Statistical Challenges in Mutational Signature Analyses of Cancer Sequencing Data

Víctor Velasco-Pardo, Michail Papathomas, and Andy G. Lynch

Abstract Cancer is a disease driven and characterised by mutations in the DNA. Different categorisations of DNA mutations have allowed the identification of patterns that can act as signatures for the processes that have governed the life of the cancer. Over the last decade, research groups have identified more than 100 such signatures. Mutational signature analyses are improving our understanding of cancer aetiology and have the potential to play a role in diagnosis, prognosis, and treatment choice. Consisting of the estimation of probability mass functions or weights determining non-negative weighted combinations, they are perhaps unique among comparable analyses in the medical literature, in that no confidence intervals or other representations of uncertainty are demanded when reporting the results. Here, we review the key statistical challenges for the field, assess the potential of existing approaches to adapt to those challenges, and comment on what we think are promising directions. As we deal with data that are noisy and heterogeneous, we evaluate how to present them so that models use all the information available. Often posed as a matrix factorisation problem, we argue that a fully probabilistic approach is required to quantify uncertainty around model parameters and to underpin principled study design. Lastly, we argue that novel methodology is required to evaluate uncertainties in analyses where prior information is available.

Keywords Biostatistics · Bioinformatics · Cancer · Genomics · Next-generation sequencing · Whole genome sequencing

V. Velasco-Pardo (✉) · M. Papathomas · A. G. Lynch
School of Mathematics and Statistics, University of St Andrews, St Andrews, U.K.
e-mail: vvp1@st-andrews.ac.uk

M. Papathomas
e-mail: m.papathomas@st-andrews.ac.uk

A. G. Lynch
e-mail: andy.lynch@st-andrews.ac.uk

A. G. Lynch
School of Medicine, University of St Andrews, St Andrews, U.K.

R. Bispo et al. (eds.), *Recent Developments in Statistics and Data Science*,
Springer Proceedings in Mathematics & Statistics 398,
https://doi.org/10.1007/978-3-031-12766-3_17

1 Introduction

Cancers can result from relatively few changes to a cell's DNA, but typically carry many additional somatic (i.e. occurring within the life of the patient) mutations. We can identify these mutations by sequencing and then comparing DNA from the cancer and DNA from healthy tissue from the same individual [1, 2]. "Mutation", here, refers to a wide range of events ranging from single base substitutions to larger structural variants (e.g. genomic rearrangements where large segments of a chromosome might be deleted, duplicated, or have their orientation inverted [3]). See, e.g. [4] for a review of mutation classes.

Somatic mutations are the result of biological mechanisms, termed mutational processes, associated with characteristic patterns of mutations or mutational signatures, described by means of probability mass functions over mutational categories [5]. Therefore, the catalogue of somatic mutations observed in an individual cancer genome can be thought of as a mixture of the mutational signatures that have acted on the tumour over time.

Some mutational processes act continuously throughout life [6], while others arise as a result of exposures to carcinogens [7, 8]. They might be ongoing, intermittent, or might have stopped [4]. Some processes are associated with germline mutations in tumour suppressor genes, such as BRCA1/2 [5, 9]. Cancer genomes contain the imprint of many such processes to differing degrees. Consequently, the goals of mutational signature analyses are to infer from the somatic mutations in tumours (1) the signatures of mutational processes, (2) the contribution of each process to individual cancer genomes, and (3) when those processes contributed.

To achieve those goals, a range of mathematical methods have been, and are being, developed [10–21] (for a review, see, e.g. [22, 23]). Their application to data sets of ever-increasing size and complexity has resulted in a remarkable improvement of our understanding of cancer and its causes [24]. More than a hundred inferred mutational signatures are available to the wider research community [24, 25]. In the context of personalised medicine, these have a remarkable potential to stratify cancer patients [26, 27] and to predict response to treatment [28].

1.1 Modelling Framework

Data Gathering. In the context of mutational signature analyses, we investigate data sets generated using next-generation sequencing and analysis pipelines (involving (a) sequencing, (b) alignment to a reference genome, (c) often-probabilistic mutation calling, and (d) post-processing). The output is a list of "mutations" observed in the tumour. Often, data are not solely collected for the purpose of signature analysis.

In the sequencing step, short segments of DNA from both tumour and matched healthy tissue are read as base sequences. Each of those "reads" covers 100–250 base pairs and may contain errors. We define the *coverage* of an individual base to be the number of times it has been sequenced. Additionally, we define the *sequencing*

depth of an experiment to be the average number of times a base is sequenced. While sequencing depth is typically set by the investigator, coverage is not uniform across genomic regions. In particular, regions with a high prevalence of Cs and Gs are susceptible to low coverage [29].

Sequence reads are then aligned to a *reference genome*, and aligned reads from both tissues are presented to a "mutation caller" that determines whether a mutation is present at a given locus by means of a statistical test. Thus, there must be a balance between sensitivity and specificity that will differ between cancer types. Additionally, that balance is unlikely to be uniform across mutation types. Thus, the systematic bias introduced in this step will be propagated to mutational signature analyses, affecting inferences. This problem can be exacerbated by the application of post-calling filters [30, 31].

Mutational Signatures and Mutational Catalogues. For the mutational class being considered, biologically meaningful categorisations must be defined (see, e.g. [4] for a review) and we denote the resulting categories by $k = 1, \ldots, K$. We define a mutational signature, $\boldsymbol{s}_n = (s_{1n}, \ldots, s_{Kn})^T$, to be a probability mass function over the K categories, with s_{kn} denoting the probability that a mutation generated by signature n is of type k.

