R}_+}^{(\dot{K},N)} \quad E \in \mathbb{M}_{\mathbb{R}_+}^{(N,G)}} ||\breve{M} - P \times E||_F^2$:

I. Initialize matrices P and E as random nonnegative matrices with respective sizes $\dot{K} \times N$ and $N \times G$, where N is the number of signatures.

II. Iterate until convergence, defined as 10,000 iterations without change, or until the maximum number of 1,000,000 iterations is reached:

$$e_G^N \;\leftarrow\; e_G^N\, \frac{[P^T\breve{\mathrm{M}}]_{N,G}}{[P^TPE]_{N,G}}$$

$$p_N^{\dot{K}} \leftarrow p_N^{\dot{K}}\, \frac{[\breve{M}E^T]_{\dot{K},N}}{[PEE^T]_{\dot{K},N}}$$

The notation $[AB]_{x,y}$ is equivalent to the $(x,y)^{-\mathrm{th}}$ element of the matrix C, where $C = A{\times}B$.

III. Store the identified signatures P and their respective exposures E.

STEP 4 (Iterate): Perform steps 2 and 3 for I iterations. I is determined by evaluating the convergence of the iteration-averaged signature matrix $\bar{P}$ (see below for deriving $\bar{P}$). I is selected in such a way that performing $2 * I$ iterations (*i.e.*, doubling the iterations) does not significantly change $\bar{P}$. In most cases between 400 and 500 iterations are needed, however, in some cases solutions could be found for $I \leq 100$ while in rare cases more than 1,000 iterations might be required. In general, the value of I is strongly dependent on the size and type of the initial matrix M.

STEP 5 (Cluster): The iterations performed in step 4 result in two sets of matrices, $S_P \in \mathbb{M}_{\mathbb{R}_+}^{(\dot{K},N)}$ and $S_E \in \mathbb{M}_{\mathbb{R}_+}^{(N,G)}$, that correspond respectively to the mutational signatures and their exposures generated over the I iterations. A partition-clustering algorithm is applied to the set of matrices S_P to cluster the data into N clusters. A variation of *k*-means (Jain, 2010), where each signature for $\forall P \in S_P$ is assigned to exactly one cluster, is used to partition the data. Similarities between mutational signatures are evaluated using a cosine similarity while the N centroids are calculated by averaging the signatures belonging to each cluster. The iteration-averaged matrix $\bar{P}$ is formed by combining the N centroid vectors ordered by their reproducibility (see Step 6). The error bars reported for each mutation type in each signature in $\bar{P}$ are calculated as the standard deviations of the corresponding mutation type in each centroid over the I iterations. Note that clustering the data in S_P effectively results in clustering S_E as each signature unambiguously corresponds to exactly one exposure, thus allowing derivation of $\bar{E}$.

STEP 6 (Evaluate): The reproducibility of the derived average signatures $\bar{P}$ is evaluated by examining the tightness and separation of the clusters used to form the centroids in $\bar{P}$ (see Step 5). More specifically, using cosine similarity, the average silhouette width for each of the N clusters is calculated. An average silhouette width of 1.00 is equivalent to consistently deciphering the same mutational signature, while a low silhouette width indicates a lack of reproducibility of the solution. The average silhouette width (Rousseeuw, 1987) of the N clusters is used as a measure of reproducibility for the whole solution. In addition to reproducibility, the average Frobenius reconstruction error is used to evaluate the accuracy with which the deciphered mutational signatures and their respective exposures describe the original matrix M, *i.e.*, $||M - \bar{P} \times \bar{E}||_F^2$, where a lower Frobenius reconstruction error corresponds to a better description of the original matrix. There is some association between the reproducibility of a solution and its reconstruction error. For example, solutions with very low reproducibility usually have high Frobenius reconstruction errors.

The developed framework for deciphering signatures of mutational processes relies on two input parameters, the original matrix M (size $K\times G$) and the number of mutational signatures N to be deciphered from M. However, in most cases, the value of N is unknown and needs to be determined from M. The model selection framework relies on applying the framework for deciphering signatures of mutational processes for values of N between 1 and $\min(K, G) - 1$. The reproducibility and average Frobenius reconstruction error is evaluated for each N, and the value of N is selected such that the extracted mutational signatures are reproducible and the reconstruction error is low.

2.3.5 Computational implementation of the algorithm

The framework for deciphering signatures of mutational processes — including its source code, brief documentation, and several examples of applying it to mutational catalogues — is freely available for download from:

http://www.mathworks.com/matlabcentral/fileexchange/38724

2.4 Evaluating the computational framework using simulated data

In the previous section of this chapter, I introduced a theoretical model of signatures of mutational processes operative in a cancer genome. Based on this model, I mathematically introduced the problem of deciphering mutational signatures from a set of mutational catalogues of cancer genomes. Further, I proposed an algorithm and developed a computational framework that allows to decipher these signatures. In this section, I focus on evaluating the developed approach with simulated data. The application of the approach to experimental data is performed in chapters 3 and 4.

