sigfit: flexible Bayesian inference of mutational signatures

Kevin Gori and Adrian Baez-Ortega

Transmissible Cancer Group, Department of Veterinary Medicine, University of Cambridge, Madingley Road, Cambridge, CB3 0ES, United Kingdom

The premise of mutational signature analysis is that genomes acquire somatic mutations through the collective action of discrete mutational processes, and its aim is to infer the mutational signatures of these processes, together with their respective levels of activity, from observed mutation data. Different models for mutational signature analysis have been developed in recent years, most prominently based on non-negative matrix factorisation (NMF). Here we present sigfit, an R package for mutational signature analysis that employs Bayesian inference to perform fitting of mutational signatures to data and extraction of mutational signatures from data, using both NMF-inspired and alternative probabilistic models. We compare the performance of sigfit to prominent existing software for mutational signature analysis, and find that it compares favourably. Furthermore, sigfit introduces novel probabilistic models to perform simultaneous fitting and extraction of mutational signatures, with the aim of powering the detection of rare or previously undescribed signatures. The package provides user-friendly data visualisation routines and is easily integrable with other bioinformatic R packages.

mutational signatures | cancer | Bayesian inference
Correspondence: *kcg25@cam.ac.uk*

Background

The study of mutational signatures — the idiosyncratic patterns of somatic alterations left by mutagenic processes on the genomes of cells — has acquired considerable scientific prominence in recent years, particularly within the field of cancer genomics (1, 2). The identification of mutational signatures and the quantification of their activities in a genome offer valuable insight into the molecular processes that operate on the DNA of cancer and normal cells. These may include endogenous mutagenic processes, such as those arising from deficiencies in DNA repair pathways, and exposures to exogenous carcinogens, such as tobacco smoke (2).

Most analyses of mutational signatures so far have focused on patterns of mutations defined over a categorisation of single-nucleotide variants (SNVs). Specifically, 96 trinucleotide mutation categories are defined for SNVs based on base substitution type (6 categories) and trinucleotide sequence context (i.e. the bases immediately 5' and 3' of the mutated base; 16 categories). Base substitution types are collapsed so that only pyrimidine reference bases (cytosine or thymine) are considered. (For instance, guanine-to-adenine and cytosine-to-thymine mutations represent the same base change on opposite strands, so both can be expressed as cytosine-to-thymine, or C>T.) A vector of mutation counts or mutation probabilities across these 96 mutation categories is known as a mutational catalogue, or spectrum (1, 3).

A considerable number of software tools for mutational signature analysis have entered the scientific literature over the last five years (4). Perhaps most prominent among these is the Wellcome Trust Sanger Institute Mutational Signature Framework (5). This was the first published method to perform inference of mutational signatures *de novo* (also known as signature extraction), and implements a non-negative matrix factorization (NMF) algorithm that has been repeatedly applied to blind source separation problems in biology and other fields (6). An alternative to this NMF approach is the probabilistic model introduced by the EMu software (7), which implements an expectation–maximisation (EM) algorithm to perform maximum-likelihood estimation using a Poisson mutational model, and is able to account for differences in the opportunity for mutations of each class to occur in the sequence. Later tools have aimed to estimate the proportions in which a set of predefined mutational signatures are present in a collection of catalogues (a problem known as signature fitting). Notably, signature fitting does not share the considerable sample-size requirements of signature extraction methods, and is applicable even to individual mutational catalogues. However, recurrent drawbacks of algorithms for mutational signature analysis include data overfitting (overestimating the number of mutational processes), elevated computational cost, limited ranges of accepted mutation categories, poor integration with existing statistical programming frameworks and incomplete documentation.

Here we present sigfit, a powerful, flexible and user-friendly software package for performing mutational signature analysis in the R programming language. It provides both NMF- and EMu-inspired Bayesian statistical models for signature extraction and signature fitting, as well as novel statistical models that enable simultaneous fitting and extraction of signatures, aimed at detection of rare or previously undescribed mutational patterns. These models can be seamlessly applied to any kind of categorised mutation data, including short insertions and deletions (indels), dinucleotide variants, large structural variants and copy number alterations. The package also incorporates functions for production of publication-quality graphics of input and output data, and extensive user documentation featuring usage examples and test data sets.

Implementation

Overview

Developed as an open-source extension for the popular R programming language and environment, sigfit makes use of the Stan (8) statistical modelling platform and its interface package rstan (http://mc-stan.org/rstan) to carry out Bayesian inference on probabilistic mutational signature models. The package is publicly available on GitHub (https://github.com/kgori/sigfit), and can be installed directly from R using the devtools package (9). A detailed vignette is included in the package to illustrate its use.

Bayesian statistical modelling

Two types of statistical models of mutational signatures are implemented in sigfit. The first is equivalent to the mathematical formulation of the NMF approach that was introduced by the Wellcome Trust Sanger Institute Mutational Signature Framework (5) (hereafter referred to as WTSI-NMF) and has since been adopted by other tools (4). Such a formulation interprets the observed mutational catalogues as an approximate linear combination of a collection of mutational signatures, with the signature exposures representing the mixing proportions. An equivalent statistical formulation can be defined by modelling the mutational signatures as the probability parameters of independent multinomial distributions, and treating the observed mutational catalogues as draws from a mixture of these distributions, with the signature exposures serving as mixture weights. For a general case with G catalogues, N signatures and K mutation categories, the data and parameters in the model are as follows:

$S_{N\times K}$ is the matrix of mutational signatures,
$E_{G\times N}$ is the matrix of signature exposures,
$M_{G\times K}$ is the matrix of observed mutational catalogues,
$s_n = [s_{n1}, s_{n2}, \ldots, s_{nK}]$ are the mutation probabilities in signature n (n-th row of S),
$e_g = [e_{g1}, e_{g2}, \ldots, e_{gN}]$ are the signature exposures in catalogue g (g-th row of E),
$m_g = [m_{g1}, m_{g2}, \ldots, m_{gK}]$ are the observed mutation counts in catalogue g (g-th row of M).

