Bioinformatics, 33(1), 2017, 8–16
doi: 10.1093/bioinformatics/btw572
Advance Access Publication Date: 1 September 2016
Original Paper

OXFORD

Genome analysis

signeR: an empirical Bayesian approach to mutational signature discovery

Rafael A. Rosales[1,†], Rodrigo D. Drummond[2,†], Renan Valieris[2], Emmanuel Dias-Neto[3,4] and Israel T. da Silva[2,5,*]

[1]Departamento de Computação e Matemática, Universidade de São Paulo, Ribeirão Preto, SP 14040-901, Brazil, [2]Laboratory of Bioinformatics and Computational Biology, A. C. Camargo Cancer Center, São Paulo, SP 01509-010, Brazil, [3]Laboratory of Medical Genomics, A. C. Camargo Cancer Center, São Paulo, SP 01509-010, Brazil, [4]Laboratory of Neurosciences (LIM27), Department and Institute of Psychiatry, Faculty of Medicine, University of São Paulo, São Paulo, SP 05403-903, Brazil and [5]Laboratory of Molecular Immunology, The Rockefeller University, New York, NY 10065, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.
Associate Editor: Alfonso Valencia

Received on June 22, 2016; revised on August 11, 2016; accepted on August 26, 2016

Abstract

Motivation: Mutational signatures can be used to understand cancer origins and provide a unique opportunity to group tumor types that share the same origins and result from similar processes. These signatures have been identified from high throughput sequencing data generated from cancer genomes by using non-negative matrix factorisation (NMF) techniques. Current methods based on optimization techniques are strongly sensitive to initial conditions due to high dimensionality and nonconvexity of the NMF paradigm. In this context, an important question consists in the determination of the actual number of signatures that best represent the data. The extraction of mutational signatures from high-throughput data still remains a daunting task.

Results: Here we present a new method for the statistical estimation of mutational signatures based on an empirical Bayesian treatment of the NMF model. While requiring minimal intervention from the user, our method addresses the determination of the number of signatures directly as a model selection problem. In addition, we introduce two new concepts of significant clinical relevance for evaluating the mutational profile. The advantages brought by our approach are shown by the analysis of real and synthetic data. The later is used to compare our approach against two alternative methods mostly used in the literature and with the same NMF parametrization as the one considered here. Our approach is robust to initial conditions and more accurate than competing alternatives. It also estimates the correct number of signatures even when other methods fail. Results on real data agree well with current knowledge.

Availability and Implementation: `signeR` is implemented in R and C++, and is available as a R package at http://bioconductor.org/packages/signeR.

Contact: itojal@cipe.accamargo.org.br

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is a collection of diverse pathological entities that harbour and probably derive from a complex collection of genomic alterations. Today, it is widely accepted that the accumulation of these alterations, including somatic mutations, is one of the major causes of the malignant transformation triggering the expansion of tumour cell clones. As tumors evolve, these mutations are found across many genomic loci, but also tend to preferentially affect certain pathways (Ciriello *et al.*, 2012). The diversity and complexity of somatic mutational processes in these clones is a conspicuous feature orchestrated by DNA damaging agents and repair processes, including the exposure to exogenous or endogenous carcinogenic/mutagenic agents, retro-insertion of endogenous retroviruses, defects in DNA mismatch repair enzymes and enzymatic modifications of the DNA among others (Roberts and Gordenin, 2014). The actual identification of the underlying mutational processes is central to an understanding of cancer origin and evolution (Alexandrov, 2013; Alexandrov and Stratton, 2014; Helleday *et al.*, 2014; Roberts and Gordenin, 2014).

Most somatic mutations comprise single base substitutions, insertions and deletions, rearrangements and copy number variations (CNV). Single base substitutions fall into one of six possible base changes, namely `C:G>A:T`, `C:G>G:C`, `C:G>T:A`, `T:A>A:T`, `T:A>C:G` and `T:A>G:C`. According to Alexandrov *et al.* (2013), this set may be further enlarged by including the $5'$ and $3'$ neighbouring bases of each substitution site, leading to an alphabet $\mathcal{A}$ with 96 trinucleotide mutation types. More generally, the definition of $\mathcal{A}$ could in principle accommodate mutations of various other kinds such as indels, rearrangements, copy number changes and even wider neighbouring contexts. Once $\mathcal{A}$ is properly defined, the counts for the mutations found in G different genomes are assembled into a $K \times G$ matrix M with $K = |\mathcal{A}|$. A crucial assumption consists in viewing the counts in M as the additive effect of N mutational processes, each defined as a $K \times 1$ vector of mutational rates. The later defines what is known as a mutational signature. More precisely, the mutations across all genomes result as the linear combination of N basis vectors of dimension $K \times 1$, with mixture coefficients defined by N exposure vectors of dimension $1 \times G$. If the basis vectors are merged into a $K \times N$ matrix of signatures P, and the coefficient vectors into a $N \times G$ matrix of exposures E, then the data can be simply factored as $M = PE$. An example of this is shown in Figure 1.

For any given mutation-count matrix there are essentially two interrelated questions that should be addressed: 1. the determination of the underlying signatures and exposures to best account for the observations, and 2. the determination of the actual number of signatures N. Nik-Zainal *et al.* (2012) and Alexandrov *et al.* (2013) addressed the first issue by using nonnegative matrix factorization (NMF) techniques. NMF as conceived by Lee and Seung (2001) finds the factors P, E that approximately solve the following nonconvex optimization problem

$$\min_{P \geq 0, E \geq 0} ||M - PE||, \tag{1}$$

for a given fixed rank N and an appropriately chosen norm. To deal with the second question, Nik-Zainal *et al.* (2012) and Alexandrov *et al.* (2013) perform the factorization of the same data for various ranks, namely for $1 \leq N \leq \min\{K, G\} - 1$. The rank is then determined rather indirectly by studying the clustering properties of the obtained factors via a criterion developed by Brunet *et al.* (2004) or by using the residual sum of squares, Hutchins *et al.* (2008).