2.4.1 Generating the simulation data

Signatures of mutational processes with different exposures are randomly generated and used to simulate mutational catalogues of cancer genomes. The simulated mutational catalogues are leveraged to assess the ability of the developed approach to decipher the mutational signatures with which the data are simulated. In most cases (unless specified otherwise in the text), the signatures of mutational processes are stochastically generated over the alphabet Ξ_{96} with similarities between them comparable to those previously observed in breast cancer genomes (Nik-Zainal et al., 2012). Similarly, unless specified otherwise, the exposures to mutational processes are uniformly distributed across the set of simulated cancer genomes while the total number of mutations in each mutational catalogue is drawn from a distribution comparable to the distribution of the total substitutions found in many human cancer genomes (Greenman et al., 2007; Nik-Zainal et al., 2012; Stratton, 2011; Wood et al., 2007). For every mutational process with signature $P_1 = [p_1^1, p_1^2, \dots p_1^K]^\mathrm{T}$, defined over an alphabet Ξ with K letters, contributing e_g^1 mutations in a cancer genome g, each mutation is assigned to one of the K mutation types according to the discrete probability density function of P_1. Poisson noise and additive white Gaussian noise are added to every simulated mutational catalogue. Lastly, each simulation scenario is repeated 100 times and the standard deviations of the results over these 100 repeats are reported as error bars in the respective figures.

2.4.2 Extracting mutational signatures from 100 simulated cancer genomes

An example of applying the developed theoretical approach to a set of 100 simulated mutational catalogues of cancer genomes is shown in Figure 2.5. Similar to many human cancer genomes (Greenman et al., 2007; Nik-Zainal et al., 2012; Stratton, 2011; Wood et al., 2007), every simulated genome contains between 500 and