We now consider the mutational catalogues of G cancer patients and assume that they have been exposed to N mutational processes. The observed number of mutations of category k in patient g, m_{kg}, is approximately

$$m_{kg} \approx \sum_{n=1}^{N} s_{kn} e_{ng} \tag{1}$$

where e_{ng} denotes the exposure of patient g to signature n, that is, the number of mutations attributed to that signature. In matrix form,

$$\boldsymbol{M} \approx \boldsymbol{S} \times \boldsymbol{E} \tag{2}$$

where $\boldsymbol{M} = [\boldsymbol{m}_1 \cdots \boldsymbol{m}_G]$, $\boldsymbol{S} = [\boldsymbol{s}_1 \cdots \boldsymbol{s}_N]$, and $\boldsymbol{E} = [\boldsymbol{e}_1 \cdots \boldsymbol{e}_G]$.

1.2 Mathematical Approaches to Mutational Signatures

We will consider two problems. The first, termed de novo signature extraction, consists in estimating $\boldsymbol{S}$ and $\boldsymbol{E}$ for known $\boldsymbol{M}$. The second, termed refitting, consists in estimating $\boldsymbol{E}$ for known $\boldsymbol{M}$ and $\boldsymbol{S}$.

De Novo Signature Extraction. This problem, consisting of estimating $\boldsymbol{S}$ and $\boldsymbol{E}$ given $\boldsymbol{M}$ in (2), was originally posed as the following non-convex optimisation problem:

$$\arg \min_{\boldsymbol{S} \geq 0, \boldsymbol{E} \geq 0} ||\boldsymbol{M} - \boldsymbol{S}\boldsymbol{E}|| \tag{3}$$

where $||\cdot||$ denotes an appropriate norm. This approach, termed Non-Negative Matrix Factorisation (NMF) [32], is taken by the original and arguably most popular method, `SigProfiler` [10, 24]. Several other software packages are available implementing similar solutions based on NMF [11, 25, 33–36]. An alternative method is `EMu` [14], which considers the exposures to be nuisance parameters and uses the EM algorithm to estimate the matrix $\boldsymbol{S}$.

A slightly different approach is to place (2) in a Bayesian setting, as done by `SignatureAnalyzer` [12, 13], `signeR` [15], and `sigfit` [16]. Briefly, prior distributions are placed on the elements of $\boldsymbol{S}$ and $\boldsymbol{E}$, and a likelihood function is assumed for the elements of $\boldsymbol{M}$. `SignatureAnalyzer` performs Maximum A Posteriori estimation of $\boldsymbol{S}$ and $\boldsymbol{E}$ using the methodology developed by Tan and Févotte [37]. Alternatively, the other two methods use different MCMC algorithms [38–40] to draw from the posterior distributions of $\boldsymbol{S}$ and $\boldsymbol{E}$.

Those methods also differ in their model selection criterion (Table 1). For brevity, we refer the reader to [22] for a thorough albeit somewhat dated summary.

Table 1 Overview of methods for de novo mutational signature analysis. The third column indicates, if relevant, a point estimation criterion, a posterior sampling method, and a model selection criterion. NMF, PCA, MLE, MAP, EM, BIC, and HMC stand for Non-negative Matrix Factorisation, Principal Component Analysis, Maximum Likelihood Estimation, Maximum A Posteriori, Expectation Maximisation, Bayesian Information Criterion, and Hamiltonian Monte Carlo

Software	Method	Estimation methods
`SigProfiler` [24]	NMF [32]	MLE – Ad hoc
`SomaticSignatures` [11]	NMF/PCA [32]	Optimisation – –
`SignatureAnalyzer` [12, 13]	Bayesian NMF [37]	MAP – Not needed
`EMu` [14]	Poisson model	MLE (EM) – BIC
`signeR` [15]	Bayesian NMF [38, 39]	– Gibbs BIC
`sigfit` [16]	Bayesian NMF	– HMC (`stan` [40]) Ad hoc
`SparseSignatures` [17]	Sparse NMF	– – Cross validation

The Bayesian Nonparametric Alternative. An alternative approach to the methods described above is the one by Roberts [18], implemented in the `R` package `hdp`, using the methodology of Teh et al. [41]. Here, we are not presented with vectors of counts but with lists of mutations.

Specifically, we are presented with a data set $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_J)$ where $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jn_j})^T$ is the list of mutations observed in the jth patient. Within this framework, patients are assumed to be exchangeable, i.e. the joint probability distribution $p(X)$ does not depend on the ordering of patients. Similarly, mutations are assumed to be partially exchangeable, meaning that $p(X)$ is independent of the ordering of mutations within a specific patient. Observations are assumed to be drawn from a categorical distribution:

$$x_{ji}|\boldsymbol{\theta}_{ji} \sim \text{Categorical}(\boldsymbol{\theta}_{ji}) \tag{4}$$

The parameters $\boldsymbol{\theta}_{ji}$ of the discrete distributions are drawn from G_j, a realisation of the Dirichlet Process associated with the jth patient, whose base measure G_0 is distributed according to a "global" DP with base measure H and concentration parameter γ. Formally,

$$\boldsymbol{\theta}_{ji}|G_j \sim G_j \tag{5}$$

$$G_j|\alpha, G_0 \sim \text{DP}(\alpha, G_0) \tag{6}$$

$$G_0|\gamma, H \sim \text{DP}(\gamma, H) \tag{7}$$

where $\text{DP}(\cdot, \cdot)$ denotes a Dirichlet Process [41]. That is a nonparametric hierarchical prior that does not assume a fixed number of components and has three hyperparameters: H is the mean of the prior distribution over the signatures, and γ and α control the variability around that mean at the global and patient level, respectively. Often, H is conveniently set to Dirichlet$(1, \ldots, 1)$, a flat prior over the $(K-1)$-simplex, and non-informative Gamma hyper-priors are placed on γ and α. As with any Bayesian analysis, a sensitivity analysis is required to assess the prior choice for H. The model of Eqs. (4)–(7) is referred to as the Hierarchical Dirichlet Process Mixture Model (HDPMM).