An alternative approach to mutational signature discovery, and to NMF in general, follows from a statistical interpretation of the problem posed by (1) in which M is assumed to be a random matrix distributed according to a member of the exponential family parameterised by P and E. The optimization problem posed by (1), under the norm induced by a specific Bregman divergence (see Banerjee *et al.*, 2005), turns out to be equivalent to the maximum likelihood estimation of P and E. For instance, if M is Poisson distributed with rate PE, then the likelihood maximization with respect to P and E is equivalent to the minimization of (1) under the norm defined by the Kullback-Leibler divergence. The maximization of a Gaussian likelihood is equivalent to the minimization under the Frobenius norm. A key aspect of this perspective is that it allows to treat the determination of the factorization rank N as a model selection problem. The statistical interpretation was developed by Cemgil (2009), Févotte and Cemgil (2009) and Schmidt *et al.* (2009) in the general NMF context and then considered by Fischer *et al.* (2013) for the mutational signature application. Fischer *et al.* (2013) modelled M as Poisson distributed and then considered the estimation of P and E by using an expectation maximisation (EM) algorithm. The number of mutational signatures was estimated by considering an (unnecessary) saddlepoint approximation to the Bayesian information criterion (BIC). Recently, Shiraishi *et al.* (2015) and Rosenthal *et al.* (2016) also considered a statistical approach to the determination of mutational signatures. The former, however, considers a different NMF parametrization in which the features composing each mutation type in $\mathcal{A}$ are assumed independent. The latter considers the simpler problem of the estimation of E for given M, P and N.

In this article we consider an empirical Bayesian treatment to the NMF model as initially described by Alexandrov *et al.* (2013) for

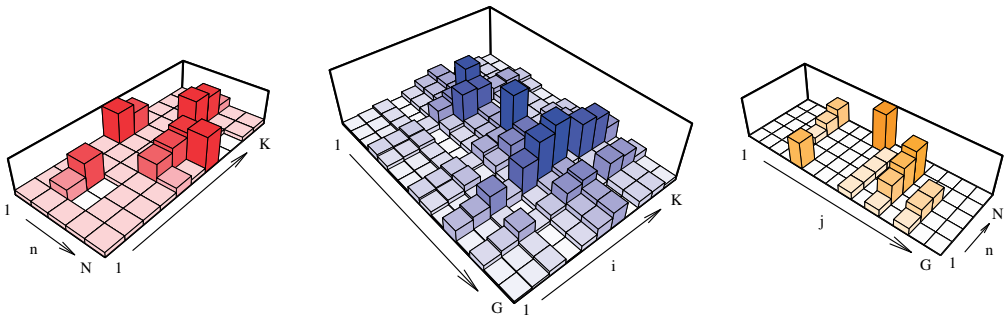

Fig. 1. A factorization for a mutation counts matrix M. The mutation matrix shown at the centre is defined over an alphabet with $K = 11$ symbols, $1 \leq i \leq 11$, and $G = 15$ genomes, $1 \leq j \leq 15$. The matrices at the left and the right represent respectively a signature and an exposure matrix P and E, obtained for a factorization with rank $N = 5$

the estimation of mutational signatures. Following Fischer *et al.* (2013), our model also incorporates the genome frequencies of the triplets where each mutation type in $\mathcal{A}$ can occur, which are known as opportunities. These enter the model as a matrix of weights W, leading to observations generated at rate $PE \circ W$ with $\circ$ as Hadamard's element wise matrix product. Both, the effectiveness and the advantages of our method are shown by using real and synthetic datasets.

2 Approach

2.1 Hierarchical model

2.1.1 Likelihood and latent variables

Let $p_{in} = (P)_{in}$ be the i, n-entry of P and likewise let $e_{nj} = (E)_{nj}$ and $w_{ij} = (W)_{ij}$. The random variables M_{ij} are assumed to be independent and Poisson distributed with rates $(PE \circ W)_{ij} = w_{ij} \sum_{n=1}^{N} p_{in} e_{nj}$. For a given sample of M, say m, this formulation is sufficient to define the likelihood function $\mathcal{L}(\theta, W; m)$ if one identifies the matrices P, E as model parameters θ, for $\theta \in \Theta = \mathbb{R}_{+}^{K \times N} \times \mathbb{R}_{+}^{N \times G}$. The opportunities are either known or set to $W = \mathbf{1}$, and hence regarded as fixed parameters. To simplify notation we omit any further reference to W. An expression for $\mathcal{L}(\theta; m)$ is presented as supplementary material by the equation (s_1).

This relatively simple model allows for a latent variable representation in which the observed counts are expressed as the sum of $N \geq 1$ independent Poisson variables

$$M_{ij} = Z_{i1j} + Z_{i2j} + \cdots + Z_{iNj}, \tag{2}$$

each with rate respectively equal to $p_{in} e_{nj} w_{ij}$. This description is an immediate consequence of the properties of sums of independent Poisson random variables. Biologically, this accounts for the observation that the total number of mutations of a specific type, say (i, j), arise as the linear combination of N mutational processes Z_{inj}, $n = 1, \ldots, N$. From a statistical perspective, (2) enables a data augmentation scheme that becomes instrumental for a Bayesian treatment to NMF. As observed by Cemgil (2009), this allows the implementation of several powerful techniques such as the Expectation Maximisation (EM) algorithm, Markov chain Monte Carlo (MCMC) and variational Bayesian approximations. Our approach to NMF fully exploits the data augmentation scheme defined by (2). Hereafter we denote by Z the random tensor $\{Z_{inj} : 1 \leq i \leq K, 1 \leq n \leq N, 1 \leq j \leq G\}$ and then let z be a generic value for Z.

2.1.2 Priors and hyperpriors

We consider conjugate priors for the matrices P and E by modelling each of their entries as being independent Gamma distributed random variables. Specifically, p_{in} is Gamma distributed with shape $\alpha_{in}^{p} + 1$ and rate $\beta_{in}^{p} \geq 0$ for $\alpha_{in}^{p} \geq 0$. Likewise, e_{nj} are Gamma with shape α_{nj}^{e} $+1$ and rate $\beta_{nj}^{e} \geq 0$ for $\alpha_{nj}^{e} \geq 0$. Shape parameters are shifted by 1 to ensure bounded values for the Gamma densities, improving stability of the computational methods described in subsequent sections. Let A_p and B_p be $K \times N$ matrices respectively with entries α_{in}^{p} and β_{in}^{p} and A_e, B_e be $N \times G$ matrices with elements α_{nj}^{e} and β_{nj}^{e}, and then let $\psi = (A_p, B_p, A_e, B_e)$ denote the hyperparameters.

A further hierarchy in our model is set by considering the distributions for the hyperparameters ψ. By conjugancy to the prior, we define the entries of B_p as being independent and distributed according to a common Gamma distribution with shape and rate $a_p > 0$, $b_p > 0$. Similarly, the elements of B_e are Gamma distributed with shape $a_e > 0$ and rate $b_e > 0$. The situation for the matrices A_p and A_e is however different. While a Gamma distribution for the entries of A_p and A_e is conjugate to the Gamma prior (see Miller, 1980), the resulting full conditional distribution necessary to draw inferences about A_p and A_e does not have a standard form. This fact has long been recognized in the Poisson hierarchical model (George *et al.*, 1993) and may be dealt with by choosing any parametric family of distributions with the appropriate support. Here we take the elements of A_p and A_e as independent and exponentially distributed with rates $\lambda_p > 0$ and $\lambda_e > 0$. Let η be the vector of hyperprior parameters $(a_e, b_e, a_p, b_p, \lambda_p, \lambda_e)$ defined on $\Lambda = (\mathbb{R}^{+} \backslash 0)^6$.