In a signature fitting model, the signatures matrix S is fixed *a priori* and the exposures matrix E is inferred, whereas in a signature extraction model both S and E are inferred.

To illustrate the NMF-inspired approach to inference implemented in sigfit, we first describe a generative process producing mutations in a set of mutational catalogues (Fig. 1A). In this process, N conditionally independent mutational signatures are sampled from a Dirichlet distribution parameterised by α. For each catalogue, a set of exposures (which function as mixture weights) is sampled from a Dirichlet distribution parameterised by κ. For each of the I mutations occurring in this catalogue, an indicator value θ is drawn from a categorical distribution, parameterised by e_g, to specify which of the signatures produces the current mutation. Finally, the form that the mutation takes is drawn from a categorical distribution parameterised by the indicated signature, $s_{\theta_{gi}}$. This process is repeated for all I mutations in each of the G catalogues.

The inferential model is a simplification of the generative process. Since the assignment of individual mutations to specific signatures is not the aim of mutational signature analysis, considerable computational savings are achieved by marginalising out the assignment and only considering the sufficient statistics — the counts of each mutation category per catalogue — and modelling the likelihood of each catalogue m_g via a multinomial distribution, the parameters of which, P (with p_g being the g-th row of P), are calculated as the matrix product of the exposures, E, and the mutational signatures, S (Fig. 1B). The model is formalised as follows:

$$m_g \sim \text{Multinomial}(p_g)$$
$$\text{where } P = E\ S$$
$$s_n \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$$
$$e_g \sim \text{Dirichlet}(\kappa_1, \kappa_2, \ldots, \kappa_N)$$
$$g \in [1\,..\,G];\ n \in [1\,..\,N]$$

This multinomial model by default imposes uninformative Dirichlet prior distributions on signatures and exposures. The hyperparameters of the Dirichlet priors, α_k and κ_n, are all by default assigned a value of 0.5, corresponding to Jeffrey's transformation-invariant prior (10). However, users can make use of these parameters to specify custom prior distributions on signatures and exposures.

In addition to the NMF-inspired Dirichlet–Multinomial model above, sigfit also offers a Bayesian interpretation of the Poisson model implemented in the EMu software (7). This models the mutation counts in each category as draws from a Poisson distribution with expected value equal to the product of the mutational signatures, the degree of activity of each signature, and the opportunity for each mutation type across the relevant sequence (usually, the entire genome or exome). For the common case of 96 trinucleotide mutation types, such mutational opportunities correspond to the frequencies of the corresponding reference trinucleotides across the genome or exome. This allows the model to accommodate variability in the frequencies of trinucleotides among samples, such as that induced by copy number variation. For a case with with G catalogues, N signatures and K mutation categories, this model is formalised as follows (Fig. 1C):

$$m_{gk} \sim \text{Poisson}(p_{gk})$$
$$\text{where } p_{gk} = o_{gk} \textstyle\sum_{n=1}^{N} a_{gn} s_{nk}$$
$$a_{gn} \sim \text{Half-Cauchy}(0, 1)$$
$$s_n \sim \text{Dirichlet}(\alpha_1, \alpha_2, \ldots, \alpha_K)$$
$$g \in [1\,..\,G];\ n \in [1\,..\,N];\ k \in [1\,..\,K]$$

Here, m_{gk} is the observed number of mutations of type k in catalogue g, p_{gk} is the expected number of mutations of type k in catalogue g, s_{nk} is the mutation probability of mutation type k in signature n, a_{gn} is the activity of signature n in catalogue g (part of an activity matrix, $A_{G\times N}$), and o_{gk} is the mutational opportunity of mutation type k in catalogue g (part of an opportunity matrix, $O_{G\times K}$). The matrix of signature activities in this Poisson model, A, plays an analogous role to the exposures, E, in the previous model, with the dif-

ference that signature activities are not constrained to sum to unity (but must still be non-negative), and hence are assigned an uninformative half-Cauchy prior distribution.

One immediate advantage of this formulation is that the inferred mutational signatures are not contingent on the sequence composition of the genome under study (as this is captured by the mutational opportunities), and so the signatures are directly applicable to both genomes and exomes, and even across species (11), as long as the relevant mutational opportunities are available. Furthermore, sigfit provides a function ('convert_signatures') for converting mutational signatures between the opportunity-dependent Dirichlet–Multinomial formulation and the opportunity-independent Poisson formulation, allowing users to transition between models, or to adapt existing signatures to different sets of mutational opportunities.

Basic usage

sigfit accepts input data as a table of mutations, from which mutational catalogues can be derived using the 'build_catalogues' function, or directly as a matrix of catalogues. The mutation table must include at least four character fields: sample ID, bases of the reference and alternate alleles, and trinucleotide sequence context of each mutation. If mutations are located in transcribed genomic regions, an additional field can be used to indicate the transcriptional strand of each mutation. We note that the 'build_catalogues' function can currently process only mutation data defined over the common 96 trinucleotide mutation types; for data following different categorisations, a matrix of catalogues must be directly provided.