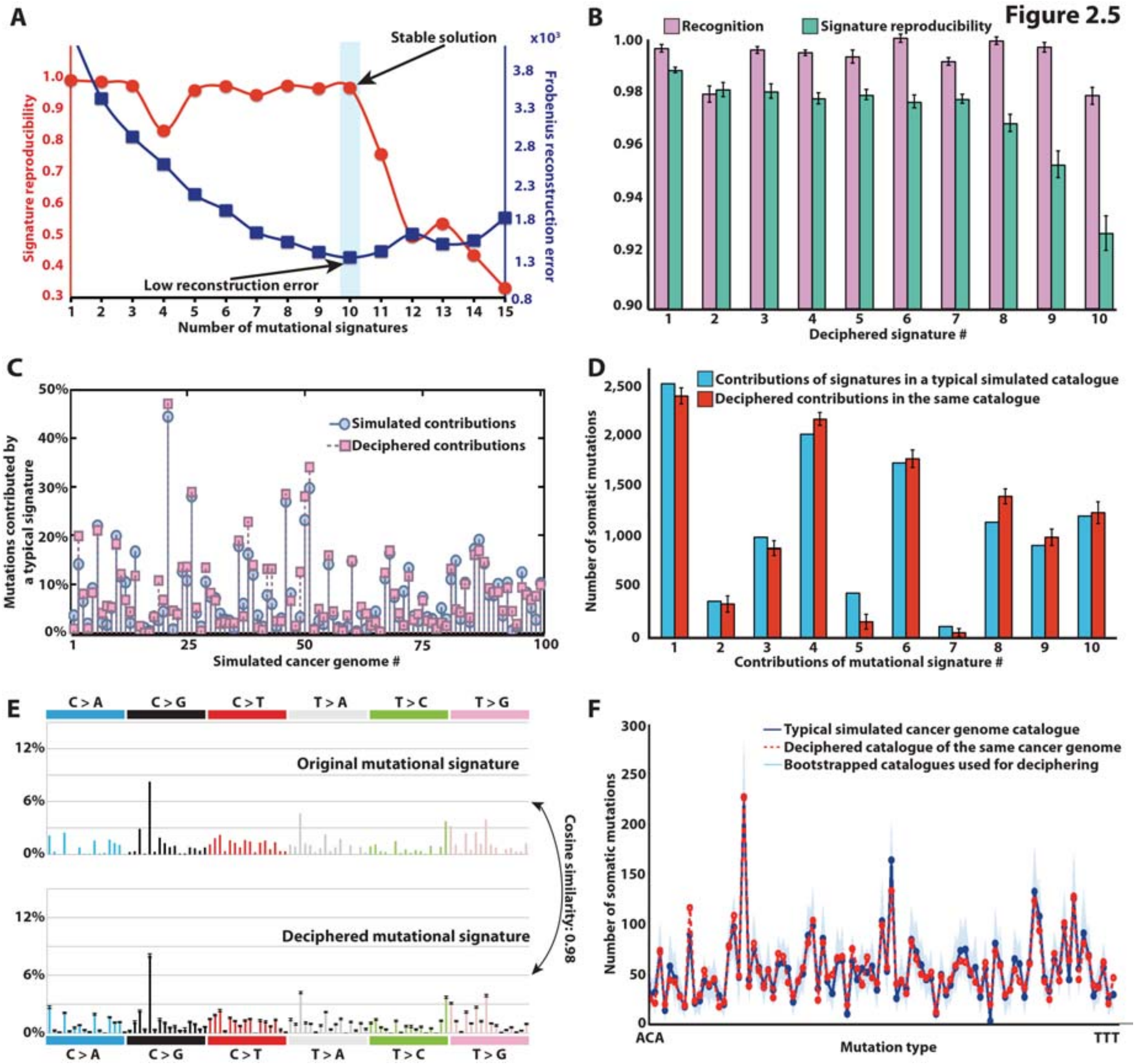


Figure 2.5: Deciphering mutational signatures from a set of 100 simulated mutational catalogues. ***(A)*** Identifying the number of processes operative in a set of 100 simulated cancer genomes based on reproducibility of their signatures and low error for reconstructing the original catalogues. ***(B)*** Comparison between the ten deciphered signatures and the ten signatures used to simulate the catalogues. Signature recognition, measured using cosine similarity, and signature reproducibility, measured using average silhouette width, is given for each mutational signature. ***(C)*** Comparison between deciphered and simulated contributions of one of the ten mutational processes in all cancer genomes. ***(D)*** Comparison between deciphered and simulated contributions of all signatures in a typical cancer genome. ***(E)*** Comparison between the profiles of typical deciphered and simulated signature. ***(F)*** Comparison between the mutational catalogues of a typical deciphered (red line) and simulated (dark blue line) cancer genome.

50,000 substitutions. The simulated mutations are generated using 10 mutational processes with distinct signatures each with 96 mutation types (*i.e.*, signatures are defined over Ξ_{96}).

Identifying the number, N, of mutational processes operative in a set of cancer genomes is required prior to deciphering their signatures. The developed model selection approach identifies N by applying the method for different values of N (see section 2.3.4). For every N, the similarity between the extracted processes (*i.e.*, process reproducibility) is evaluated from the stochastically initialized iterations. Further, for every N, the model selection approach assesses the average Frobenius reconstruction error of the averaged deciphered signatures $\bar{P}$ and their exposures $\bar{E}$, *i.e.*, $||\ M\ -\ \bar{P}\times\bar{E}||_F^2$. Low reconstruction error is indicative of an accurate description of the original cancer genome catalogues. N is selected such that the extracted processes are reproducible and the reconstruction error is low. Over-fitting the mutational signatures is avoided by bootstrapping the data (in each iteration) before applying NMF to them (see section 2.3.4).

For the 100 simulated cancer genomes, the approach is able to identify reproducible solutions for N between 2 and 10 (Figure 2.5A). Increasing the number of signatures from 2 to 10 substantially reduces the reconstruction error, but increasing beyond 10 does not further reduce it (Figure 2.5A). This indicates that the computational approach can optimally distinguish the signatures of 10 mutational processes, precisely the number originally used to simulate the mutational catalogues of these 100 cancer genomes. The 10 deciphered signatures are very reproducible (average silhouette width > 0.96) as well as extremely similar (average cosine similarity $>$ 0.98) to the ones used to generate the 100 mutational catalogues (Figure 2.5B). Further, the computational approach is able to accurately identify the number of mutations contributed by each of the 10 processes in each of the genomes. Comparison between original and deciphered exposures of one of the signatures in all genomes is shown in Figure 2.5C and a comparison of the contributions of all 10 signatures in a single genome is shown in Figure 2.5D. A typical comparison between an original and deciphered signature is shown in Figure 2.5E and a typical comparison between an original and reconstructed mutational catalogue of a genome is depicted in Figure 2.5F. In summary, the applied approach is able to accurately identify the underlying mutational signatures and their respective exposures in this set of 100 simulated mutational catalogues.

2.4.3 Identifying factors that influence extraction of mutational signatures

To identify factors that affect the ability to extract mutational signatures, signatures of mutational processes and their respective exposures are simulated under a number of different scenarios. The original signatures used to simulate the data are compared to the deciphered signatures in order to evaluate both the limitations and robustness of the developed computational framework. All comparisons between mutational signatures are done using a cosine similarity as previously described in section 2.2.2.