2.2 Bayesian treatment

We follow an empirical Bayesian approach in which the parameters θ, the hyperparameters ψ and the hyperprior parameters η are all estimated from the data. Inferences about the mutational signatures and their exposures are driven by the posterior distribution for the NMF model by combining MCMC and EM techniques as encouraged by Casella (2001). Specifically, for a given value of η we consider a Metropolized Gibbs sampler targeted towards the conditional posterior $\pi(\theta | M, \eta)$. This entails the iterative generation of a sequence of samples $(Z^{(r)}, \theta^{(r)}, \psi^{(r)})$, $r \geq 1$, from the set of full conditional distributions

$$Z^{(r+1)} \sim \pi(Z | \theta^{(r)}, \psi^{(r)}, M, \eta), \quad \theta^{(r+1)} \sim \pi(\theta | Z^{(r+1)}, \psi^{(r)}, M, \eta),$$
$$\text{and} \quad \psi^{(r+1)} \sim \pi(\psi | Z^{(r+1)}, \theta^{(r+1)}, M, \eta).$$

These samples are used to update the value of η via a stochastic EM step and the later is then used to draw a subsequent sequence of samples for Z, θ and ψ. The successive iteration of these steps defines a convergent sequence $(\eta^{(u)})$, $u \geq 1$, allowing the estimate $\hat{\eta} = \eta^{(U)}$ for sufficiently large U. A final set of MCMC samples for Z, θ and ψ drawn by conditioning on $\hat{\eta}$ provide the following estimate of the required posterior $\pi(\theta | M, \eta)$.

$$\hat{\pi}(\theta | M, \hat{\eta}) = \frac{1}{R} \sum_{r=1}^{R} \pi(\theta | Z^{(r)}, \psi^{(r)}, M, \eta^{(U)}) \tag{3}$$

The MCMC samples are used to compute estimates and all other related posterior statistics for the signatures and their exposures. The sampled matrices are first rescaled such that the columns of P sum to 1 and the product PE is left unaltered. The full set of rescaled samples is used to exploit the posterior distribution of those parameters, allowing to define two novel applications described in Sections 2.4 and 2.5. Point estimates for the signatures and their exposures are considered for visualization and comparison purposes. These are computed as the sample median and denoted hereafter by $\widehat{P}$ and $\widehat{E}$ respectively.

Details about the implementation of Gibbs sampler are relatively standard and are included in Section 2 of the supplementary material. The following section presents and justifies the MCMC EM approach.

2.2.1 MCMC EM

For a given data sample m and $Z = z$, direct use of Bayes theorem allows to express the marginal likelihood for η, i.e. the function $\mathcal{L} : \Lambda \to R$ induced by $\eta \mapsto p(M = m | \eta)$, as

$$\mathcal{L}(\eta; m) = \frac{\mathcal{L}(\eta; m, z, \theta, \psi)}{\pi(z, \theta, \psi | m, \eta)}. \tag{4}$$

Here $\mathcal{L}(\eta; m, z, \theta, \psi)$ is defined as being equal to $p(M = m, Z = z, \theta, \psi | \eta)$ but considered as a function of η. This quantity can be evaluated by observing the conditional decomposition

$$\mathcal{L}(\eta; m, z, \theta, \psi) = p(M = m, Z = z|\theta)p(\theta|\psi)p(\psi|\eta),$$

where $p(\theta|\psi)$ and $p(\psi|\eta)$ stand respectively for the prior and the hyperprior distributions, and $p(M = m, Z = z|\theta)$ is the complete data likelihood. An expression for the later is included as supplementary material by equation (s_2). Taking logarithms and integrating with respect to the posterior distribution $\pi(Z, \theta, \psi|m, \eta)$ with $\eta = \eta_0$ at both sides of (4) gives

$$\mathbb{E}[\ln \mathcal{L}(\eta; m)|\eta_0] = \mathbb{E}[\ln \mathcal{L}(\eta; m, Z, \theta, \psi)|\eta_0] - \mathbb{E}[\ln \pi(Z, \theta, \psi \mid m, \eta)|\eta_0].$$

This expression is the basic identity on which the EM algorithm is built and justifies therefore the convergence of the sequence

$$\eta^{(u+1)} = \arg\max_{\eta \in \Lambda} \mathbb{E}[\ln \mathcal{L}(\eta; m, Z, \theta, \psi)|\eta^{(u)}], \quad u \geq 0, \qquad (5)$$

towards the maximum likelihood estimate of η for any $\eta^{(0)} = \eta_0 \in \Lambda$. The integral involved in the above expectation cannot be computed directly but it may be estimated via Monte Carlo, leading to the sequence

$$\widehat{\eta}^{(u+1)} = \arg\max_{\widehat{\eta}^{(u)} \in \Lambda} \left\{ \frac{1}{R} \sum_{r=1}^{R} \ln \mathcal{L}(\widehat{\eta}^{(u)}; m, Z^{(r)}, \theta^{(r)}, \psi^{(r)}) \right\}. \qquad (6)$$

The maximization steps involved in (6) are relatively simple to implement and further detailed in Section 3.1 of the supplementary material.

The procedure described is valid because the sampler developed throughout generates $(Z^{(r)}, \theta^{(r)}, \psi^{(r)})$ approximately from the posterior distribution that is actually used to define the expectation in (5), see for instance Fort and Moulines (2003). This raises however the issue as to in what sense $\widehat{\pi}(\theta|M, \widehat{\eta})$ defined by (3) can be regarded as an estimate for $\pi(\theta|M, \eta)$. The answer to this is provided by the following result. Let $f(\theta|M, \eta)$ be the density of the posterior distribution $\pi(\theta|M, \eta)$.

THEOREM 1. For any measurable set $B \subseteq \Theta$, $\widehat{\pi}(B|M, \widehat{\eta})$ converges in total variation towards $\pi(B|M, \eta)$ as $R, U \to \infty$, that is

$$\lim_{U,R \to \infty} \sup_{B} \left| \int_B \left[\widehat{\pi}(\theta|M, \widehat{\eta}) - f(\theta|M, \eta) \right] d\theta \right| = 0.$$

The proof to this is included in Section 3.2 of the accompanying supplementary material.