This method has several advantages over the ones reviewed above: First, the number of components (signatures) is inferred from the data, rather than fixed. Second, it naturally models the hierarchical nature of patient data. Further, it assumes naturally that the number of components grows with the number of observations, explicitly modelling the rate of growth. However, the assumption that the number of clusters grows logarithmically with the number of patients and doubly-logarithmically with the number of mutations is unchecked [42]. The main disadvantage is that, even if MCMC samplers are available, inference from the raw MCMC output is non-trivial as it requires a post-processing procedure that is currently not available.

Additionally, it should be noted that the HDPMM allows for the assumption of exchangeability at the patient level to be relaxed by extending the hierarchy of Dirichlet Processes. Patients can then be considered partially exchangeable and grouped,

Table 2 Overview of challenges, grouped by proposed statistical solution

Proposed statistical approach	Challenge
Constructing the matrix M	1. Accounting for bias and variance in M
	2. Recognising intra-tumour heterogeneity
	3. Accounting for opportunities
	4. Going beyond the 96 categories
Bayesian nonparametrics	5. Uncertainty in the number of signatures
	6. Uncertainty around the signatures
	7. Sample size calculations
Novel statistical methodology	8. Uncertainty around the exposures
	9. Obtaining separated signatures
	10. Partial information about the signatures

e.g. according to the tissue where the tumour arose [18]. However, to relax the assumption of exchangeability at the mutation level would be more challenging.

Refitting of Mutational Signatures. This is a simpler problem which consists of solving for $\boldsymbol{e}_g$ for a single patient g in (2), assuming $\boldsymbol{m}_g$ and $\boldsymbol{S}$ are known. The most popular approach is perhaps `deconstructSigs` [19]. Alternatively, one can solve (2) using, e.g. non-negative least squares [20, 43]. An attempt to quantify uncertainty by using the Bootstrap within the context of refitting has been provided by `SignatureEstimation` [20]. A Bayesian alternative that also enforces sparsity in the solution is `sigLASSO` [21]. For brevity, we do not detail these approaches here.

Statistical Challenges. Despite the advances in this area over the last decade, it is a concern that within this field, uncertainty quantification is not receiving enough attention. Even if the effort to develop new methods has been substantial, recognition of uncertainty within the discipline is surprisingly limited. While previous reviews have focused on a mathematical description of the methods [22] and their performance [23], here we focus on the key statistical challenges for the field, enumerated in Table 2. In the forthcoming sections, we describe these challenges, highlighting the potential of different methods to address these challenges.

The first group of challenges (Sect. 2) concerns the uncertainties arising from data collection. The second group (Sect. 3) concerns uncertainties in de novo analyses and how accounting for them could inform data collection. We will argue that the Bayesian Nonparametric approach is suitable to address those challenges. The third group (Sect. 4) concerns uncertainty in analyses where partial information is available. While we will highlight that progress has been made, the need to address these challenges demands the development of new methodology.

2 Challenges in Constructing $\boldsymbol{M}$

2.1 Challenge 1: Accounting for Bias and Variance in $\boldsymbol{M}$

Sequencing experiments are stochastic events, and the identification of mutations, necessary for constructing $\boldsymbol{M}$, is often based on probabilistic models [31]. $\boldsymbol{M}$ is itself therefore also an observation of a random variable. While uncertainty around the mutation calls is unavoidable, it can be reduced by increasing sequencing depth [29]. High sequencing depth increases the chance of calling subclonal mutations (see also Sect. 2.2) and reduces disagreements between mutation callers [31]. Typically, it is beneficial to increase the depth of sequencing as it results in the identification of mutations that are present in a fraction of cells. However, the benefits of doing so are marginal after a certain depth threshold, which differs across individual tumours [30]. Therefore, allocating extra resources to recruit more patients might be more cost-efficient.

As well as exhibiting variation, $\boldsymbol{M}$ will be a biased estimate of the true value. Different callers [31] and sequencing pipelines [30] can return systematically different results. Genomic context affects the power to detect mutations (via variation of sequencing coverage [44]) and the false discovery rate [31], meaning that some classes of mutation are less likely to be called correctly than others. There is potential for novel statistical developments to estimate more accurate catalogues.

Going back to the identification of mutations present in a small fraction of cells, these are more likely to have occurred more recently—and thus they are more likely to be overlooked due to insufficient coverage. If there is a change in mutational patterns over time [45], then this will cause a bias in $\boldsymbol{M}$. On the other hand, if the tumour has recently diverged into subclones, then recent mutational processes might have their impact measured on each subclone, and these processes will be over-represented relative to the truth for any cell present.

2.2 Challenge 2: Recognising Intra-Tumour Heterogeneity

Intra-tumour heterogeneity (ITH) poses a difficulty with mutational signature analyses that is not always acknowledged. Briefly, tumours are heterogeneous mixtures of cells, and we are often able to identify mutations only at the patient level (i.e. not with single cell resolution). We can sometimes infer whether a mutation is *clonal* (meaning it is present in every sampled cancer cell) or *subclonal*. Every subclonal mutation belongs to one or more *subclones*, subpopulations of cells that carry the same variants. Subclones can be inferred by clustering on the space of the *cancer cell fraction* (CCF), the unobserved proportion of tumour cells in which a mutation is present [46].