The main functionality of the package can be accessed via the functions 'fit_signatures' (signature fitting), 'extract_signatures' (signature extraction) and 'fit_extract_signatures' (simultaneous fitting and extraction with Fit-Ext models; see details below). These functions carry out Markov chain Monte Carlo (MCMC) sampling on the corresponding Bayesian models, described in the previous section. For signature fitting, a matrix of signatures must be provided, whereas for signature extraction, the number of signatures to extract must be specified. If a continuous range of signature numbers (e.g. 2–10) is specified for extraction, sigfit will extract signatures for each value in the range, and suggest the most likely true number of signatures. This is done heuristically, by finding the number of signatures which minimises the second derivative of the reconstruction accuracy function; the latter is defined, by default, as the cosine similarity between the original catalogues and the catalogues that have been reconstructed using the inferred signatures and exposures (two identical or parallel vectors have a cosine similarity of one, whereas two perpendicular vectors have a cosine similarity of zero).

sigfit provides functions to plot mutational catalogues and signatures ('plot_spectrum'), signature exposures ('plot_exposures') and spectrum reconstructions ('plot_reconstructions'). Moreover, all the relevant plots can be produced at once, directly from the output of the signature fitting and extraction functions, using the 'plot_all' function (Fig. 2A–C). The package also includes the set of 30 mutational signatures available in the COSMIC catalogue (https://cancer.sanger.ac.uk/cosmic), as well as several test data sets. The usage of the functions in sigfit is covered in detail in its associated vignette, which can be accessed through the 'browseVignettes' function, and in the R function documentation.

Combined mutational signature fitting and extraction

In addition to models for conventional extraction and fitting of mutational signatures, sigfit implements a novel formulation that allows fitting of predefined signatures and extraction of undefined signatures within a single Bayesian inference process. This formulation, which we have dubbed Fit-Ext, can be understood as a generalisation of the fitting and extraction models described above, where the signatures matrix S is composed of N signatures known *a priori* (modelled as data), and M additional undefined signatures (modelled as parameters). Through Dirichlet–Multinomial and Poisson formulations analogous to the ones introduced above, the Fit-Ext models perform fitting of the N fixed signatures and extraction of the M parametric signatures simultaneously. The definition of static signatures entails a substantial dimensionality reduction, thus enhancing statistical power for deconvolution of any additional rare signatures that may be present in the mutational catalogues.

We consider the Fit-Ext models to be of use in two particular scenarios. The first of these involves the case where a small sample size precludes signature extraction (i.e. only signature fitting is feasible), yet there is qualitative evidence of the presence of one or more novel signatures (not featured in COSMIC or other repositories) in the data. By fitting the set of known mutational signatures and extracting one or more additional signatures from the data, the novel signatures can be inferred even from single mutational catalogues. The second scenario involves the identification of very rare or weak signatures in large cohorts, or in studies of differential mutagenesis (for instance, between primary tumours and their metastases). In these cases, even if signature extraction is applicable, very rare signatures may not be amenable to conventional signature extraction due to insufficient statistical support. In this scenario, provided that there is evidence of the presence of infrequent mutational patterns which cannot be captured by the signature extraction models, the Fit-Ext models may be used to re-fit the previously inferred signatures and extract the rare signatures. However, we note that such approach is likely to be appropriate only if there is strong information about the presence and number of rare signatures; otherwise, the signatures inferred by the Fit-Ext models will likely be composed of uninformative mutational noise.

Extended mutational signature families

Recent genomic studies of cancer (12–16) have focused on mutational patterns beyond SNVs, defining mutational signatures over categories of indels, large structural variants and

copy number alterations. Whereas some software packages for mutational signature analysis are limited to specific sets of mutation categories (commonly, the 96 SNV categories described above), sigfit can inherently infer signatures over any set of features of interest. Furthermore, the package is already prepared for the inference of transcriptional-strand-wise SNV signatures, which distinguish between the transcribed and untranscribed strands for mutations in transcribed genomic regions: if an input mutation table is provided with an additional column containing strand information, sigfit will automatically define 192 strand-wise SNV categories, infer strand-wise signatures and adapt its plotting functions correspondingly (strand-wise test data sets are also provided). If a set of mutational catalogues is given with a number of mutation categories other than 96 or 192, sigfit will infer signatures using the categories provided. This flexibility makes the package suitable for identification of signatures over a range of unorthodox mutation categories, or even for entirely unrelated signal deconvolution problems.

Results

Comparison with existing methods

Using previously published mutational catalogues from 119 breast cancer genomes and 88 liver cancer genomes (3), we benchmarked sigfit against the software tools that originally introduced the two signature formulations implemented in the package: WTSI-NMF (5) and EMu (7) (Fig. 3A–E).

When comparing the Dirichlet–Multinomial and Poisson signature extraction models in sigfit to the original WTSI-NMF and EMu methods, we found that the approaches for estimating the most likely number of signatures in the latter two tools did not agree in any of the data sets. For the set of 119 breast cancer catalogues, sigfit (both models) and WTSI-NMF estimated the most likely number of signatures to be $N = 6$, whereas EMu estimated $N = 5$ as the best value. Conversely, for the set of 88 liver cancer catalogues, sigfit (both models) and EMu estimated the best number of signatures to be $N = 6$, whereas WTSI-NMF estimated it as $N = 10$. The contrast arises from the different approaches followed by each tool: while sigfit selects the number of signatures that minimises the second derivative of the overall cosine similarity between the original and reconstructed catalogues (essentially the 'elbow method', used in the clustering literature as a heuristic to estimate the number of clusters in a data set; see example in Fig. 2D), WTSI-NMF minimises the Frobenius reconstruction error while maximising signature reproducibility (5), and EMu uses the Bayesian information criterion (BIC) to correct for increased model complexity (7). The fact that sigfit consistently estimated the same optimal number of signatures for the Dirichlet–Multinomial and Poisson models, while inferring highly similar signatures and exposures to those obtained by the other methods (Fig. 3A–C), suggests that the criterion followed by sigfit to estimate the most plausible number of signatures may be more robust.