The implementation of the MCMC EM used to estimate η and generate the samples for (Z, θ, ψ) conditionally on η for the rank N is shown in Algorithm 1. Model parameters θ are initialized by sampling from the prior or by considering the optimization approach to (1), implemented in R via the NMF package by Gaujoux and Seoighe (2010). The constants R_0 and R_2 are set to 1000 and R_1 and U are set to 100, but all may also be changed by the user.

Algorithm 1 MCMC EM

1: **Input** : M, W, N, `start`
2: **Initialize** η, ψ, θ, Z :
3: $u \leftarrow 0$; $\epsilon \leftarrow 1$
4: $\eta^{(0)} \leftarrow (1, \ldots, 1)$
5: $\psi^{(0)} \leftarrow \psi^{prior} \sim p(\psi|\eta^{(0)})$
6: **if** `start` = "`NMF`" **then**
7: $\theta^{(0)} \leftarrow \text{NMF}(M/W)$
8: else
9: $\theta^{(0)} \leftarrow \theta^{prior} \sim p(\theta|\psi^{(0)})$
end if
10: $Z^{(0)} \leftarrow Z^{conditional} \sim p(Z|\theta^{(0)}, \psi^{(0)}, M, \eta^{(0)})$
11: **Burning phase:** run the Gibbs sampler for R_0 iterations
12: **while** $\epsilon > 0.05$ **and** $u \leq U$ **do**
13: Iterate the Gibbs sampler to get $(Z^{(r)}, \theta^{(r)}, \psi^{(r)})$, $1 \leq r \leq R_1$
14: Update $\widehat{\eta}^{(u)}$ according to the MCMC EM formula (6)
15: $\epsilon \leftarrow ||\widehat{\eta}^{(u)} - \widehat{\eta}^{(u-1)}||_\infty$
16: $u \leftarrow u + 1$
end while
17: **Final run:** Set $\widehat{\eta} \leftarrow \widehat{\eta}^{(u)}$ and iterate the Gibbs sampler to obtain the final sequence of samples $(Z^{(r)}, \theta^{(r)}, \psi^{(r)})$, $1 \leq r \leq R_2$

2.3 Model dimension

The samples for θ generated by the last iteration of the MCMC EM analysis are considered for the estimation of the number of mutational signatures N. To this end let $T = \min\{K, G\} - 1$ and then for each $1 \leq k \leq T$ let

$$\text{BIC}_k^{(r)} = 2\ln \mathcal{L}(m; \theta_k^{(r)}) - k(G + K)\ln G, \quad r = 1, \ldots, R$$

with $\theta_k^{(r)}$ as the sequence of sampled signatures and exposure matrices of rank k. It is important to observe that in contrast to the approach in Fischer *et al.* (2013), the evaluation of the BIC here does not requires any further approximation because the likelihood $\mathcal{L}(\theta_N^{(r)}; m)$ is directly available, see (s_1). Let $\overline{\text{BIC}}_k$ be the median of $\{\text{BIC}_k^{(r)}, r = 1, \ldots, R\}$. The number of signatures N is estimated as

$$N = \{1 \leq k \leq T : \overline{\text{BIC}}_k \geq \overline{\text{BIC}}_q, \, q \neq k\}.$$

The evaluation of the Bayesian information criterion for all $1 \leq k \leq T$ can be expensive for relatively large T because this involves a full MCMC EM analysis at each k. The Algorithm 2 describes a simple mode finding strategy that works for unimodal BIC sequences and reduces the overall computational cost.

2.4 Differential exposure score

The exposure matrices $E^{(r)}$, $1 \leq r \leq R$, keep important information about the contribution of each signature across the genome samples. This can be associated with independent knowledge such as clinical data in order to check how the activity of each mutational process correlates to the latter. In particular, when *a priori* information motivates the division of samples in two or more categories, we propose the use of the Kruskal–Wallis test to check whether exposure values are significantly different among categories. This test is applied to each sample of the exposure matrix $E^{(r)}$ generating a set of P-values, $p^{(r)}$, $1 \leq r \leq R$, for each signature. The median of minus the logarithm of these values defines what we call the *Differential Exposure Score* (DES). Signatures with a DES above a prescribed level are considered as differentially active among groups.

2.5 Genome sample classification

The samples $E^{(r)}$, $1 \leq r \leq R$, can also be used for the classification of genomes. The same as in Section 2.4, assume there exists prior

Algorithm 2 Model Selection

1: $\delta \leftarrow 2^d$ for $d \in N$ such that $\frac{1}{8}T < \delta \leq \frac{1}{4}T$
2: $\mathcal{T} \leftarrow \{1, 1+\delta, 1+2\delta, \ldots, T\}$
3: $i \leftarrow 1$
4: $k^* \leftarrow k \leftarrow \mathcal{T}[1]$
5: Compute $\overline{\text{BIC}}_{k^*}$ via Algorithm 1 with $N \leftarrow k^*$
6: **while** $k < (k^* + 2\delta)$ **and** $i < length(\mathcal{T})$ **do**
7: $k \leftarrow \mathcal{T}[i+1]$
8: Compute $\overline{\text{BIC}}_k$ via Algorithm 1 with $N \leftarrow k$
9: **if** $\overline{\text{BIC}}_k > \overline{\text{BIC}}_{k^*}$ **then** $k^* \leftarrow k$
10: $i \leftarrow i+1$
end while
11: **while** $\delta > 1$ **do**
12: $\delta \leftarrow 2^d/2$
13: $\mathcal{T} \leftarrow \{k^* - \delta, k^*, k^* + \delta\}$
14: $k^* \leftarrow \{k \in \mathcal{T} : \overline{\text{BIC}}_k \geq \overline{\text{BIC}}_q, q \in \mathcal{T}\}$
end while
15: **Final estimate:** $N \leftarrow k^*$

information motivating the division of samples in two or more categories and suppose there are genomes for which this information is not available (unlabelled samples). Assigning those genomes to one of the categories based on their mutational profiles could be of interest, especially in clinical settings. We propose the use of classification algorithms, e.g. *k*-Nearest Neighbours to accomplish this task. The selected algorithm is applied to each sample $E^{(r)}$ in order to classify the unlabelled genomes according to their exposures to mutational processes. This procedure generates a set of possible classifications for each unlabelled sample, and one can apply a majority rule to find the final group assignment. Unlabelled samples are assigned to a group if there is more than 75% of agreement among possible classifications, otherwise they are labeled as *undefined*. This approach clearly provides a valuable tool for prognostic.

