Bioinformatics, 34(2), 2018, 330–337
doi: 10.1093/bioinformatics/btx604
Advance Access Publication Date: 22 September 2017
Original paper

OXFORD

Genome analysis

Detecting presence of mutational signatures in cancer with confidence

Xiaoqing Huang†, Damian Wojtowicz*,† and Teresa M. Przytycka*

National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Christina Curtis

Received and revised on August 4, 2017; editorial decision on September 18, 2017; accepted on September 21, 2017

Abstract

Motivation: Cancers arise as the result of somatically acquired changes in the DNA of cancer cells. However, in addition to the mutations that confer a growth advantage, cancer genomes accumulate a large number of somatic mutations resulting from normal DNA damage and repair processes as well as carcinogenic exposures or cancer related aberrations of DNA maintenance machinery. These mutagenic processes often produce characteristic mutational patterns called mutational signatures. The decomposition of a cancer genome's mutation catalog into mutations consistent with such signatures can provide valuable information about cancer etiology. However, the results from different decomposition methods are not always consistent. Hence, one needs to be able to not only decompose a patient's mutational profile into signatures but also establish the accuracy of such decomposition.

Results: We proposed two complementary ways of measuring confidence and stability of decomposition results and applied them to analyze mutational signatures in breast cancer genomes. We identified both very stable and highly unstable signatures, as well as signatures that previously have not been associated with breast cancer. We also provided additional support for the novel signatures. Our results emphasize the importance of assessing the confidence and stability of inferred signature contributions.

Availability and implementation: All tools developed in this paper have been implemented in an R package, called SignatureEstimation, which is available from https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi\#signatureestimation.

Contact: wojtowda@ncbi.nlm.nih.gov or przytyck@ncbi.nlm.nih.gov

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

1 Introduction

Knowledge of elementary mutational processes underlying cancer cells is essential for understanding the etiology of cancer and its progression. The ever-growing amount of sequencing data allows for the analysis of cancer genomes not only from the perspective of highly mutated genes, but also from the perspective of broader mutational patterns. It is increasingly recognized that there is a whole spectrum of mutational processes that contribute to the mutation landscape of cancers. In addition to cancer driving mutations, cancer genomes harbor somatic mutations acquired during the normal cell cycle as well as those triggered by cancer related aberrations of DNA maintenance machinery such as mismatch repair, or by carcinogenic exposures such as tobacco smoking, ultraviolet light, replication stress. Each of these processes often leads to distinctive pattern of mutations—the so-called mutational signature. Computational methods developed to uncover such signatures from catalogs of somatic mutations (Alexandrov *et al.*, 2013a,b; Helleday *et al.*, 2014; Alexandrov and Stratton, 2014; Fischer *et al.*, 2013; Goncearenco *et al.*, 2017; Nik-Zainal *et al.*, 2012), including the classical nonnegative matrix factorization (*NMF*) approach, build

on the assumptions that the mutations observed in a cancer genome are a result of several mutational processes and various genomes might experience different exposure to each of the contributing mutagens. Previous analyses of cancer genomes from the perspective of mutational signatures have been very informative. In particular, the APOBEC mutational signature acts as a footprint of the activity of the APOBEC family of cytidine deaminases. Playing a key factor in many human cancers (Haradhvala *et al.*, 2016; Kanu *et al.*, 2016; Roberts *et al.*, 2013), APOBEC activity has been proposed to be the direct cause of some cancer driving mutations (Burns *et al.*, 2013; Henderson *et al.*, 2014; Kim *et al.*, 2017). As another example, a recent analysis of mutational signatures in the genomes of tobacco smoking-associated cancers provided a novel insight into how smoking increases cancer risk (Alexandrov *et al.*, 2016).