Notably, EMu estimates BIC values and selects the optimal number of signatures prior to signature extraction, whereas both WTSI-NMF and sigfit perform extraction for different numbers of signatures and measure the goodness of fit of each solution. Together with the fact that the expectation–maximisation algorithm is not guaranteed to converge to a global maximum of the likelihood (17), this can result in cases where EMu converges to a local, but not global, maximum for some values of N, resulting in incorrect estimation of the number of signatures that best explain the data. This is exemplified in the set of 119 breast cancer catalogues, where EMu converged to a locally optimal solution for $N = 6$ signatures (Fig. 3D), which, as mentioned above, led it to propose $N = 5$ as the best number of signatures. Interestingly, the difference between the solutions reported by sigfit and EMu for $N = 6$ is more prominent in terms of the inferred exposures than in the inferred signatures (Fig. 3A, C). By contrast, WTSI-NMF and both models in sigfit found virtually identical solutions for $N = 6$ signatures in this data set (Fig. 3A, C), which were considerably better than the one reported by EMu in terms of reconstruction accuracy (Fig. 3B). When applied to the set of 88 liver cancer catalogues, however, all the methods showed remarkably similar reconstruction accuracy distributions (Fig. 3B).

It is worth noting that, whereas EMu performs efficient maximum-likelihood estimation via EM (7), the WTSI-NMF method entails computing a considerable number of bootstrap replicates in order to identify stable clusters of mutational signatures (5). This causes WTSI-NMF to be very computationally expensive, and probably best suited for highly parallel computing infrastructures. By contrast, the models in sigfit, which exploit the efficient No-U-Turn-Sampler algorithm for MCMC sampling implemented in the Stan framework (8), incur only moderate memory and CPU demands that are easily met by laptop or desktop computers, while providing virtually identical solutions to those obtained by WTSI-NMF (Fig. 3A–C).

In addition to its favourable performance in signature extraction problems, when fitting the entire set of 30 COSMIC signatures to a previously published set of catalogues from 21 breast cancer genomes (18), sigfit reported only six significantly active signatures (Fig. 3E), five of which have been previously described to be active in breast cancer (3, 12). By contrast, optimisation-based methods such as least-squares regression and quadratic programming are notoriously prone to overfitting (19), especially when applied to non-convex functions like those involved in signature extraction and fitting problems. Hence, we believe that sigfit's capability to estimate realistic contributions of known signatures to sets of catalogues (or even to single catalogues), without the need to constrain the input set of candidate signatures to prevent overfitting, makes the package valuable for small-sized genomic studies with insufficient statistical power for signature extraction.

Fit-Ext model validation on simulated mutation data

As discussed above, the Fit-Ext models included in sigfit can be of use in situations where there is qualitative evidence of the presence of rare or previously undescribed signatures, yet lack of statistical power precludes conventional extrac-

tion of such signatures (or even signature extraction altogether). To demonstrate the effectiveness of the combined fitting–extraction approach in such situations — specifically, in the case where signature extraction is not possible — we repeatedly applied the Dirichlet–Multinomial Fit-Ext model to individual simulated mutational catalogues, which were drawn from a multinomial distribution defined as a random mix of three COSMIC signatures (signatures 1, 2 and 7) and one simulated novel signature (Fig. 4A). The novel signature was drawn from a Dirichlet distribution with very low concentration parameter values ($\alpha_1, ..., \alpha_{96} = 0.05$), and bore no resemblance to any signature in COSMIC (cosine similarity < 0.327). In each of 100 simulation experiments, sigfit was applied to such artificial catalogue in two ways. First, the Dirichlet–Multinomial signature fitting model was used to fit all 30 COSMIC signatures to the simulated catalogue; this resulted in inaccurate catalogue reconstruction (median cosine similarity of 0.887 between the original and reconstructed catalogues), as none of the COSMIC signatures captured the distinctive features of the simulated novel signature (Fig. 4B). Secondly, the Dirichlet–Multinomial Fit-Ext model was applied to fit all COSMIC signatures to the simulated catalogue, while simultaneously extracting one additional signature. Because a single catalogue was used in each simulation, the original and reconstructed catalogues were always identical, as any potential reconstruction error arising from random sampling was incorporated into the extracted signature (Fig. 4B). Despite this, the Fit-Ext results show that the simulated novel signature could be extracted with high accuracy in the majority of cases (median cosine similarity 0.948; Fig. 4C). This illustrates the model's potential in scenarios where available signatures do not entirely explain the mutational patterns found in the data.

Conclusions

Although the number of available tools for mutational signature analysis is growing rapidly (4), many existing methods exhibit recurrent conceptual or practical limitations, including excessive computational requirements, convergence to local optima, low user-friendliness and rigidity in the types of data and analyses supported. By exploiting the versatility of the R programming language and the robustness of the Bayesian inference machinery offered by the Stan (8) framework, sigfit provides new methods for flexible, simple and efficient analysis of mutational signatures. Furthermore, sigfit is, to the best of our knowledge, the first package to allow simultaneous fitting and extraction of mutational signatures, enhancing statistical power for the discovery of rare or novel signatures that cannot be deconvoluted using standard approaches. In addition, the popular R programming environment facilitates reproducibility and integration of data and results with different bioinformatic analysis workflows.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgements

The authors would like to thank Maximilian Stammnitz for testing the software, and Daniel Gaffney for helpful suggestions and bug reports.