3 Methods

3.1 Data

Single base substitutions were mapped onto trinucleotide sequences by including the 5′ and 3′ neighbouring base context to construct a $96 \times G$ matrix of mutation counts. The (i, j)th element of the opportunity matrix was computed as the frequency of the triplets in the jth genome where the ith mutation can occur. A dataset containing 183916 somatic point mutations from $G = 21$ breast cancer genomes was obtained from ftp://ftp.sanger.ac.uk/pub/cancer/NikZainalEtAl, by following the instructions in Nik-Zainal *et al.* (2012), Table S1. A second dataset containing 38 157 curated somatic mutations identified in $G = 114$ gastric cancer genomes were retrieved from the portal of The Cancer Genome Atlas (TCGA). This data is a subset of the original data described in Bass *et al.* (2014), restricted to samples grouped according to Laurén's classification. Additional details can be found as supplementary material.

3.2 `signeR`

All analyses described here are implemented in the open-source R language. Low level functions for the generation of random samples were coded in C^{++}. The design of our algorithm provides a combination of speed and low memory overhead enabling the execution on a standard computer. The stand alone algorithm, `signeR`, is available at http://bioconductor.org/packages/signeR. This package allows the extraction of the sequence context of somatic variants necessary to construct the *M* matrix from VCF files and also provides a variety of graphics to facilitate the interpretation of results. Further instructions about how to install and run this software are included as supplementary material.

4 Results

4.1 The 21 breast cancer data

The analysis of the 21 breast cancer data made by considering opportunities revealed 5 distinct signatures (Fig. 2A) that agree well with existing knowledge as documented in Sanger's catalogue of somatic cancer mutations (COSMIC, http://cancer.sanger.ac.uk/cosmic/signatures), and in Helleday *et al.* (2014) and Alexandrov (2013). The number of signatures necessary to describe the data was obtained by considering the median BIC value out of the set of values computed via the MCMC samples. The BIC boxplots obtained by varying *N* from 1 to 12 are shown in Figure 2B. Signatures S_1 and S_5 (respectively signature 2 and 13 in COSMIC) are attributed to activity of the APOBEC family of cytidine deaminases. Signature S_2 (signature 1 in COSMIC) is associated with a process initiated by the spontaneous deamination of 5-methylcytosine and correlates with the patient age at cancer diagnosis. Signature S_3 (signature 3 in COSMIC) is associated with failure of DNA double-strand break-repair by homologous recombination, whereas signature S_4 has not been reported previously.

Results for the differential exposure score obtained while grouping the data into two categories defined by samples with and without germinative mutations in *BRCA1* and *BRCA2* genes are presented in Figure 2C. The proteins encoded by the genes *BRCA1* and *BRCA2* play important roles in maintaining genomic stability and are involved in a variety of cellular processes such as damage, signalling and DNA repair (Liu and West, 2002). Disruption of such processes often leads to a rapid and widespread accumulation of somatic mutations in cancer cells (Lord and Ashworth, 2016). DES highlights three mutational signatures (Fig. 2: S_3–S_5), one is associated with inactivating mutations in *BRCA1*/*BRCA2* genes (S_3), another is implicated with the activity of APOBEC genes (S_5). Taken together, these findings are consistent with existing knowledge and, at the same time, they reinforce and help demonstrating the failure in the response to DNA damage by homologous recombination in BRCA-defective breast cancer. Interestingly, the signature S_4 is preferentially exposed in patients with deletions and mutations in gene p53 (Tables S.1 and S.2, supplementary material), associated with more aggressive tumours in triple negative breast cancers (Dang and Peng, 2013). We observe that the signatures S_1 and S_2, which are not differentially exposed, have been reported as being present in most cancer types by Sanger's catalogue. We conclude that the DES method is quite effective at revealing genotype-phenotype relationships between groups of interest.

A leave-one-out cross-validation strategy was applied to test our classification approach by examining the same sample categories used in the DES analysis. This study is motivated by the fact that patients with mutational profiles similar to those found in genomes with mutated BRCA genes could respond to treatments targeting defective DNA double-strand break repair mechanisms (Lord and Ashworth, 2016). Each one of the 21 genome samples had its label removed and was then subsequently classified based on the remaining 20 samples. Results presented in Figure 2D show that only one