ITH in De Novo Signature Extraction. All de novo methods ignore ITH. They consider, explicitly or implicitly, mutations to be exchangeable at the patient level, ignoring their clonal status. Ideally, we would relax the assumption of exchangeability by incorporating available information regarding ITH. An interesting approach has been taken in recent studies of normal and non-neoplastic colon biopsies [47, 48] and consists of extending the tree-like hierarchical structure of the HDPMM to a further level. Then, mutations are grouped according to their subclone, which is in turn grouped according to patients. However, it remains to be shown whether this approach is applicable to cancer data.

ITH in Signature Refitting. By combining the estimation of subclones with refitting methods we can learn about the evolution of cancers [45]. One approach is to infer the subclones and then apply a refitting algorithm to each of them [49]. An alternative is implemented by `TrackSig` [50], and consists of sorting mutations by CCF (a surrogate for "age"). Refitting is then applied to "time points" of 100 mutations each. Lastly, subclones are inferred at boundaries between time points.

The first approach fails to propagate the uncertainty around subclones to the second step of the analysis. Performing inference on the subclones and the subclone-specific exposures jointly, as done by `TrackSig`, seems sensible but the current approach ignores uncertainty in the estimation of the CCF.

2.3 Challenge 3: Accounting for Opportunities

A mutation category implies a "reference state" and a "variant state". For example, consider the category "A[C>T]G" in the standard categorisation of SBSs. That category implies a reference state "ACG" and a variant state "ATG". Reference states are not uniformly distributed across the human genome and their distribution varies across cancer patients (due to copy number variation and loss of heterozygosity events).

Fischer et al. [14] have proposed to adjust the observed number of mutations of category k by the relative prevalence of that category's reference state. That relative prevalence is termed "opportunity" and, for patient g, is denoted o_{kg}. Adjusting for opportunities, (1) becomes

$$m_{kg} \approx o_{kg} \sum_{n=1}^{N} s_{kn} e_{ng} \tag{8}$$

While this approach is available in several de novo methods [14–16], it does not seem to be widely used in practice.

Opportunities, when measured, are informative about the distribution of mutations that might occur contemporaneously, but are used to analyse mutations that have occurred in the past. Copy-Number gains change the opportunities for late mutations, while loss of heterozygosity events and copy number losses effectively change the opportunities for early events. By contrast, other processes can gradually

shift the balance of opportunities. An SBS event can change three local contexts, so a hypermutation event with 1,000,000+ similar mutations would noticeably change the opportunities.

2.4 Challenge 4: Going Beyond the 96 Categories

As mentioned in Sect. 1.1, signature analyses are applicable to a range of mutational classes. Most, though, have been performed on single base substitutions (SBS) for which a canonical categorisation with 96 categories is available. Six basic categories result from considering the pyrimidine in the mutated base pair and the base to which it mutates (C>A, C>G, C>T, T>A, T>C, T>G). Considering this and the four possible nucleotides before and after the mutated base, we obtain the most common categorisation, with $4 \times 6 \times 4 = 96$ mutation types.

Further Categorisations of SBS. We could consider four flanking bases instead of two. The number of categories in this taxonomy is then $6 \times 4^4 = 1536$. While it has been shown that the two bases immediately flanking the mutated base carry a stronger signal, in some cases using this extended taxonomy has led to further resolution [24]. This taxonomy comes with its own challenges. First, we would not expect MCMC-based methods to scale to this level of resolution. Second, we would expect matrix $\boldsymbol{M}$ to contain many zeroes, requiring methods that can account for such sparsity.

A related problem is that there is currently no distance structure between mutation categories. A mutation A[C>T]G is as different from C[C >T]G as it is from T[T>A]T. While the NMF approach offers no obvious way of creating such distance structure, the one-dimensional categorical observations $x_{ji} \in \{1, \ldots, 96\}$ in the HDPMM could be replaced with three-dimensional observations $\boldsymbol{x}_{ji} = (x_{ji1}, x_{ji2}, x_{ji3})$ with $x_{ji2} \in \{1, \ldots, 6\}$ and $x_{ji1}, x_{ji3} \in \{1, \ldots, 4\}$.

Integrating Mutation Classes. Whether it would be informative for signatures to integrate all the mutation classes is a matter of debate [4, 24]. A cross-class categorisation, such as the one with 1,697 categories proposed by Alexandrov et al. [24], ignores the difference in noise and degree of sparsity between mutational classes. Performing separate analyses for each class followed by post-hoc association analysis of exposures has the drawback of ignoring uncertainty in signature attribution. Instead, we would suggest a strategy of information sharing, using class-specific categorisations and catalogues to extract signatures, but incorporating an association parameter that would quantify which signatures of diverse classes tend to occur together.

Accounting for Genomic Properties. So far, we have considered mutations from a given patient to be exchangeable. That is reasonable if we lack information to distinguish them, other than the category we are measuring. However, that is not entirely true, as each mutation has *genomic properties* (e.g. chromosome, chromatin

state, proximity to a particular binding site, etc.) that we might be able to measure. Those properties can help elucidate the aetiology of a signature, as well as help determine whether a signature is an artefact of the extraction algorithm.

Categorisations can be augmented to account for these genomic properties, but increasing the number of categories comes at a price. With that strategy, we are likely to be able to consider one genomic property at a time. Vöhringer et al. have suggested an alternative based on *non-negative tensor factorisation*, `TensorSignatures` [51]. This method scales to a large number of genomic properties. However, it has the disadvantage of not being a probabilistic method. Further methods may arise, in the spirit of `TensorSignatures`, perhaps modelling mutation categories and genomic properties with a joint probability distribution and thus relaxing the assumption of exchangeability.

3 Challenges Addressed with Bayesian Nonparametrics

3.1 Challenge 5: Uncertainty in the Number of Signatures

Parametric methods such as those based on NMF, reviewed in Sect. 1.2, assume a fixed number of signatures. Therefore, uncertainty around the number of signatures is not modelled or evaluated. Moreover, it has been argued that uncertainty around the model dimension should be disregarded as its influence in the estimation of the main signatures is marginal [4].