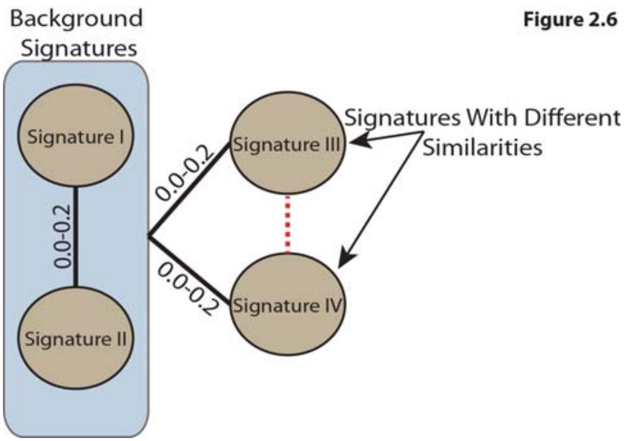


Figure 2.6: Design for simulating four mutational signatures with different similarities between them. Signatures I and II differ significantly from each other as well as from the other two Signatures (cosine similarity between 0.00 and 0.20). Signatures III and IV are simulated with varying similarities between them.

To evaluate how the degree of similarity between mutational signatures affects their extraction, sets of four randomly generated signatures are simulated; two of the signatures are very different from any of the other signatures, while the similarity of the remaining two to each other is varied (Figure 2.6). Hence, Signatures I and II are

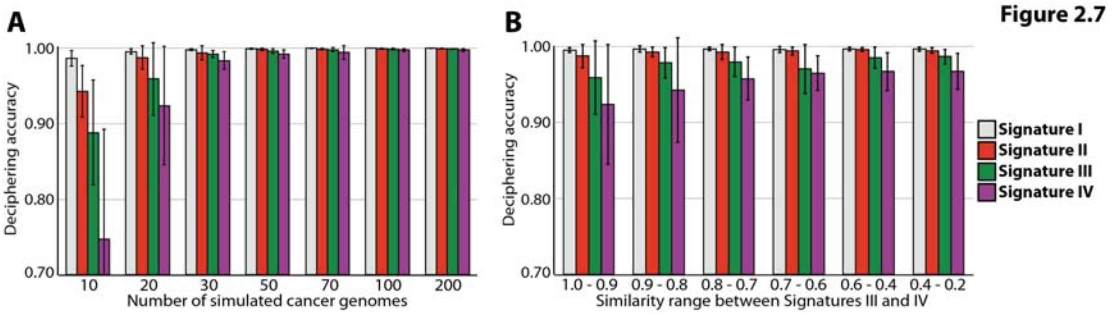


Figure 2.7: Deciphering mutational signatures with different similarities between them. ***(A)*** Different numbers of mutational catalogues are examined while Signatures III and IV are simulated with very similar profiles. ***(B)*** The mutational catalogues of 20 cancer genomes are simulated while the similarity between Signatures III and IV is varied.

simulated such as the cosine similarity between each of these signatures and any other signature is always within the range of 0.00 and 0.20 (*i.e.*, signatures with very different mutational profiles) while the similarity range between Signatures III and IV is varied, as described below.

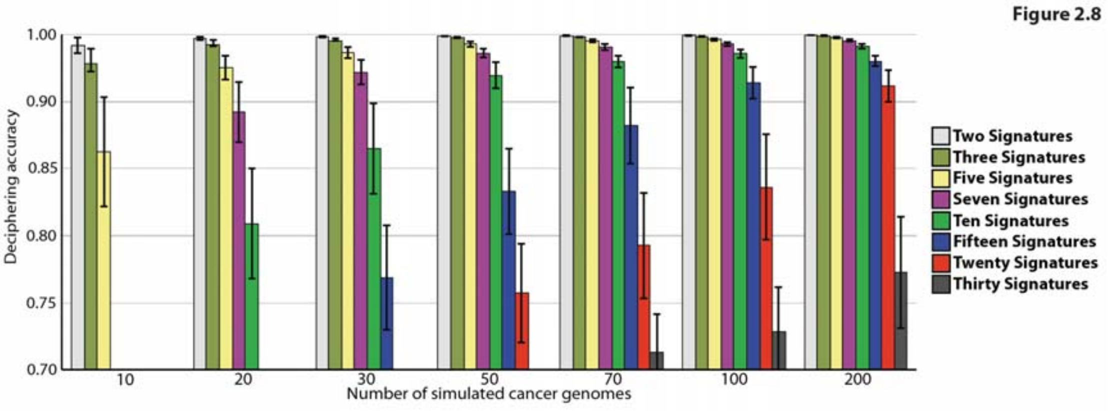


Figure 2.8: Deciphering mutational signatures from different sets of cancer genomes. Evaluating the effect of deciphering between two and thirty mutational signatures from sets of mutational catalogues derived from 10, 20, 30, 50, 70, 100, and 200 cancer genomes.