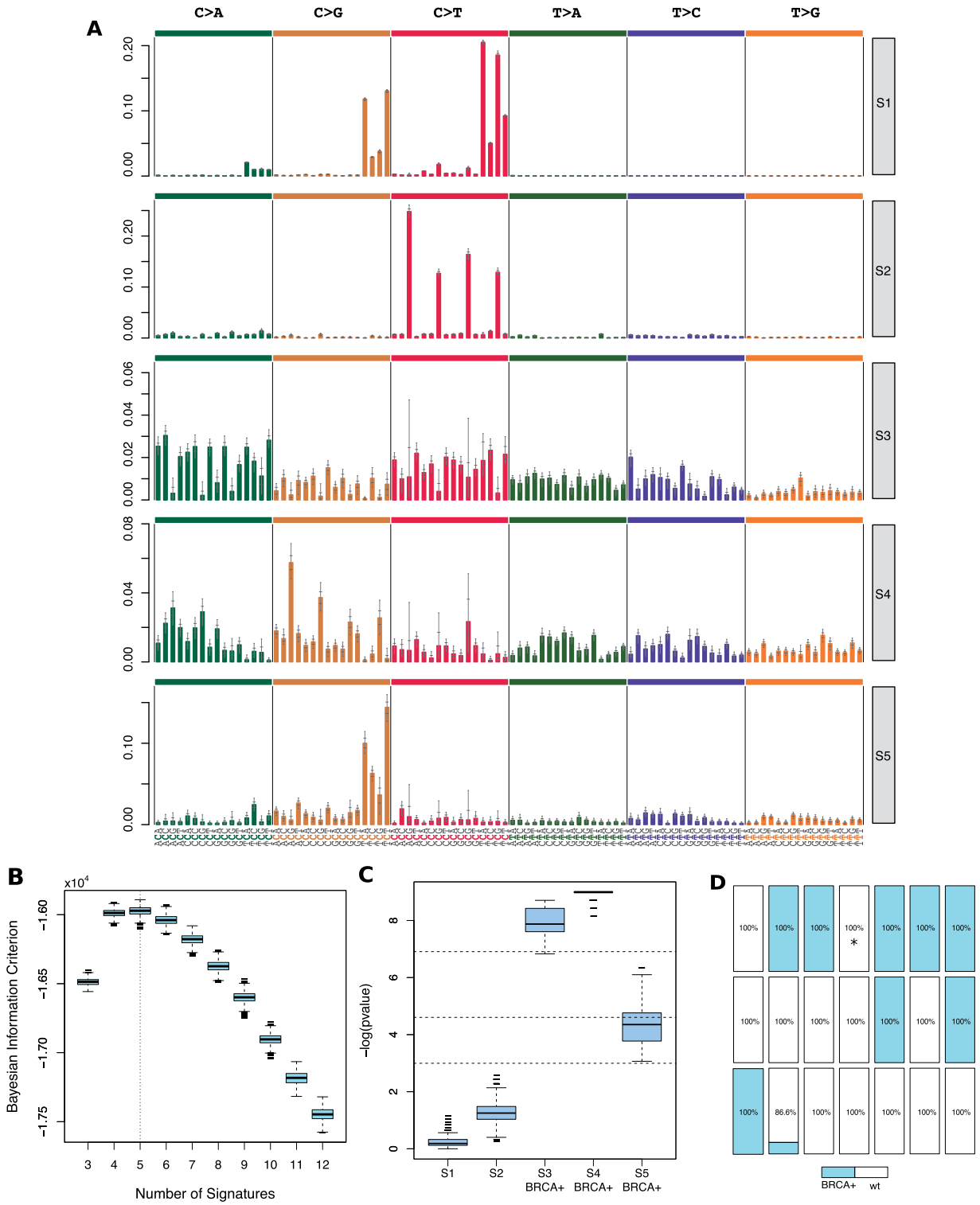

Fig. 2. Results for the 21 breast cancer data. **A** presents the five signatures obtained for the highest NMF model rank according to the BIC score presented in **B**. Signatures are labelled according to the order induced by the total signature exposure defined as $\widehat{e}_n = \sum_j \widehat{e}_{nj}$, with S_1 being the most exposed signature. Bars are located at the MCMC sample median, i.e. $\widehat{P}$, while other horizontal level marks are located at the sample percentiles 0.05, 0.25, 0.75 and 0.95. B: Boxplots for the $\text{BIC}_k^{(r)}$, $1 \leq r \leq R$, values obtained at various NMF ranks, N. **C**: Differential exposure scores - signatures showing median of log-P-values above thresholds were selected as differentially active among groups, and labels show group where they were most active. Dashed horizontal lines are located at the levels 0.05, 0.01 and 0.001. **D**: Classifications obtained for each breast cancer genome based on the remaining 20 samples. The sample marked as '*' was the only misclassification found, 'wt' stands for wild type *BRCA1*/*BRCA2* condition. Percentages show the proportion of agreement among classifications for each genome sample (Color version of this figure is available at *Bioinformatics* online.)

sample carrying a mutation at the *BRCA1* gene was misclassified, thus reinforcing the efficacy of our classification approach.

4.2 Simulation study

Synthetic datasets mimicking real observations were assembled by taking a group of four mutational signatures commonly found in breast cancer genomes. These include the signatures 1, 2, 3 and 13 described in Sanger's catalogue. Throughout, let $\tilde{P}$ denote the resulting signature matrix. The exposures are generated by maximising the likelihood $\mathscr{L}(\theta; m)$ for a given data sample m with respect to the exposure matrix E by assuming the data as being generated by $\tilde{P}$. The maximization of the likelihood in (s_1) with P fixed at $\tilde{P}$ is achieved by using R's **nlopt** package, but it may also be made by using Lee and Seung (2001) multiplicative update algorithm to solve (1). A matrix of simulated mutation counts $\tilde{m}$ is finally generated by sampling each entry $\tilde{m}_{ij}$ from a Poisson distribution with rate $(\tilde{P}\tilde{E} \circ W)_{ij}$.

Two synthetic datasets $\tilde{m}_1$ and $\tilde{m}_2$ were generated by using the 21 breast cancer data respectively without and with opportunities, i.e. by setting $W = \mathbf{1}$ or taking W from the real data. The former is used to compare the estimates for θ produced by `signeR` and the method in Alexandrov *et al.* (2013). The latter was used to establish a comparison between `signeR` and the method in Fischer *et al.* (2013). We refer to these two alternative methods respectively as LBA and EMu. The dataset $\tilde{m}_1$ was analyzed 100 times by both `signeR` and LBA and $\tilde{m}_2$ 500 times by `signeR` and EMu. The NMF rank was set to 4 for all analyses made by LBA. Whereas all analyses made with `signeR` (500/500) correctly estimated four signatures, only 51/500 of the analyses performed by EMu detected four signatures, the remaining 449/500 analyses estimated only three. The accuracy of each method was compared by the sum of squared errors between $\tilde{P}$ and the estimated signature matrix $\widehat{P}$, defined by squaring the Frobenius norm of $\widehat{P} - \tilde{P}$. This actually led to the consideration of

$$\min_{\sigma} \|\tilde{P} - \widehat{P}[\sigma]\|_F^2 = \min_{\sigma} \sum_{in} |\tilde{p}_{in} - \widehat{p}_{i\sigma(n)}|^2,$$

where $\widehat{P}[\sigma]$ is a permutation of the columns of $\widehat{P}$ introduced to account for the order in which each method exports the signatures. Only those runs where EMu correctly estimated the dimension of P where included for this analysis. These results are presented in Figure 3A, B. Clearly, the estimates produced by `signeR` are more accurate than those obtained by EMu ($P < 2.2e{-}16$, Wilcoxon rank sum test with continuity correction) or by LBA ($P < 2.2e{-}16$). For the analysis with opportunities, the mean and the standard deviation for the sum of squared errors for `signeR` are 0.095 and 0.016, and for EMu respectively 0.23 and 0.007. For the analysis without opportunities the values for `signeR` are 0.044 and 0.029 and for LBA, 0.203 and 0.012.

Further insights into the differences between `signeR` and EMu can be gained by inspection of the likelihood at the estimates for P and E, despite `signeR` being not just a likelihood maximization technique. An analogous comparison between `signeR` and LBA was not considered because LBA sets by default various entries of P to 0, leading to an undefined log likelihood. The estimates obtained by EMu cluster about two different likelihood values (Fig. 4), the lower one is identified with those runs where only three signatures are found and the higher with those instances with four signatures. The estimates obtained for `signeR` cluster at a single likelihood value. These results reveal the stability of `signeR` as opposed to EMu. It is well known that the optimization approach to NMF, as posed by (1), is very sensitive to the initial condition because of high dimensionality and because (1) does not have a unique global minimum. This problem has deserved special attention in the optimization community and several initialization strategies in this context have been suggested, see for instance Berry *et al.* (2007); Boutsidis and Gallopoulos (2008). Both the methods by Fischer *et al.* (2013) and Alexandrov *et al.* (2013) do not take these into account and consider instead a random initialization for the matrices P and E.