We argue that as the number of signatures is unknown, there is uncertainty about the true model dimension. This uncertainty can be modelled and evaluated after collecting data. A Bayesian clustering approach relaxes the assumption of a fixed number of signatures and lets this number be a parameter whose value is to be learned. This is achieved by placing a prior on the number of signatures. A nonparametric prior implies that the model dimension increases with the number of observations [52]. The assumed rate of growth depends on the chosen nonparametric prior, as briefly discussed for the HDPMM in Sect. 1.2.

The latter approach has, in our opinion, several advantages. First, avoiding an upper bound on the number of signatures is intuitively appealing, as we expect to see more signatures as more observations arrive. However, the assumption about the rate of growth is rather strong and must be checked. Second, it allows for inference to be performed on model parameters and model dimension jointly. Hence, uncertainty intervals around model parameters will reflect the uncertainty around the number of signatures (see also Sect. 3.2).

Provided with a data set, a sampler for the HDPMM will produce draws from a posterior distribution, each of them with a different number of signatures. From those draws, it is straightforward to produce a (marginal) posterior distribution over the number of signatures. As that posterior will help quantify the strength of the

signal in the data set, it must be reported along with the "most representative set of signatures". Relatedly, the required evaluation of uncertainty around signatures in that representative set is not trivial (see Sect. 3.2).

3.2 Challenge 6: Uncertainty Around the Signatures

Contrary to the usual practice in the biomedical literature, estimates of mutational signatures have typically been reported without intervals of uncertainty [5, 9, 24]. This is undesirable, as we are often interested in the possible range of values that might have generated the data. First, even if we were only interested in the "centre" of the signatures, uncertainty in estimating that centre is unavoidable. Second, if there is any randomness in the biological mechanism under which mutational processes generate mutations, we would expect them to leave slightly different "fingerprints" in each patient. Uncertainty intervals around signature probabilities should reflect that variability.

The Bayesian paradigm provides a natural setting to quantify that uncertainty. While this has been proposed in two contexts, Bayesian NMF [15, 16] and Bayesian clustering [18], we believe that the latter is more promising. This is because the Bayesian clustering approach accounts for the uncertainty in the model dimension when reporting uncertainty around the signatures (see Sect. 3.1). This can be useful considering study design (see Sect. 3.3).

The Bayesian clustering framework provides a posterior over the space of possible *partitions*. At every iteration of the MCMC sampler, every mutation is allocated to a cluster which is, in turn, characterised by $\boldsymbol{\theta}_{ji}$ in (5)–(7). The random vector $\boldsymbol{\theta}_{ji}$ represents the signature attributed to mutation x_{ji}. For ease of interpretation, a representative clustering must be determined from the MCMC output. An objective criterion must be defined to determine that "most representative set of signatures".

Once a representative set has been derived, the MCMC output can be used to determine the strength of the signal. If a signature is needed to explain the data, it will appear consistently across iterations of the sampler, and hence credible intervals around it will be narrow. Conversely, if a signature appears in the best set but does not appear throughout the MCMC output (e.g. because it might emerge admixed with similar signatures), it will be reported with wide credible intervals.

Such an approach, while needing development, would differ from the post-processing method of Roberts [18] that disregards uncertainty in clustering by assuming that every reported signature *is present across iterations of the sampler*. Rather, one of the strengths of the Bayesian clustering approach is that it allows one to assess *whether a given signature is present across iterations*.

3.3 Challenge 7: Sample Size Calculations

Since the first collection of 5 mutational signatures was found on a data set of 21 breast cancer whole genomes [9], the number of known mutational signatures has grown with the number of cancer genomes available for analysis. The first pan-cancer mutational signature study reported 21 SBS signatures in 507 genomes and 6535 exomes [5, 10], while the most recent large-scale study has reported 49 SBS signatures in 4645 genomes and 19184 exomes [24], suggesting that the rate at which new mutational signatures can be found shrinks as the number of patients and observed mutations grows. Heterogeneity within the cohort is also known to influence the power to extract signatures.

While we would expect the inventory of mutational signatures to keep increasing as new tumour samples are observed, it is good practice to make sample size calculations before collecting new samples. When making sample size calculations, it is advisable to consider (1) the number of new individuals recruited, (2) the number of mutations observed in each patient, and (3) heterogeneity within the cohort.

Whereas methods based on Non-negative matrix factorisation do not provide an obvious way of informing study design, the fully probabilistic approach of the HDPMM could be used to inform future sample collection. In particular, we would be interested in assessing the posterior probability of discovering a new signature, conditional on the data already observed and L future observations $\boldsymbol{x}_{J+1}, \boldsymbol{x}_{J+2}, \ldots, \boldsymbol{x}_{J+L}$.

The scaling properties of the HDPMM [42, 52], explained in Sects. 1.2 and 3.1, can be applied to assess that probability. Related probabilistic questions on future data collection could be answered, for example regarding heterogeneity within the cohort. This approach has been successful in other problems, such as single-cell sequencing experiments with competing budget constraints [53]. However, to avoid making false inferences, we must check that the newly discovered signatures are likely to be genuine, considering the level of support for them by the observed data.

4 Challenges Requiring a New Modelling Approach

4.1 Challenge 8: Uncertainty Quantification Around Exposures

Remember that the goal of a refitting analysis is to solve for $\boldsymbol{e}_g$ in (2) for a single patient g. In Sect. 1.2, we have briefly reviewed the mathematical methods available for performing this task. To date, it remains the case that most point estimates in refitting analyses are reported without an uncertainty interval (see, e.g. [54]).