Given the growing interest in studies of mutational signatures in cancers and the need for a better understanding of the relation between detected signatures and biological causes, it is important to confidently associate known signatures to patients and assess patients' exposure to each of these signatures. As a step in this direction, Rosenthal *et al.* recently developed an approach, called *deconstructSigs* (*dS*), which determines a linear combination of the predefined signatures that best reconstructs the patient's mutational profile given a patient's catalog of mutations and a set of mutational signatures (Rosenthal *et al.*, 2016). However, this decomposition is not always in agreement with the signature contributions provided by the nonnegative matrix factorization approach. Importantly, while *deconstructSigs* is based on a heuristic, the decomposition problem can be solved using a quadratic programming (*QP*) approach (Goldfarb and Idnani, 1983) that rapidly converges to the optimal solution. We observed that the optimal decomposition is sometimes strikingly different from the decomposition provided by either of the two approaches in terms of signature presence or inferred contribution of signatures in tumor samples. This prompted us to address the confidence and stability of the decomposition problem. In particular, our approach allows us to assess if a cancer genome was exposed to a given set of mutational signatures and to quantify the confidence in the estimated contribution of each mutational signature.

There are two complementary aspects related to evaluating the confidence and stability of the decomposition of a mutational catalog into mutational signatures. The first is to measure decomposition variability and accuracy when the patient's mutational catalog is perturbed, as mutational catalogs might be noisy and incomplete. In addition, mutational processes act in a stochastic way. Given that quadratic programming can quickly compute optimal solutions for the original and the perturbed data, this question can be answered using a bootstrap analysis. The second and complementary perspective arises from the possibility that equivalent approximate solutions might exist. Specifically, different linear combinations of signatures might equally well approximate a mutational profile observed in a given patient. If quantitatively different suboptimal solutions exist in close proximity to the optimal solution, our confidence in the biological relevance of the optimal decomposition is reduced. To address this concern, we used a simulated annealing based method to randomly explore alternative decompositions whose error is close to the optimal one.

We applied both approaches to the whole-genome dataset of somatic mutations from 560 breast cancer (BRCA) patients (Morganella *et al.*, 2016; Nik-Zainal *et al.*, 2016). Identifying both very stable signatures (e.g. APOBEC related signatures 2 and 13) and highly unstable signatures (such as signatures 3, 5 and 8), we found that unstable signatures can be decomposed into other signatures with relatively small error, thereby explaining the lack of stability of these signatures. Next we re-analyzed the BRCA data and found signatures that have not been associated with breat cancer by previous analyses. In addition to statistical validation of these signatures, we evaluated their association with genomic features, which provides additional support for these signatures. Our results emphasize the importance of assessing the confidence and stability of inferred signature contributions for the interoperability of decomposition results.

2 Approach

Given the mutational catalog of a cancer genome (a set of somatic mutations), we strive to identify operative mutagenic processes and to quantify the genome's exposure to each of them. The imprint of a particular mutational process, referred to as its signature, is defined by the relative frequency of all types of nucleotide substitutions typically within the context of specific flanking residues. Here, we utilized mutational signatures that have already been discovered (Alexandrov *et al.*, 2013a). The goal is to approximate a genome mutational profile (i.e. observed mutation frequencies) as a linear combination of the signatures. Each signature coefficient of the linear combination is interpreted as the genome exposure (i.e. the fraction of mutations in the genome) to a mutagenic process represented by the respective signature.

Formally, let $\mathbf{M}$ $(K \times G)$ be a matrix containing observed mutational profiles from G samples (e.g. 560 breast cancer patients), where each profile contains the frequencies of K mutation types (e.g. 96 substitution types in the trinucleotide context) computed from each patient's catalog of mutations; and let $\mathbf{P}$ $(K \times N)$ be a matrix of N predefined mutational signatures (e.g. 30 COSMIC signatures) that specifies the probabilities of generating each of K mutation types by separate mutational processes. The objective is to find a nonnegative exposure matrix $\mathbf{E}$ $(N \times G)$ that contains, for each of G samples, the exposure to each of N signatures by minimizing a Frobenius norm (Alexandrov *et al.*, 2013 b,a; Nik-Zainal *et al.*, 2012): $\min_{\mathbf{E}} \|\mathbf{M} - \mathbf{PE}\|_F$; the value of the objective function represents the error of the inferred decomposition. For each patient g $(g = 1, \ldots, G)$, the objective function can be written as:

$$\min \sqrt{\sum_{k=1}^{K} (\mathbf{m}_{kg} - \sum_{n=1}^{N} (\mathbf{e}_{ng} \mathbf{p}_{kn}))} \qquad \text{s.t.} : \begin{cases} \sum_{n=1}^{N} \mathbf{e}_{ng} = 1 \\ \mathbf{e}_{ng} \geq 0. \end{cases}$$