Sets of Signatures III and IV are simulated with a cosine similarity ranging between 0.90 and 1.00 (*i.e.*, signatures with extremely similar profiles). In addition, different numbers of mutational catalogues are examined (Figure 2.7A). The performed simulations indicate that 30 mutational catalogues are sufficient for adequately identifying the four mutational signatures, while 50 or more cancer genomes allow to perfectly decipher signatures that are extremely similar (Figure 2.7A). Further simulations are carried out in which sets of mutational catalogues of 20

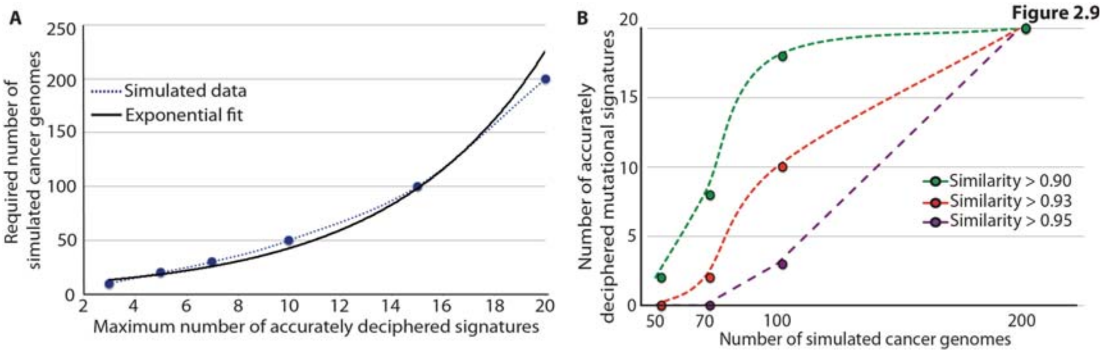


Figure 2.9: Dependencies between mutational signatures and mutational catalogues of cancer genomes. ***(A)*** Exponential dependency between accurately deciphered signatures (*i.e.*, cosine similarity between simulated and deciphered signature ≥ 0.95) and the number of mutational catalogs needed to decipher these signatures. ***(B)*** Identification of the maximum number of accurately deciphered signatures (cosine similarity between simulated and deciphered signature shown in the legend) from sets of mutational catalogues simulated using the signatures of 20 mutational processes.

cancer genomes are evaluated with a varied distance range between Signatures III and IV. Interestingly, even though 20 mutational catalogues are insufficient to decipher the profiles of very similar looking signatures, they are suitable for effectively extracting signatures that have similarities ≤ 0.70 (Figure 2.7B).

The number of available cancer genomes mathematically limits the number of signatures that can be extracted from the mutational catalogues of these genomes. For example, accurately deconvoluting signatures of 15 mutational processes from the mutational catalogues of only 10 cancer genomes is ineffective. To evaluate the effect of the number of mutational catalogues on extracting mutational signatures, simulations with different numbers of cancer genomes generated using a varying number of mutational signatures are performed. Between 10 and 200 sets of mutational catalogues are simulated using up to thirty mutational signatures (Figure 2.8). Interestingly, the number of mutational catalogues required to accurately decipher the signatures operative in them increases exponentially with the number of signatures (Figure 2.9A). Thus, while mutational catalogues from 100 cancer genomes are necessary to extract the signatures of fifteen mutational processes, at least 200 cancer genome catalogues are required to deconvolute twenty signatures (Figure 2.8). Nevertheless, it is possible to decipher at least some of the 20 mutational signatures from a set of 100 or fewer mutational catalogues (Figure 2.9B).

The number of somatic mutations in each cancer genome affects the ability to decipher the signatures of the operative mutational processes. In all previous