References

1. Alexandrov, L.B. and Stratton, M.R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.*, 24:52–60, February 2014.
2. Petljak, M. and Alexandrov, L.B. Understanding mutagenesis through delineation of mutational signatures in human cancer. *Carcinogenesis*, 37(6):531–540, 2016.
3. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjörd, J.E., Foekens, J.A., Greaves, M. et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, August 2013.
4. Baez-Ortega, A. and Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Briefings in bioinformatics*, 2017.
5. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. and Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1): 246–259, 2013.
6. Devarajan, K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS computational biology*, 4(7):e1000029, 2008.
7. Fischer, A., Illingworth, C.J.R., Campbell, P.J. and Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.*, 14(4): R39, April 2013.
8. Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P. and Riddell, A. Stan: A probabilistic programming language. *J. Stat. Softw.*, 20, 2016.
9. Wickham, H., Hester, J. and Chang, W. *devtools: Tools to Make Developing R Packages Easier*, 2018. R package version 1.13.5.
10. Jeffreys, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 453–461, 1946.
11. Stammnitz, M.R., Coorens, T.H., Gori, K.C., Hayes, D., Fu, B., Wang, J., Martin-Herranz, D.E., Alexandrov, L.B., Baez-Ortega, A., Barthorpe, S. et al. The origins and vulnerabilities of two transmissible cancers in tasmanian devils. *Cancer cell*, 33(4):607–619, 2018.
12. Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47, 2016.
13. Waszak, S.M., Tiao, G., Zhu, B., Rausch, T., Muyas, F., Rodriguez-Martin, B., Rabionet, R., Yakneen, S., Escaramis, G., Li, Y., Saini, N., Roberts, S.A., Demidov, G.M., Pitkanen, E., Delaneau, O., Heredia-Genestar, J.M., Weischenfeldt, J., Shringarpure, S.S., Chen, J., Nakagawa, H. et al. Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. *bioRxiv*, 2017. doi: 10.1101/208330.
14. Li, Y., Roberts, N., Weischenfeldt, J., Wala, J.A., Shapira, O., Schumacher, S., Khurana, E., Korbel, J.O., Imielinski, M., Beroukhim, R. and Campbell, P. Patterns of structural variation in human cancer. *bioRxiv*, 2017. doi: 10.1101/181339.
15. Zou, X., Owusu, M., Harris, R., Jackson, S.P., Loizou, J.I. and Nik-Zainal, S. Validating the concept of mutational signatures with isogenic cell models. *Nature communications*, 9(1): 1744, 2018.
16. Macintyre, G., Goranova, T., De Silva, D., Ennis, D., Piskorz, A.M., Eldridge, M., Sie, D., Lewsley, L.A., Hanif, A., Wilson, C., Dowson, S., Glasspool, R.M., Lockley, M., Brockbank, E., Montes, A., Walther, A., Sundar, S., Edmondson, R., Hall, G.D., Clamp, A. et al. Copy-number signatures and mutational processes in ovarian carcinoma. *bioRxiv*, 2017. doi: 10.1101/174201.
17. Dempster, A.P., Laird, N.M. and Rubin, D.B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.
18. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K.W., Mudie, L.J., Varela, I. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, May 2012.
19. Harrell, F.E. Regression modeling strategies. *as implemented in R package 'rms' version*, 3(3), 2014.