There are many approaches to solve such a minimization problem. In this paper, we implemented two methods based on quadratic programming (*QP*) and simulated annealing (*SA*); for details see Supplementary Material. We used two publicly available R packages (https://cran.r-project.org): *quadprog* and *GenSA* for QP and SA, respectively. Both algorithms can find the optimal solutions. QP is extremely fast and stable, but it requires a predefined signature matrix P to be full column rank and relies on the problem formulation as a minimization of the Frobenius norm. SA can be widely used on a not-well-defined signature matrix, as well as a wider range of error measures, but it is slower than QP in converging to the optimal solution. Importantly, SA can also be used to randomly explore the landscape of suboptimal solutions that are close to the optimal decomposition.

3 Materials and methods

3.1 Confidence and stability of signature contributions

While the optimal method of quantifying the uncertainty attributed to noise in biological data requires the replication of measurements,

bootstrapping serves as a pragmatic alternative. To determine how estimates of signature contribution were distributed and assess their confidence and stability, we perturbed the original mutational catalog of each patient 1000 times by random re-sampling with replacement and estimated signature contributions in each bootstrap sample using the QP method. Based on the distribution of signature contributions, one can estimate bootstrap confidence intervals for each signature contribution and the empirical probability that a signature contribution is above a specific threshold. To assess the stability (bias) of contributions for different signatures, i.e. to test how much the contributions of bootstrap experiments vary from the original contributions, we computed the mean squared error (*MSE*) of the difference between the bootstrap estimates and the optimal contributions in the original data.

Apart of decomposition stability with respect to the input data, we tested the stability of the optimal solution itself to explore hidden dependencies between signatures. We sampled the space of suboptimal decompositions, i.e. approximate decompositions with slightly higher error than the optimal solution, using the simulated annealing approach—modifying the SA approach we used to find the optimal solutions by enforcing a stopping rule to report suboptimal solutions when a given value of the objective function (decomposition error) is reached. The decomposition error threshold was set to be higher than the error of the optimal solution by a factor of 1, 3 or 5%. The calculations were repeated 1000 times for each patient to randomly sample the space of suboptimal decompositions with each given threshold of decomposition error. The obtained distributions of signature contributions were then used, similarly to the bootstrap analysis, to assess the confidence and stability of signature contributions.

3.2 Genomic features of signatures

Morganella *et al.* (2016) showed that breast cancer mutational signatures exhibit distinct relationships with genomic features related to transcription, replication and chromatin organization. These analyses indicate that, in addition to the sequence context captured by the nucleotides flanking each mutation, the general genomic context also matters. In particular, if the mutational profile of a patient is shaped by several mutational processes, the mutations from the contributing processes are not shuffled randomly over the genome. Instead, mutations from each operating process are often clustered together. Maximal stretches of adjacent substitutions in a sample generated by the same mutational process on the same reference allele are called processive groups. Our analysis of breast cancer data (see Results) has identified previously missed signatures, thus, in addition to statistical analysis, we tested these signatures for expected genomic features, following procedures of Morganella *et al.* as summarized below (for details see Morganella *et al.*, 2016).

Each base substitution was associated with the mutational signature with the highest *a posteriori* probability of generating this mutation; the *a posteriori* probability was computed based on signature exposure levels inferred by QP method. Mutations (in the pyrimidine context) within protein coding genes (Ensembl release 60) were classified based on whether they are located on the transcribed/non-coding strand or the non-transcribed/coding strand. Transcriptional strand bias was computed for each signature separately as a ratio of the number of mutations on the transcribed strand to the total number of mutations in all samples. For processive groups, we counted the number of groups of different lengths (the number of maximal successive mutations) for each signature separately. To assess the significance of the observed group counts, we compared them with the numbers of processive groups in 100 randomized datasets, where the order of mutations was shuffled with respect to the original data.

3.3 Datasets

The whole-genome dataset of somatic mutations from 560 breast cancer (BRCA) patients (Morganella *et al.*, 2016; Nik-Zainal *et al.*, 2016) was downloaded from the ICGC Data Portal (release 23), https://dcc.icgc.org. We classified all somatic substitutions into 96 mutation types in the trinucleotide context (6 substitutions from a pyrimidine base pair times 4×4 nucleotide types at both 5′ and 3′ sides of substitution). For each patient, we computed its mutational profile from the patient's catalog of mutations, i.e. the number of mutations of each type. Each patient's mutational profile was normalized by the total number of mutations that each patient possessed.