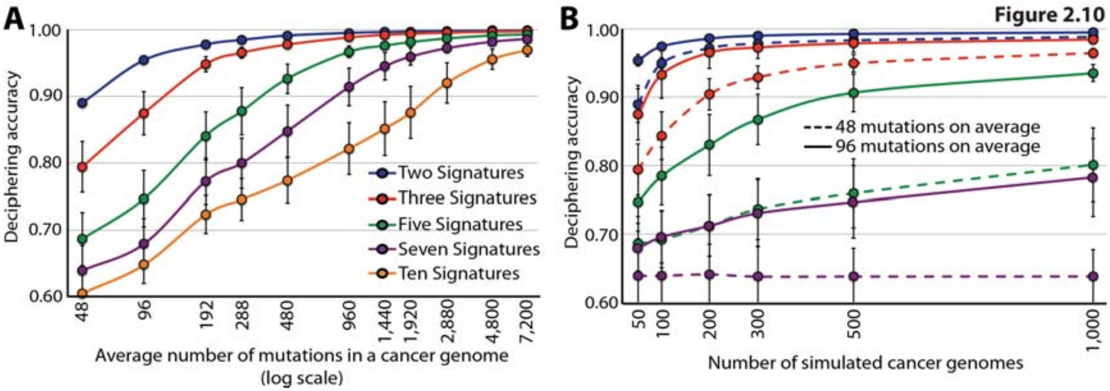
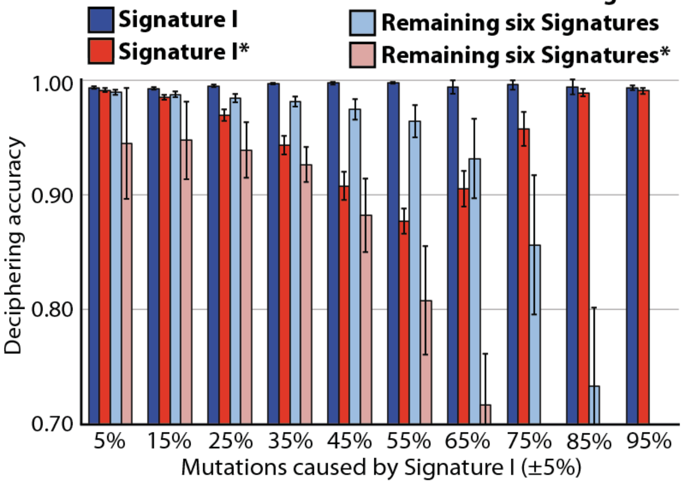


Figure 2.10: Dependencies between mutational signatures and numbers of somatic mutations. ***(A)*** Evaluating the effect of deciphering different number of mutational signatures from sets of mutational catalogues derived from 50 cancer genomes. The catalogues are simulated with different average number of mutations in a cancer genome. ***(B)*** Evaluating the effect of deciphering 2, 3, 5, or 7 mutational signatures from large sets of mutational catalogues containing small number of average mutations per cancer genome. The line colours correspond to the ones in the legend of panel A.

simulations, it is assumed that the distributions used to simulate the number of somatic mutations in cancer genomes are similar to those of some common cancers such as breast and prostate cancer. However, recent studies have demonstrated that there is substantial heterogeneity between the mutational burdens across major cancer types (Alexandrov et al., 2013a; Lawrence et al., 2013). In this section, simulations of 50 mutational catalogues with different average numbers of somatic mutations are performed. Each mutational catalogue is simulated using between two and ten mutational signatures. Obviously, having more somatic mutations (*i.e.*, more data for each sample) allows to better distinguish the profiles of the mutational signatures. As such, the focus of these simulations is to examine how lower average numbers of mutations (*i.e.*, between 48 and 7,200 mutations) affect the ability of the approach to identify mutational signatures. The results indicate that two or three signatures can be effectively extracted from catalogues with less than a hundred somatic mutations (Figure 2.10A). In contrast, extracting seven or more mutational signatures requires an average of at least 1,000 mutations per catalogue.

Figure 2.11



*Signature I's contributions are fixed in each of the cancer genomes

Figure 2.11: Deciphering mutational signatures with different contributions in mutational catalogues. Fifty mutational catalogues are simulated using mutational signatures with different contributions. Signature I's contributions are fixed to contribute a fixed percentage of all mutations in either the whole set of mutational catalogues (*i.e.*, the overall contribution is fixed but different genomes can have different contributions of Signature I; blue bars) or in each individual cancer genome (*i.e.*, Signature I's contributions are fixed in every single mutational catalogue; red bars).

The combined protein coding exons (the "exome") constitute only ~1% of the human genome. The analysis of exomes compared to whole-genome sequences is often perceived as advantageous because of lower costs and because a substantial proportion of cancer-causing driver somatic mutations may be found using this strategy. As a result, many more exome sequences of cancers have currently being generated than whole-genomes. To further evaluate the applicability of the approach to only parts of the genome (and more specifically exome sequences), large sets of mutational catalogues simulated with small average numbers of somatic mutations are examined. The results reveal that at least 500 mutational catalogues with an average of 96 mutations per catalogue (a total of ~50,000 mutations) are needed to decipher five mutational processes (Figure 2.10A), but these five mutational processes can be more easily deciphered from 50 cancer genomes containing an average of 480 mutations (a total of ~25,000 mutations, Figure 2.10B). This result indicates that it is more effective to decipher mutational signatures from a small number of catalogues containing many mutations than from many catalogues containing few mutations.