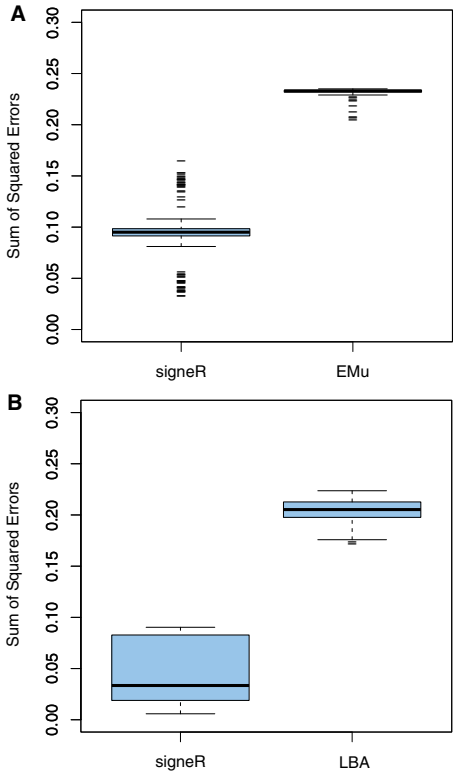

Fig. 3. Sum of squared errors for the estimated signatures. **A**: comparison between `signeR` and EMu by modelling a synthetic dataset with opportunities. **B**: comparison between `signeR` and LBA by modelling a dataset without opportunities

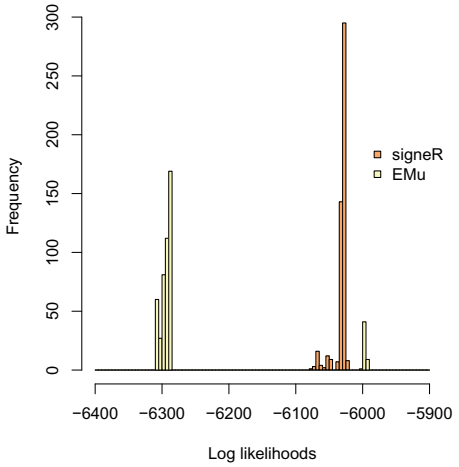

Fig. 4. Histogram of Log likelihood values at the estimates for P and E obtained by the analysis of a 4 signatures synthetic dataset via `signeR` and EMu. The analysis was made by including a mutation opportunity matrix W

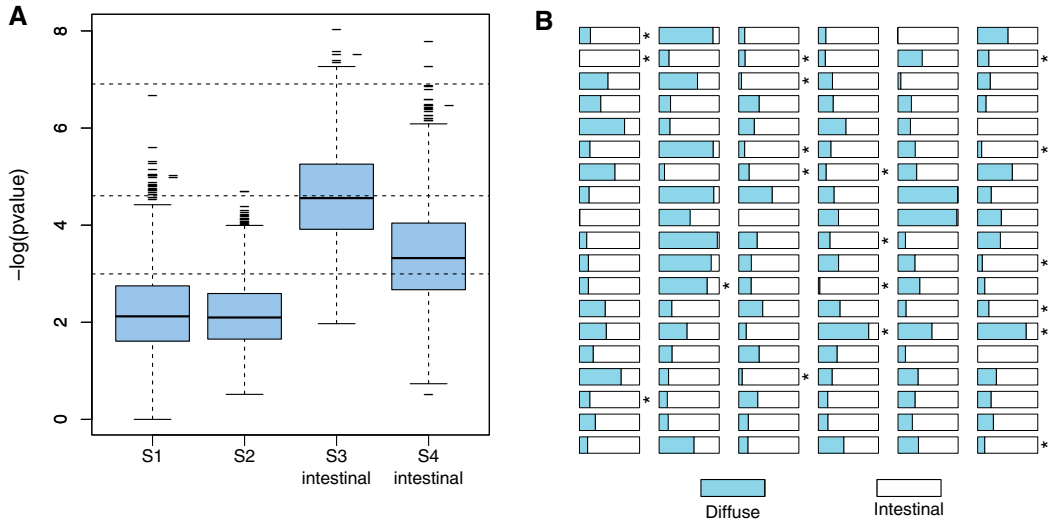

Fig. 5. A: DES analysis for the gastric cancer dataset. **B**: posterior sample classification for each gastric cancer genome, based on the remaining samples. The colors for each sample show the proportion of agreement among the a posteriori classifications. Samples are ordered by column, according to Table S.6 in the supplementary material. Misclassifications are marked with *

4.3 Gastric cancer dataset

To further demonstrate our framework this section presents the analysis of 114 gastric cancer (GC) genomes sequenced at TCGA. Although heterogeneous, most GC cases can be grouped as intestinal or diffuse subtypes, according to the classic Lauren's histological classification. Here we aim at the identification of signatures that show significantly different levels of exposure in the Laurén's subtypes. The analysis with `signeR` reveals the data as being best described by 4 signatures (supplementary Figs S.9, S.10). The DES presented in Figure 5 shows that two of these signatures are differentially enriched for the intestinal type which has a better prognosis, Ma *et al.* (2016). These signatures are described at the COSMIC database as Signatures 3 and 17 respectively. None of the signatures found is significantly more active in the diffuse group. We also considered the sample classification approach by using a weight parameter to restrict the analysis to the signatures highlighted in the DES analysis. By following the same cross-validation approach as the one considered in the breast dataset, we were able to classify 65% of the samples (Fig. 5B). Among these our classification reached 75% of success. We conclude that the analyses pursued here provide an additional roadmap for patient stratification that could aid in planning therapeutic strategies.

5 Discussion

The detection of mutational signatures from whole genome sequencing data has significantly helped to advance the understanding of mutagenesis and the development of cancers. In this article we present a new method to identify mutational signatures based upon an empirical Bayesian treatment to the Poissonian NMF model. The empirical approach requires minimal intervention on the part of the user and is specially suited for the applied practitioner. A key aspect of our analysis is that it addresses the model selection problem directly, i.e. the estimation of the underlying number of signatures, without using further approximations or ad hoc heuristics previously considered. In addition, we introduce two concepts, namely the Differential Exposure Score (DES) and posterior sample classification, which may have potential impact in clinical practice.