The patterns of 30 known and validated mutational signatures were retrieved from the COSMIC website (release 80), http://cancer.sanger.ac.uk. This set of signatures was deciphered from dozens distinct types of human cancer, although not all signatures are present in every cancer genome (Alexandrov *et al.*, 2013a,b; Alexandrov and Stratton, 2014; Helleday *et al.*, 2014; Nik-Zainal *et al.*, 2012, 2016). The most recent analysis of the breast cancer whole-genome sequences by Nik-Zainal *et al.* (Nik-Zainal *et al.*, 2016) revealed 12 signatures found in breast cancer patients—signatures 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26 and 30 in the set of 30 COSMIC signatures—and provided exposures of each breast cancer patient to each signature as estimated by the NMF-based Mutational Signatures Framework (Alexandrov *et al.*, 2013a). We refer to this decomposition as *NMF decomposition*. We also applied the deconstructSigs approach to this dataset using the 12 known breast cancer signatures to infer their exposure contributions (Rosenthal *et al.*, 2016); no signatures were discarded to minimize decomposition error. We refer to these estimates as *dS decomposition*.

4 Results

4.1 Differences in signature exposures inferred by different decomposition methods

The most commonly used tool for discovering mutational signatures from the catalog of mutations, the Mutational Signatures Framework (Alexandrov *et al.*, 2013a) based on NMF technique, additionally provides estimates of the number of mutations generated by each discovered signature. The newly developed tool—deconstructSigs (dS)—can be used to decompose a small tumor sample or a patient's mutation profile into already known signatures (Rosenthal *et al.*, 2016). As both methods are based on heuristics, they do not always show consistent levels of signature exposures. In addition, the decomposition problem into known signatures can be solved optimally using existing optimization techniques based on quadratic programming or simulated annealing (see Approach section) to provide the theoretically optimal decomposition solution. We ran all four methods and compared them on the largest available whole-genome dataset containing somatic mutations from 560 breast cancer patients (Nik-Zainal *et al.*, 2016) (Fig. 1).

Results across all methods are often very similar; for example the methods show similar contributions of all signatures for patient PD24196 in Figure 1A. However, in many cases their results are significantly different (see patients PD8609 and PD13608 in Fig. 1A), where contributions of some signatures vary greatly between methods. QP and SA show very consistent results. NMF identifies the