The strength of exposure of a mutational process in a set of genomes also influences the ability to decipher its signature. Two types of simulations of seven signatures operating with different strengths in 50 mutational catalogues are performed. In the first type, the percentage of exposures of Signature I in *all samples* is simulated as a constant parameter with values between 5% and 95% of all mutations (Figure 2.11). In contrast, in the second type of simulation, the exposures to Signature I are kept as a constant parameter in *every sample*, again, with values between 5% and 95% of all mutations (Figure 2.11). The results demonstrate that signatures contributing <5% of all mutations can be difficult to distinguish. Similarly, deciphering the members of a set of mutational signatures that have similar exposures with respect to each other over a set of cancer genomes is challenging (Figure 2.11). To overcome this problem, it may be advantageous to combine sets of mutational catalogues in which mutational processes are

Figure 2.12

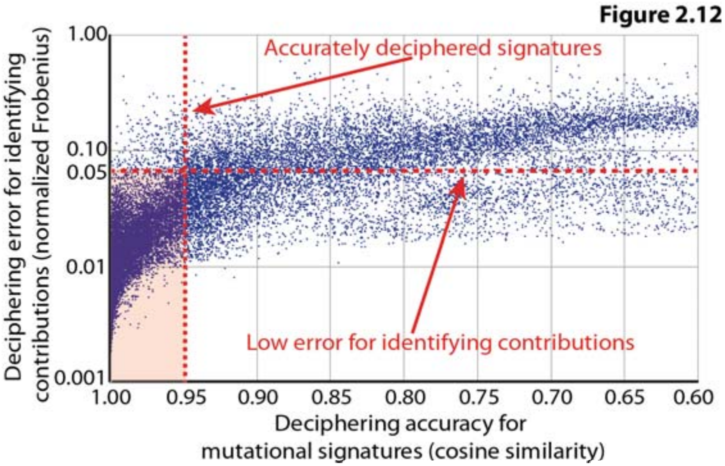


Figure 2.12: Deciphering errors of exposures and accuracy of mutational signatures. Comparison, across all previously performed simulations, between the accuracy of the deciphered mutational signatures and the deciphering error for identifying the contributions of these signatures. The deciphering Frobenius reconstruction error is calculated and averaged for each contribution and normalized based on the numbers of mutations in the respective mutational catalogue.

more likely to be active in different proportions (*e.g.*, from different cancer types). However, combining sets of mutational catalogues in this way ought to be considered with caution as the number of cancer genomes required for the extraction of signatures increases exponentially with the number of operative signatures and more cancer types may well entail more signatures (Figure 2.8 and Figure 2.9).

In addition to deciphering mutational signatures, the developed computational approach identifies the number of somatic mutations that each signature contributes to each mutational catalogue. In general, one would expect that the developed algorithm is, at least to some degree, symmetrical. Thus, when the algorithm correctly identifies the mutational signatures, it should also accurately estimate the contributions of these signatures (see section 2.3.4 in regards to the symmetric clustering of the data extracted in the sets of signatures, S_P, and the sets of exposures, S_E). Evaluating the average deciphering error for identifying contributions, for all previously performed simulations, confirms that the majority of accurately deciphered mutational signatures (*i.e.*, cosine similarity between simulated and extracted signatures ≥ 0.95) are associated with a low error (*i.e.*, normalized Frobenius error rate $\leq 5\%$) for their respective signature contributions (Figure 2.12). Further examination reveals that only very few of the accurately extracted signatures are associated with a normalized Frobenius error rate $\geq 5\%$ (Figure 2.13A). Interestingly, the analysis indicates that the contributions of signatures generating large numbers of mutations (> 200) are generally associated with lower error rates (Figure 2.13B).