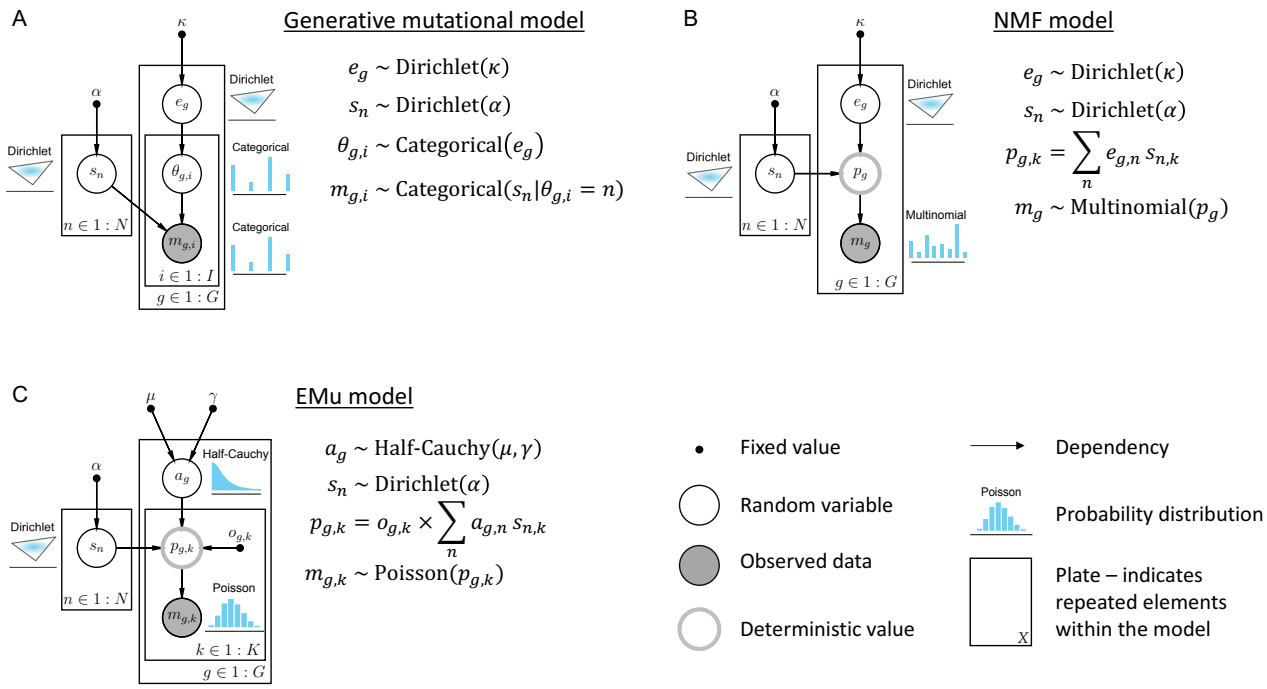

Figure 1. Bayesian model diagrams. (A) A generative model of mutation with multiple mutational processes. Each mutational catalogue (indexed by g) accumulates I mutations. Each mutation $m_{g,i}$ is produced by mutational process n of N processes. $\theta_{g,i}$ encodes the choice of mutational process for mutation $m_{g,i}$ in catalogue g. θ is assigned by a draw from a categorical distribution, parameterised by probabilities e_g, which are in turn drawn from a Dirichlet distribution parameterised by κ. The type of mutation is chosen by drawing from a categorical distribution, parameterised by probabilities s_n, in turn drawn from a Dirichlet distribution parameterised by α. **(B)** The NMF-inspired inferential model is based on the generative model, with the distinction that, as the allocation of individual mutations to specific processes is not of interest, the latent parameter θ is marginalised out. The process of marginalisation is to matrix-multiply the exposures (E, where e_g is the g-th row) and the signatures (S, where s_n is the n-th row), to yield a matrix of probabilities (P, where p_g is the g-th row). P is denoted by a node with a grey outline to indicate that it is a deterministic value, to distinguish it from random variable nodes, which take a black outline. The likelihood of the observed mutational catalogue m_g is obtained from the probability mass function of a multinomial distribution parameterised by p_g. **(C)** The EMu-inspired inferential model differs from the NMF-inspired model in that the mutation counts per mutation category, $m_{g,k}$, are modelled as if generated by a Poisson distribution, parameterised by the matrix product of the activities (A) and the signatures (S), element-wise multiplied by the opportunities matrix (O). A half-Cauchy distribution is used as the choice of prior for A, as the activities are not required to sum to unity. Plate notation diagrams were produced using daft-pgm (http://daft-pgm.org/)

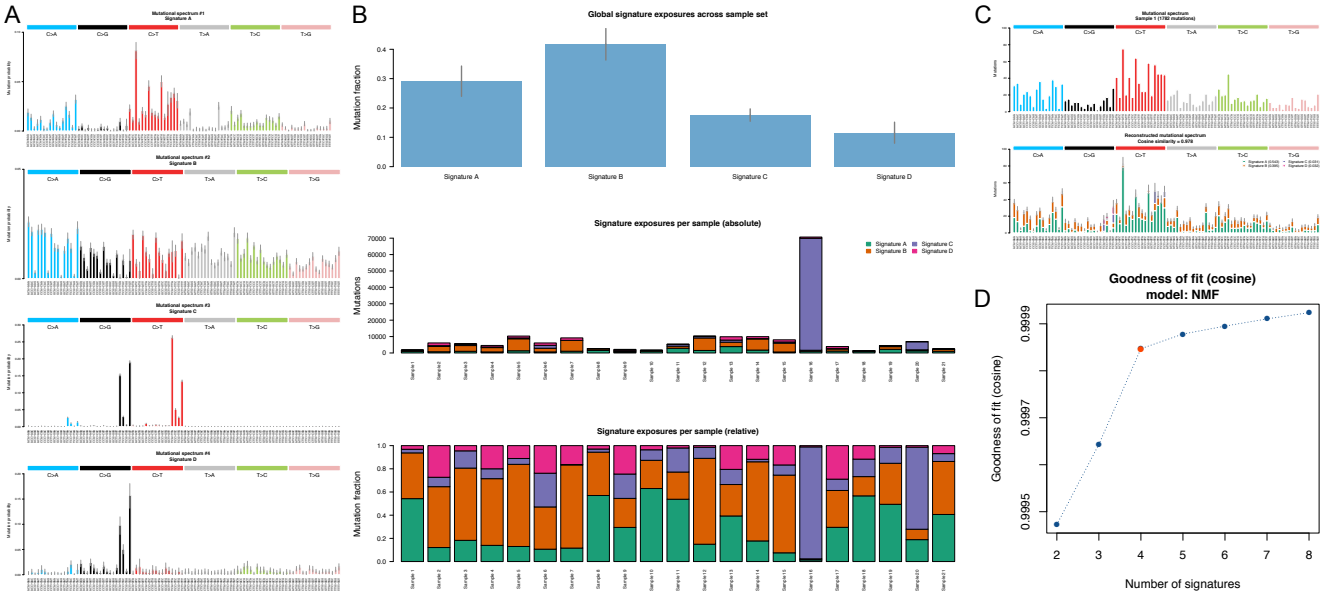

Figure 2. Plotting capabilities of sigfit. The plots in panels (A–C) were produced using the 'plot_all' function in sigfit, from the results obtained by extracting four signatures from a set of 21 breast cancer catalogues (18). **(A)** The four signatures extracted by sigfit; the most similar signatures in COSMIC are signatures 1, 3, 2 and 13, respectively. **(B)** Signature exposures across the set of catalogues as a whole (top), and within each catalogue, both as absolute mutation counts (middle) and in relation to the total number of mutations (bottom). **(C)** Plot of the reconstruction of an example mutational catalogue from the estimated signatures and exposures. The original catalogue is shown above the reconstructed one. **(D)** Evaluation of the reconstruction accuracy for a range of numbers of signatures, plotted using the 'plot_gof' function in sigfit. A value of four signatures is automatically suggested as the point maximising the approximated second derivative of the curve (indicated in orange).