So far, there has been one attempt to provide confidence intervals around the estimates of a refitting analysis, provided by `SignatureEstimation` [20], which uses the bootstrap to produce confidence intervals around the exposure estimates.

There is a concern though that this approach accounts at best for a fraction of the uncertainty.

Avoiding False Exposures and Obtaining a Sparse Solution. Because signatures overlap, different weighted combinations of signatures can explain a mutational catalogue equally well. Thus, it has been argued that $\boldsymbol{S}$ should include only the signatures that one could reasonably expect to see in the tissue where the tumour arose [4]. Moreover, any extra signature added to the $\boldsymbol{S}$ matrix will result in a fitted vector that better resembles the observed vector.

Those two difficulties are acknowledged and addressed by Alexandrov et al. [24]. Their solution consists in (a) including in $\boldsymbol{S}$ all the signatures that have been previously found in the relevant tissue, (b) removing signatures from $\boldsymbol{S}$ sequentially, until the removal of a single signature results in a reduction in the cosine similarity ≥ 0.01, and (c) adding to $\boldsymbol{S}$ the signatures that result in an increase in cosine similarity of ≥ 0.05, even if they have not been previously associated with the relevant tissue.

However, that approach is not without problems. First, the inference is based on ad-hoc rules and relies on cut-offs that appear arbitrary. The first suggestion from a statistical point of view would be to elucidate an informative prior distribution over the exposures. If prior information is limited to the tissue in which the tumour was observed, it might be possible to adopt a hierarchical modelling approach, with the ambition to borrow information across patients. Further, a penalty parameter could be included, ensuring that over-fitting is avoided.

Assessing All Sources of Uncertainty. In principle, to avoid underestimating uncertainty, all its sources should be modelled explicitly. Degasperi et al. [25] have argued that, even if most signatures occur in more than one tissue, the profile of each signature is tissue-specific. Therefore, the matrix $\boldsymbol{S}$ should contain signatures as extracted from tumours of the relevant tissue only. While this seems sensible, we would go further and argue that, if there is any randomness in the mechanism under which a given mutational process generates mutations, then the fingerprint of that process must differ at least slightly between patients. This must be accounted for when allocating mutations to signatures.

Another source of uncertainty that is often ignored has been termed "sampling uncertainty" by Li and colleagues [21]. It formalises the idea that uncertainty in the estimated exposures will decrease as more mutations are observed. A response to that is their method, `sigLASSO`. However, even if this method accounts for such "sampling uncertainty" in its modelling, it reports point estimates only. This is an appealing idea that could be incorporated into the other methods.

4.2 Challenge 9: Obtaining Separated Signatures

If we are looking to extract a representation of the true exposures and signatures, then it should be noted that two true but distinct signatures can be similar. This has

been highlighted as problematic, as the presence of similar signatures in the matrix $\boldsymbol{S}$ prevents unambiguous attribution of mutations to signatures [24]. We should also note that the interpretation of similarity is very much dependent on the vector space in which we are representing signatures, which is a restrictive space due to the non-negativity constraint.

To avoid such ambiguity in post-hoc refitting analysis, we can impose a sparsity constraint on de novo methods by adding a penalty term to the optimisation problem (3), as suggested by Lal et al. [17]:

$$\lambda \sum_{n=1}^{N} ||\boldsymbol{s}_n||_1 \tag{9}$$

where $||\cdot||_1$ is the L1 norm and λ can be interpreted as the data set's *degree of sparsity*. This approach results in extracting signatures that are sparse, thus making pairs of signatures more likely to be *separated*. It should be noted however that, by imposing a sparsity constraint, a restriction that may not be supported by evidence is introduced for computational and interpretational convenience.

By shrinking the signature parameters towards zero, the aforementioned sparsity constraint results in a rather strong restriction over a space that is already restrictive. This has implications for the stability of present and future signatures: presented with additional data carrying novel signatures, a de novo method may fail to find space to accommodate those novel signatures, potentially distorting old ones.

4.3 Challenge 10: Partial Information About the Signatures

With the methodology available to date, a researcher has two options when attempting to analyse data—to rely on an external collection of signatures to perform a refitting analysis or to perform a de novo analysis. However, there are situations where it would be more natural to assume an intermediate setting, where the signatures are neither known nor unknown.

In this context, it might make sense to consider an intermediate approach where partial information about the signatures is available, but they are not known precisely. This is not the same as the approach termed *fit-ext* in [16] and also implemented in [18]. That approach, consisting in setting part of the signatures matrix to point estimates derived from previous studies, ignores the uncertainty associated with those point estimates. Moreover, it does not allow for those estimates to be updated.

Rather than considering previously discovered signatures to be fixed, it seems more appropriate to incorporate knowledge obtained from previous studies through means of an informative prior distribution. This setting has, to some extent, been explored also in [16], allowing informative Dirichlet priors over both signatures and the exposures. However, there is little guidance on how to take advantage of this method. We note however two possible lines of future research within this approach. First, the Dirichlet distribution might not be flexible enough to model prior knowl-

edge about the signatures. Second, a hierarchical prior over the exposures might be worth considering, to borrow statistical strength between patients.

5 Conclusions

This review has set out what we perceive to be the main statistical challenges in the field of mutational signatures. While highlighting the achievements of the mutational signatures community in improving our understanding of cancer, we have drawn attention to the lack of estimates of uncertainty in such analyses. Motivated by this, and by related statistical challenges, we have highlighted the strengths of certain methods to address those challenges while also emphasising the need for future developments.