The effectiveness of our method is shown by the analysis of real and synthetic datasets. The results obtained with publicly available data consisting of whole genome somatic mutations of 21 breast cancers agree well with those in previous studies. A second analysis allowed to identify two signatures in a gastric cancer subtype characterized by a good prognosis. Results obtained with synthetic data show that our method presents several advantages when compared to the two other techniques mostly used in the literature and with the same NMF parameterization as the one considered here. When compared to Fischer *et al.* (2013), our method always estimates the correct number of signatures and even in those cases where the former estimates the correct model dimension, our estimates are more accurate. A comparison against the results produced by Alexandrov *et al.* (2013) with the correct dimension also shows that our estimates are more accurate.

The estimation of the NMF rank is perhaps the most challenging question regarding statistical inferences in the mutational signature paradigm. Our approach to model choice relies on the use of the Bayes Information criterion, which is a rough but simple approximation to Bayes factors, see Kass and Raftery (1995). These factors typically require the computation of the marginal likelihood, defined as the normalising constant of the posterior $\pi(Z, \theta_N, \psi_N | M, \eta, \mathcal{M}_N)$, with $\mathcal{M}_N$ as the model indicator corresponding to the factorization rank N, and θ_N and ψ_N respectively the associated parameters and hyperparameters. By setting $\eta = \widehat{\eta}$ and conditioning on $\widehat{\eta}$, Bayes theorem renders the marginal likelihood as the ratio

$$\ell(M|\widehat{\eta}, \mathcal{M}_N) = \frac{p(M, Z, \theta_N, \psi_N|\widehat{\eta}, \mathcal{M}_N)}{\pi(Z, \theta_N, \psi_N|M, \widehat{\eta}, \mathcal{M}N)}, \tag{7}$$

with η fixed, contrary to what is assumed in definition of $\mathcal{L}(\eta; m)$ by (4). The numerator is easily computed by observing the conditional decomposition defined by the hierarchical model. Indeed, this is simply given by

$$p(Z, M|\theta_N, \psi_N, \mathcal{M}_N) p(\theta_N|\psi_N, \mathcal{M}_N) p(\psi_N|\widehat{\eta}, \mathcal{M}_N).$$

The evaluation of the denominator in (7) is however more involved as it requires the joint posterior over the latent variables

and the parameters θ_N and ψ_N. This could be estimated at any point of high probability by using the output from the MCMC EM algorithm by using the approach suggested by Chib (1995) or by other means. Although promising, further work along these lines is required.

Funding

This work was partially supported by Fundação de Apoio a Pesquisa do Estado de São Paulo (FAPESP), grant 15/19324-6. ED-N is a research fellow from Conselho Nacional de Desenvolvimento Cientifico e Tecnologico, Brazil (CNPq) and acknowledges the support received from Associacão Beneficente Alzira Denise Hertzog Silva (ABADHS) and FAPESP grant 14/26897-0.

Conflict of Interest: none declared.

References

Alexandrov,L.B. *et al*. (2013) Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*., **3**, 246–259.

Alexandrov,L.B. and Stratton,M.R. (2014) Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev*., **24**, 52–60.

Alexandrov,L.B. *et al*. (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.

Banerjee,A. *et al*. (2005) Clustering with Bregman divergences. *J. Mach. Learn. Res*., **6**, 1705–1749.

Bass,A.J. *et al*. (2014) Comprehensive molecular characterization of gastric adenocarcinoma. *Nature*, **513**, 202–209.

Berry,M.W. *et al*. (2007) Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal*., **52**, 155– 173.

Boutsidis,C. and Gallopoulos,E. (2008) Svd based initialization: a head start for nonnegative matrix factorization. *Pattern Recogn*., **41**, 1350–1362.

Brunet,J.P. *et al*. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. USA*. **101**, 4164–4169.

Casella,G. (2001) Empirical Bayes Gibbs sampling. *Biostatistics*, **2**, 485–500.

Cemgil,A.T. (2009) Bayesian inference for nonnegative matrix factorisation models. *Intell. Neurosci*., **2009**, 4:1–4:17.

Chib,S. (1995) Marginal likelihood from the Gibbs output. *J. Am. Statist. Assoc*., **90**, 1313–1321.

Ciriello,G. *et al*. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*., **22**, 398–406.

Dang,D. and Peng,Y. (2013) Roles of p53 and p16 in triple-negative breast cancer. *Breast Cancer Manag*., **2**, 537–544.

Févotte,C. and Cemgil,A.T. (2009) Nonnegative matrix factorisations as probabilistic inference in composite models. In: *Proc. 17th European Signal Processing Conference (EUSIPCO'09)*, vol. 47, pp. 1913–1917.

Fischer,A. *et al*. (2013) EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol*., **14**, R39+.

Fort,G. and Moulines,E. (2003) Convergence of the Monte Carlo expectation maximization for curved exponential families. *Ann. Statist*., **31**, 1220–1259.

Gaujoux,R. and Seoighe,C. (2010) A flexible R package for nonnegative matrix factorization. *BMC Bioinf*., **11**, 367.

George,E.I. *et al*. (1993) Conjugate likelihood distributions. *Scand. J. Stat*., **20**, 147–156.

Helleday,T. *et al*. (2014) Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet*., **15**, 585–598.

Hutchins,L.N. *et al*. (2008) Position-dependent motif characterization using non-negative matrix factorization. *Bioinformatics*, **24**, 2684–2690.

Kass,E.R. and Raftery,A.E. (1995) Bayes factors. *J. Am. Statist. Assoc*., **90**, 773–795.

Lee,D.D. and Seung,H.S. (2001) Algorithms for non-negative matrix factorization. In: Leen,T. *et al*. (eds) *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge, Massachusetts, pp. 556–562.

Liu,Y. and West,S.C. (2002) Distinct functions of BRCA1 and BRCA2 in double-strand break repair. *Breast Cancer Res*., **4**, 9–13.

Lord,C.J. and Ashworth,A. (2016) BRCAness revisited. *Nat. Rev. Cancer*, **16**, 110–120.

Ma,J. *et al*. (2016) Lauren classification and individualized chemotherapy in gastric cancer. *Oncol. Lett*., **11**, 2959–2964.

Miller,R.B. (1980) Bayesian analysis of the two-parameter gamma distribution. *Technometrics*, **22**, 65–69.

Nik-Zainal,S. *et al*. (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979–993.

Roberts,S.A. and Gordenin,D.A. (2014) Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer*, **14**, 786–800.

Rosenthal,R. *et al*. (2016) deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol*., **17**, 31.

Schmidt,M.N. *et al*. (2009) Bayesian non-negative matrix factorization. In: Adali,T. *et al*. (eds) *Independent Component Analysis and Signal Separation,* vol. 5441, of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 540–547.

Shiraishi,Y. *et al*. (2015) A simple model-based approach to inferring and visualizing cancer mutation signatures. *PLoS Genet*., **11**, 1–21.