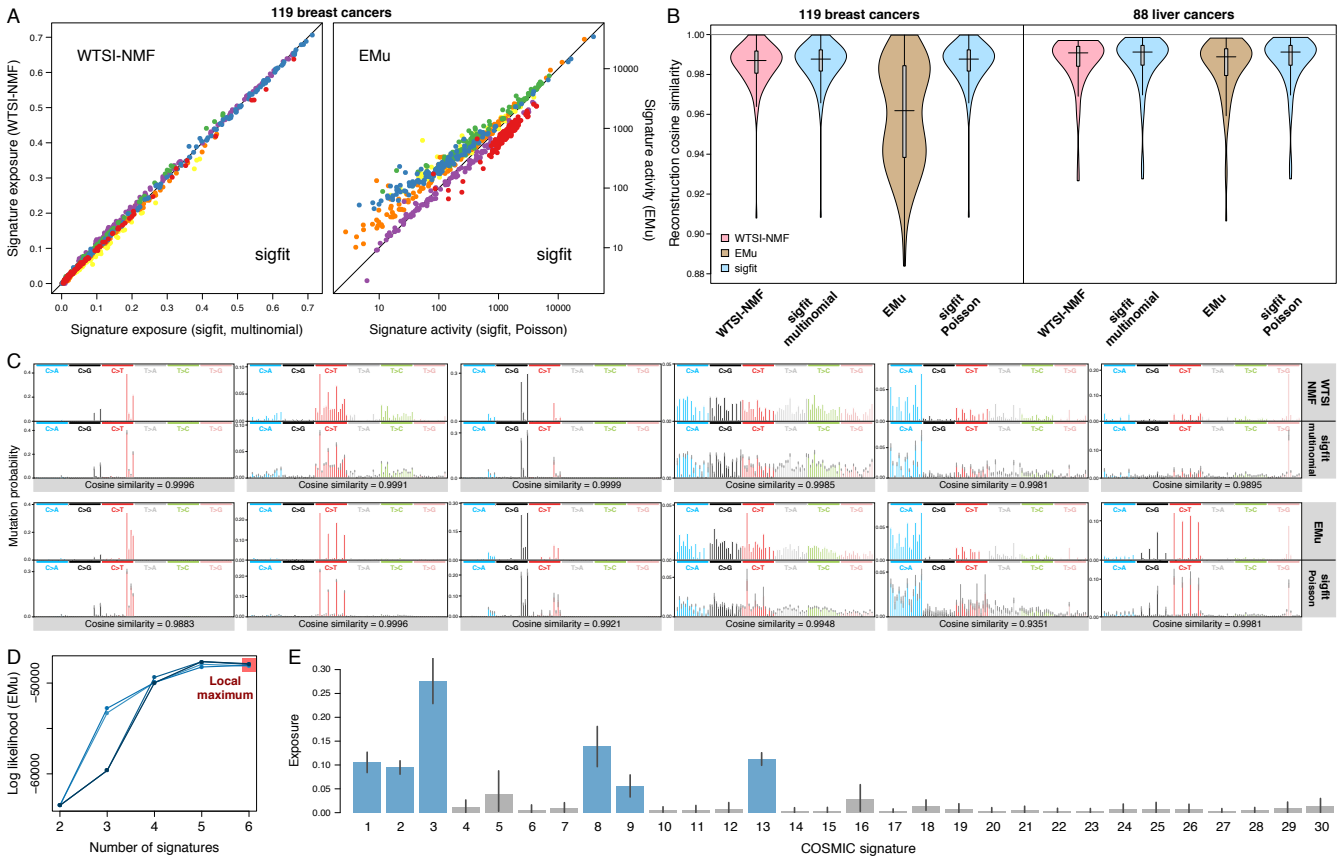

Figure 3. Evaluation of the performance of sigfit models for signature extraction and fitting. (A) Left, comparison of signature exposures (proportion of mutations attributed to each signature) as estimated by the Dirichlet–Multinomial model in sigfit (horizontal axis) and the WTSI-NMF software (vertical axis). Right, comparison of signature activities (number of mutations attributed to each signature) as estimated by the Poisson model in sigfit (horizontal axis) and the EMu software (vertical axis). The results correspond to extraction of six mutational signatures from a set of 119 breast cancer mutational catalogues (3). Dot colours distinguish between the six mutational signatures. **(B)** Distributions of reconstruction accuracy (cosine similarity between the original and reconstructed catalogues) for the models in sigfit and the WTSI-NMF and EMu software, in a set of 119 breast cancer catalogues (left) and a set of 88 liver cancer catalogues (right) (3). Boxplots are shown within each violin plot, and horizontal lines indicate median cosine similarity. Violin colours distinguish between software tools. **(C)** Comparison of mutational signatures extracted *de novo* by WTSI-NMF and the Dirichlet–Multinomial model in sigfit (top row), and by EMu and the Poisson model in sigfit (bottom row) from a set of 119 breast cancer catalogues (3). Cosine similarity between analogous pairs of signatures is shown below each. Vertical axes present mutation probabilities, and bars represent 96 trinucleotide mutation types (see main text), with colours indicating base substitution type. **(D)** Likelihood obtained by EMu over a range of numbers of signatures, for five different runs of the software. The decrease in likelihood when extracting six signatures (highlighted in red) indicates convergence of the EM algorithm to a local maximum. **(E)** Posterior mean estimates of global signature exposures inferred by the Dirichlet–Multinomial signature fitting model in sigfit, through fitting of the 30 signatures in COSMIC to a set of 21 breast cancer catalogues (18). Error bars indicate 95% highest posterior density (HPD) intervals. Grey bars indicate non-significant signature exposures, defined as exposures for which the lower end of the 95% HPD interval is below 0.01.