First, we have outlined four challenges involving potential errors or loss of information when constructing $\boldsymbol{M}$. We have highlighted that the problem of estimating the "true" $\boldsymbol{M}$ has been largely ignored (Sect. 2). As an alternative, we could have argued for a single Bayesian pipeline integrating mutation calling and signature analysis. However, that would set back the adoption of new methods, since mutation calling pipelines are established. Relatedly, we have underlined the promise of `TrackSig` in the study of tumour evolution, but further developments are required to account for all the uncertainties (Sect. 2.2). Similarly, we drew attention to the concept of mutational opportunities while calling for new developments to account for the opportunities' temporal evolution (Sect. 2.3).

Second, we have outlined three challenges related to uncertainty quantification in de novo applications. While NMF approaches have been augmented with probabilistic models, their lack of flexibility regarding model dimension is a drawback. We have argued that the Bayesian Nonparametrics approach, first suggested by Roberts, offers a more natural framework for assessing sources of uncertainty. However, we have argued that further study is needed to take advantage of the vast MCMC output resulting from this approach (Sects. 3.1 and 3.2). We have also discussed the potential of this fully probabilistic modelling to underpin study design, allowing practitioners to address trade-offs and optimise limited resources (Sect. 3.3).

Lastly, we have outlined three challenges for which no obvious statistical solution is available. We have highlighted the need for quantifying uncertainty in the context of refitting. We have also highlighted the recent application of statistical methods such as the Bootstrap to assess a fraction of such uncertainty, while identifying additional sources of uncertainty that are being ignored (Sect. 4.1). Finally, we have underlined the fit-ext approach as an attempt to pose an intermediate problem between de novo and refitting. However, that approach needs enhancement to account for the uncertainty around estimates obtained in previous studies (Sect. 4.3).

Acknowledgements We thank The Melville Trust for the Care and Cure of Cancer for providing financial support. We are grateful to the Editors and to an anonymous reviewer for valuable comments that helped to improve the manuscript.

References

1. Greenman, C., Stephens, P., Smith, R., et al.: Patterns of somatic mutation in human cancer genomes. Nature **446**, 153–158 (2007)
2. Stratton, M., Campbell, P., Futreal, P.: The cancer genome. Nature **458**(7239), 719–724 (2009)
3. Li, Y., Roberts, N., Wala, J., Shapira, O., Schumacher, S., et al.: Patterns of somatic structural variation in human cancer genomes. Nature **578**(7793), 112–121 (2020)
4. Koh, G., Degasperi, A., Zou, X., Momen, S., Nik-Zainal, S.: Mutational signatures: emerging concepts, caveats and clinical applications. Nat. Rev. Cancer **21**(10), 619–637 (2021)
5. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., et al.: Signatures of mutational processes in human cancer. Nature **500**(7463), 415–421 (2013)
6. Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., Stratton, M.R.: Clock-like mutational processes in human somatic cells. Nat. Genet. **47**(12), 1402–1407 (2015)
7. Brash, D.E., Rudolph, J.A., Simon, J.A., Lin, A., McKenna, G.J., et al.: A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. Proc. Natl. Acad. Sci. **88**(22), 10124–10128 (1991)
8. Denissenko, M.F., Pao, A., Tang, M., Pfeifer, G.P.: Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in p53. Science **274**, 430–432 (1996)
9. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., et al.: Mutational processes molding the genomes of 21 breast cancers. Cell **149**, 979–993 (2012)
10. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., Stratton, M.R.: Deciphering signatures of mutational processes operative in human cancer. Cell Rep. **3**, 246–259 (2013)
11. Gehring, J.S., Fischer, B., Lawrence, M., Huber, W.: SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics **31**, 3673–3675 (2015)
12. Kasar, S., Kim, J., Improgo, R., Tiao, G., Polak, P., et al.: Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. Nat. Commun. **6**(1), 8866 (2015)
13. Kim, J., Mouw, K.W., Polak, P., Braunstein, L.Z., Kamburov, A., et al.: Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet. **48**, 600–606 (2016)
14. Fischer, A., Illingworth, C.J.R., Campbell, P.J., Mustonen, V.: EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. **14**(4), R39 (2013)
15. Rosales, R.A., Drummond, R.D., Valieris, R., Dias-Neto, E., Da Silva, I.T.: signeR: an empirical Bayesian approach to mutational signature discovery. Bioinformatics **33**(1), 8–16 (2017)
16. Gori, K., Baez-Ortega, A.: sigfit: flexible Bayesian inference of mutational signatures. bioRxiv 372896 (2020)
17. Lal, A., Liu, K., Tibshirani, R., Sidow, A., Ramazzotti, D.: De novo mutational signature discovery in tumor genomes using SparseSignatures. PLoS Comput. Biol. **17**(6), e1009119 (2021)
18. Roberts, N.: Patterns of somatic genome rearrangement in human cancer. Ph.D. Thesis, University of Cambridge (2018)
19. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., Swanton, C.: DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. **17**, 31 (2016)
20. Huang, X., Wojtowicz, D., Przytycka, T.M.: Detecting presence of mutational signatures in cancer with confidence. Bioinformatics **34**, 330–337 (2018)
21. Li, S., Crawford, F.W., Gerstein, M.B.: Using sigLASSO to optimize cancer mutation signatures jointly with sampling likelihood. Nat. Commun. **11**(1), 3575 (2020)
22. Baez-Ortega, A., Gori, K.: Computational approaches for discovery of mutational signatures in cancer. Brief. Bioinform. **20**, 77–88 (2019)