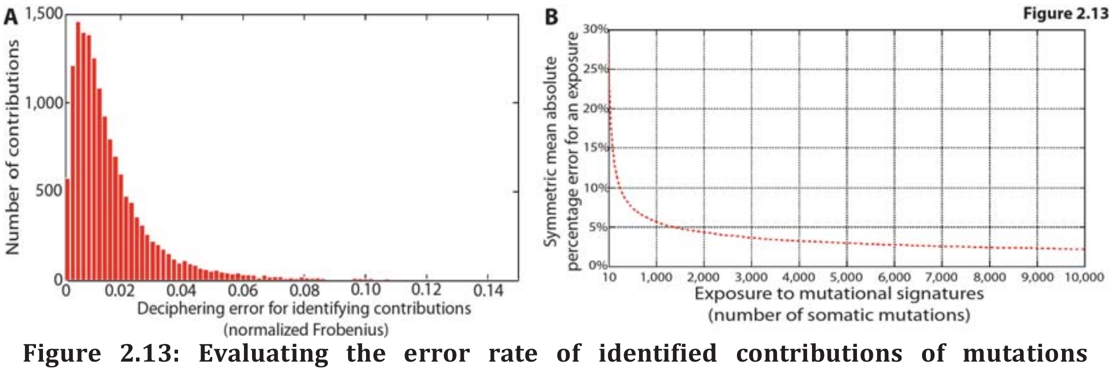


Figure 2.13: Evaluating the error rate of identified contributions of mutations signatures. ***(A)*** Distribution of the normalized Frobenius error for identifying the contributions of accurately deciphered signatures of mutational processes (*i.e.*, cosine similarity between simulated and deciphered signature ≥ 0.95). ***(B)*** Average symmetric mean absolute percentage error for identifying the contributions of accurately deciphered signatures of mutational processes (*i.e.*, cosine similarity between simulated and deciphered signature ≥ 0.95) based on the number mutations contributed by the signature.

2.5 Discussion

In this chapter, I have modelled the signatures of somatic mutational processes in cancer genomes as a blind source separation problem and introduced a computational framework that extracts these mutational signatures from the mutational catalogues obtained from cancer genome sequences. To identify these signatures, the intrinsic nonnegativity of mutations mandates employment of a method incorporating a nonnegative constraint. The extensive evaluations of the approach with simulated data demonstrate that the developed algorithm is effective in deciphering mutational signatures from mutational catalogues.

The efficiency of the algorithm could be further improved by incorporating additional constraints. For example, the current implementation of the computational framework relies on nonnegative matrix factorization, which has a natural weak sparsity constraint; however, a strong sparsity constraint could be applied to the exposure matrix E. This would guarantee that the mutational catalogue of a cancer genome is described by a minimum number of processes. Algorithms implementing this and other constraints have been previously developed (Berry et al., 2007; Gao and Church, 2005; Peharz and Pernkopf, 2012; Zheng et al., 2006) and could be applied to cancer genomics data. Nonetheless, this study demonstrates that an approach based on the simplest (*i.e.*, without additional constraints) NMF algorithm is sufficient to decipher both the signatures of the mutational processes operative in a set of cancer genomes as well as the number of mutations each signature contributes to the mutational catalogue of each cancer genome.

Parameters to which solutions are sensitive include the number of operative mutational processes, the strength of their exposures, the degree of difference between mutational signatures, the number of analysed cancer genomes, the number of mutations per cancer genome, and the number of mutation types that are incorporated into the model (Figures 2.6 through 2.13). These factors will determine the manner in which the method will be applied in the next chapters of this thesis. Importantly, the results show that, despite relatively few mutations present in each case, the approach can be applied to exome data, extracting at least some of the signatures of the operative processes.

It should be noted that when the number of samples in a dataset is too low or when the mutational burden is insufficient, the developed approach will lack the power to decipher the signatures of all operative mutational processes. Thus, in some cases, the extracted signatures will represent mixtures of multiple independent patterns of mutations and only additional samples will allow further differentiating these mutational signatures.

Diverse mutation classes can be included in this type of analysis. Thus the application of the developed approach can, if desired, be limited to single base substitutions or be widened to include double nucleotide substitutions, insertions, deletions, geographically localized forms of mutation and mutation features such as transcriptional strand-bias. Following this principle, rearrangements and copy number changes (and potentially even epigenetic modifications) could be incorporated in order to derive a comprehensive overview of operative mutational processes.

The complexity of the mutational processes operative in some cancers and the inherent challenges in extracting their attendant mutational signatures should not be underestimated. For example, tobacco smoke contains around 7,000 chemicals from which over 60 are known to be mutagenic (Rodgman and Perfetti, 2008). Thus, the mutational pattern of a lung cancer in a tobacco smoker will reflect the activity and potency of (at least) several of these chemicals. Each of these chemicals may have its unique mutational signature. A group of smokers loyal to the same brand will be simultaneously exposed to the same combination of mutagens. Analysis of tumours from this group of individuals therefore may not allow the mutagens to be distinguished from one another and the developed computational approach will extract only a single signature that encompasses the combined mutational activity of the most mutagenically potent chemicals. However, as different cigarette brands may contain different combinations and amounts of mutagens, analysis of mutational catalogues from cancers due to different tobacco brands could allow differentiation between the signatures of each of the different chemicals. An ambitious aspiration of this nature would, however, probably only be feasible with data from thousands of cases, coupled to the statistical power and resolution provided by whole-genome mutational catalogues.