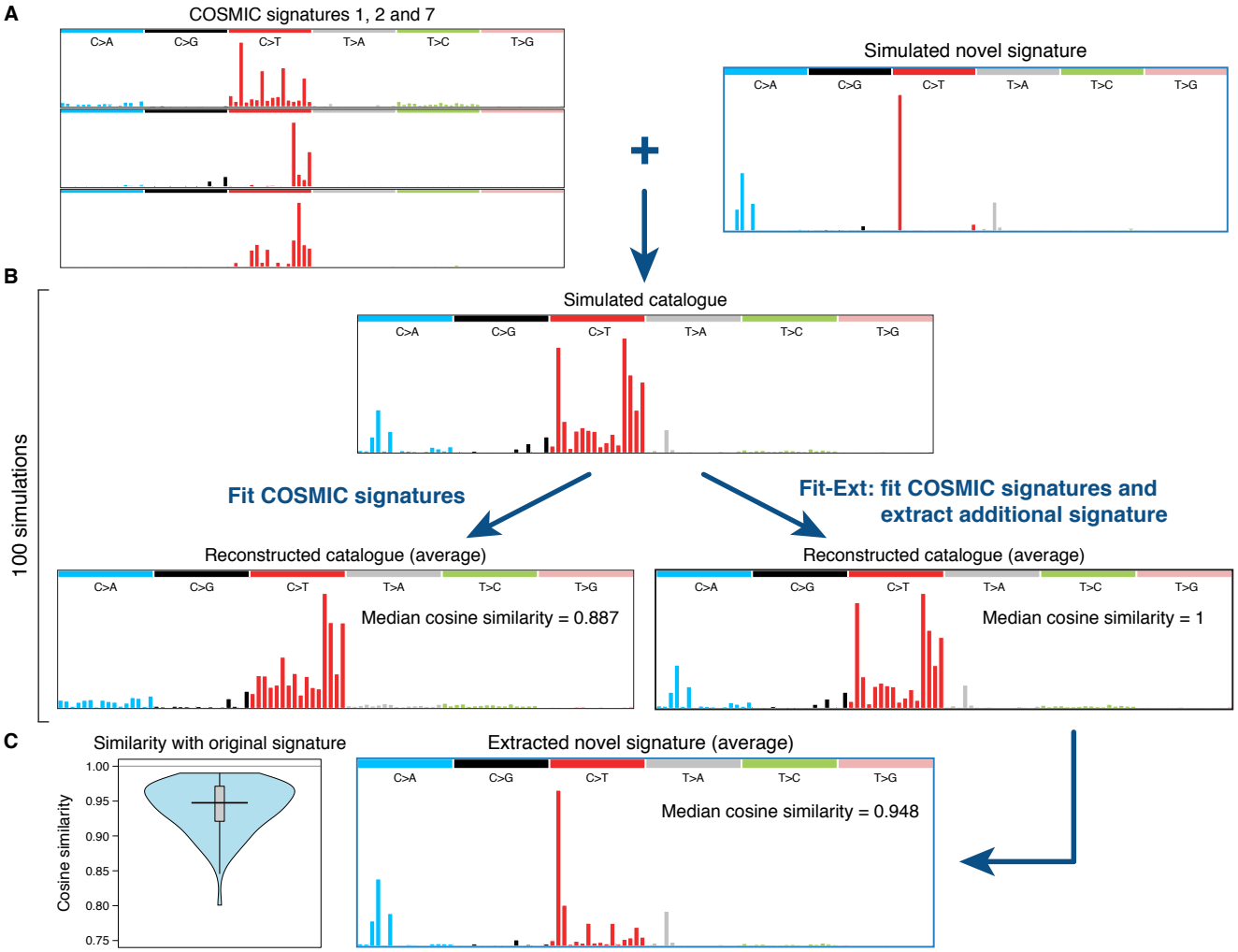

Figure 4. Evaluation of the Fit-Ext model on simulated mutation data. (A) Mutational spectra of the three COSMIC signatures (left) and the simulated novel mutational signature (right) that were used to produce simulated mutational catalogues. **(B)** In each of 100 simulations, a mutational catalogue (top) was generated from a random mixture of the signatures shown in (A), and signature analysis was performed with sigfit via two separate approaches: using the Dirichlet–Multinomial fitting model to fit all COSMIC signatures to the catalogue (bottom left), and using the Dirichlet–Multinomial Fit-Ext model to fit all COSMIC signatures and extract one additional signature (bottom right). Average reconstructed mutational catalogues (constructed from the product of signatures and inferred exposures) and median cosine similarities between the original and reconstructed catalogues across all simulations, are shown for each approach. **(C)** Average mutational spectrum of the estimated novel signature as extracted by the Fit-Ext model across all simulations. The violin plot (left) presents the distribution of the cosine similarity between the original novel signature, shown in (A), and the inferred signature across all simulations. A boxplot is shown within the violin plot, and the horizontal line indicates the median cosine similarity.