23. Omichessan, H., Severi, G., Perduca, V.: Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance. PLoS ONE **14**(9), e0221235 (2019)
24. Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., et al.: The repertoire of mutational signatures in human cancer. Nature **578**(7793), 94–101 (2020)
25. Degasperi, A., Amarante, T.D., Czarnecki, J., Shooter, S., Zou, X., et al.: A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. Nat. Cancer **1**, 249–263 (2020)
26. Davies, H., Glodzik, D., Morganella, S., Yates, L.R., Staaf, J., et al.: HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. Nat. Med. **23**, 517–525 (2017)
27. Zou, X., Koh, G.C.C., Nanda, A.S., Degasperi, A., Urgo, K., Roumeliotis, T.I., Agu, C.A., Badja, C., Momen, S., Young, J., Amarante, T.D., Side, L., Brice, G., Perez-Alonso, V., Rueda, D., Gomez, C., Bushell, W., Harris, R., Choudhary, J.S., Consortium, G.E.R., Jiricny, J., Skarne, W.C., Nik-Zainal, S.: A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. Nat. Cancer **2**, 643–657 (2021)
28. Zhao, E.Y., Shen, Y., Pleasance, E., Kasaian, K., Leelakumari, S., et al.: Homologous recombination deficiency and platinum-based therapy outcomes in advanced breast cancer. Clin. Cancer Res. **23**, 7521–7530 (2017)
29. Sims, D., Sudbery, I., Ilott, N.E., Heger, A., Ponting, C.P.: Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. **15**(2), 121–132 (2014)
30. Alioto, T.S., Buchhalter, I., Derdak, S., Hutter, B., Eldridge, M.D., et al.: A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat. Commun. **6**(1), 10001 (2015)
31. Krøigård, A.B., Thomassen, M., Lænkholm, A.-V., Kruse, T.A., Larsen, M.J.: Evaluation of nine somatic variant callers for detection of somatic mutations in exome and targeted deep sequencing data. PLoS ONE **11**, e0151664 (2016)
32. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
33. Ardin, M., Cahais, V., Castells, X., Bouaoun, L., Byrnes, G., et al.: MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. BMC Bioinform. **17**(1), 170 (2016)
34. Mayakonda, A., Lin, D.-C., Assenov, Y., Plass, C., Koeffler, H.P.: Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. **28**, 1747–1756 (2018)
35. Blokzijl, F., Janssen, R., van Boxtel, R., Cuppen, E.: MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. **10**(1) (2018)
36. Wang, S., Tao, Z., Wu, T., Liu, X.-S.: Sigflow: an automated and comprehensive pipeline for cancer genome mutational signature analysis. Bioinformatics **37**, 1590–1592 (2021)
37. Tan, V.Y., Févotte, C.: Automatic relevance determination in nonnegative matrix factorization with the /spl beta/-divergence. IEEE Trans. Pattern Anal. Mach. Intell. **35**(7), 1592–1605 (2013)
38. Cemgil, A.T.: Bayesian inference for nonnegative matrix factorisation models. Comput. Intell. Neurosci. 785152 (2009)
39. Schmidt, M.N., Winther, O., Hansen, L.K.: Bayesian non-negative matrix factorization. In: International Conference on Independent Component Analysis and Signal Separation, pp. 540–547. Springer (2009)
40. Gelman, A., Lee, D., Guo, J.: Stan: a probabilistic programming language for Bayesian inference and optimization. J. Educ. Behav. Stat. **40**(5), 530–543 (2015)
41. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical Dirichlet Processes. J. Amer. Stat. Assoc. **101**(476), 1566–1581 (2006)
42. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Ann. Stat. 1152–1174 (1974)
43. Krüger, S., Piro, R.M.: decompTumor2Sig: identification of mutational signatures active in individual tumors. BMC Bioinform. **20**(4), 1–15 (2019)

44. Barbitoff, Y.A., Polev, D.E., Glotov, A.S., Serebryakova, E.A., Shcherbakova, I.V., et al.: Systematic dissection of biases in whole-exome and whole-genome sequencing reveals major determinants of coding sequence coverage. Sci. Rep. **10**(1), 1–13 (2020)
45. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., et al.: The evolutionary history of 2,658 cancers. Nature **578**(7793), 122–128 (2020)
46. Dentro, S.C., Wedge, D.C., Van Loo, P.: Principles of reconstructing the subclonal architecture of cancers. Cold Spring Harb. Perspect. Med. **7**(8), a026625 (2017)
47. Lee-Six, H., Olafsson, S., Ellis, P., Osborne, R.J., Sanders, M.A., et al.: The landscape of somatic mutation in normal colorectal epithelial cells. Nature **574**(7779), 532–537 (2019)
48. Olafsson, S., McIntyre, R.E., Coorens, T., Butler, T., Jung, H., et al.: Somatic evolution in non-neoplastic IBD-affected colon. Cell **182**(3), 672–684 (2020)
49. Yates, L.R., Knappskog, S., Wedge, D., Farmery, J.H., Gonzalez, S., et al.: Genomic evolution of breast cancer metastasis and relapse. Cancer Cell **32**(2), 169–184 (2017)
50. Rubanova, Y., Shi, R., Harrigan, C.F., Li, R., Wintersinger, J., et al.: Reconstructing evolutionary trajectories of mutation signature activities in cancer using TrackSig. Nat. Commun. **11**(1), 1–12 (2020)
51. Vöhringer, H., Hoeck, A.V., Cuppen, E., Gerstung, M.: Learning mutational signatures and their multidimensional genomic properties with TensorSignatures. Nat. Commun. **12**(1), 1–16 (2021)
52. Teh, Y.W., Jordan, M.I.: Hierarchical Bayesian nonparametric models with applications. In: Bayesian Nonparametrics, vol. 1, pp. 158–207. Cambridge University Press, Cambridge (2010)
53. Camerlenghi, F., Dumitrascu, B., Ferrari, F., Engelhardt, B.E., Favaro, S.: Nonparametric Bayesian multiarmed bandits for single-cell experiment design. Ann. Appl. Stat. **14**(4), 2003–2019 (2020)
54. Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., et al.: Tumor and microenvironment evolution during immunotherapy with nivolumab. Cell **171**(4), 934–949 (2017)