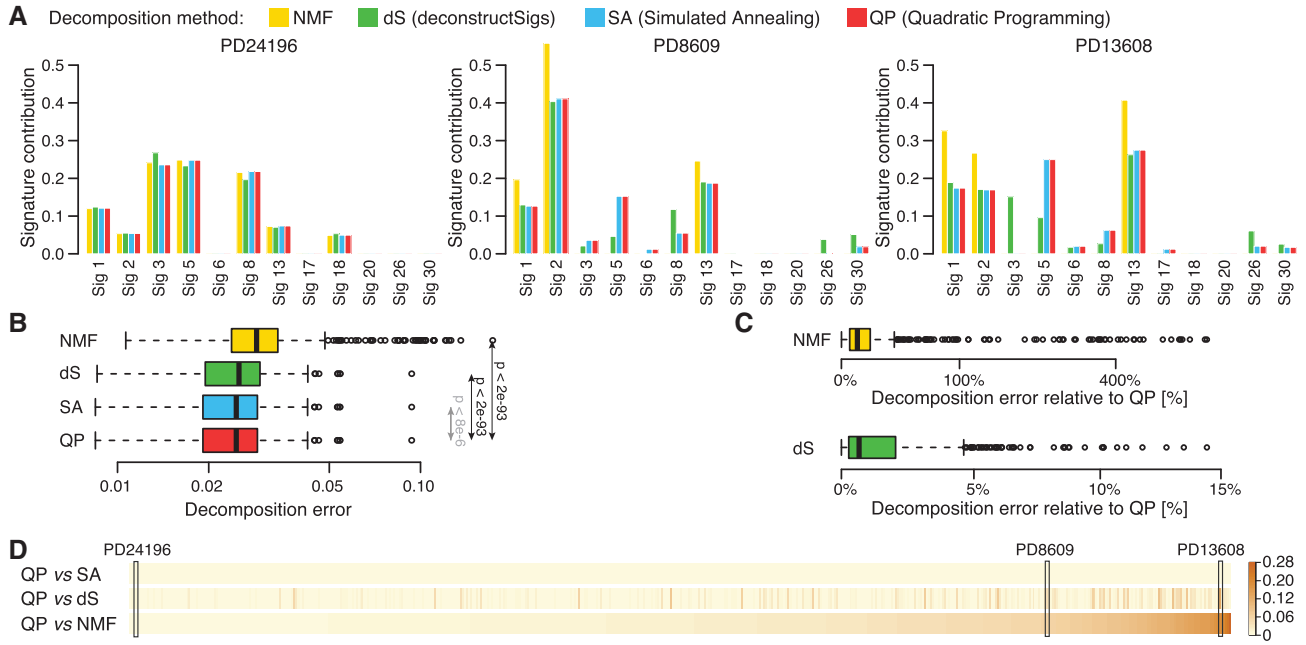

Fig. 1. Comparison of four decomposition methods. The decomposition of patient's mutational profile into 12 mutational signatures known to be present in breast cancers was inferred using four different methods for each of 560 patients. (**A**) Three examples comparing four methods shown as barplots. (**B**) Distribution of decomposition errors (the root sum-squared error between observed and inferred mutational profile) compared between four method across 560 breast cancer patients. Statistical significance of difference between decomposition errors of QP and three other methods is shown (paired Wilcoxon test); Cohen's d effect size of difference between decomposition errors of QP and SA is negligible. (**C**) Decomposition errors for NMF and dS relative to QP errors shown in the log scale (*y* axis). (**D**) Comparison of cosine distance between signature contributions inferred using QP and three other methods shown for all patients. Patients were sorted based on their distance between QP and NMF solutions. Three patients selected as the examples in (**A**) are marked

strongest mutational signatures (1, 2 and 13) in each sample and ignores signatures with little contribution. Therefore the exposures of strong signatures are over-estimated as compared to other methods. Despite this, signature 5, estimated by QP and SA to have a high contribution (about 0.15 and 0.25 in both examples), is missed. dS usually shows similar results to QP and SA, but for some signatures it differs significantly. For example, in patient PD13608, signature 3 is detected by dS with a high contribution of 0.15 although other methods yield an absence of this signature. Signature 5 is detected only at 0.1 even though QP and SA show a strong contribution of 0.25. The decomposition errors (values of the objective function) across all 560 patients for the four methods are shown in Figure 1B. Although QP finds slightly better solutions than SA in terms of optimal decomposition error in the majority of patients (75%), the Cohen's d effect size of the difference between two methods is negligible so we can assume that they perform equally well. Moreover, QP and SA always have lower decomposition errors than that of the two other methods across all 560 patients, and significantly outperform them in terms of both *P*-value (<2e-93; paired Wilcoxon test) and Cohen's d effect size (>0.4). In addition, we used simulated data to test if QP and SA recover the true composition of the sample (this information is not available in the real data). Indeed, we found that these methods had excellent performance and outperformed the dS method (see Supplementary Material and Supplementary Fig. S1). NMF has, on average, the largest decomposition error, which is much higher (by up to 750% and mean of 40%) relative to QP and SA as shown in Figure 1C. However, the main objective of NMF is to discover unknown signatures from the patient's catalogs of mutations, not to find the optimal decomposition into predefined signatures. The method focuses on the most prominent signatures in each patient, so the lower performance in terms of optimal error is not surprising. While it is not striking in Figure 1B, the decomposition error of dS is higher by up to 15% (mean 1.6%) relative to QP and SA (Fig. 1C). As it is shown in two examples in Figure 1A, even such small differences in decomposition error can lead to significant discrepancy in inferred contributions or even overall signature presence. To show the extent of differences between methods, we computed the pairwise cosine distance between exposures inferred by QP versus NMF, dS and SA for each individual patient (Fig. 1D). There are a number of patients for which the distance between solutions is large; this is not specific to particular patients, but rather to compared decomposition methods.

The above results show that for some patients the signature exposures inferred by different methods differ significantly, even when there is no striking difference in decomposition errors. Moreover, it appears that some signatures (e.g. 3 and 5) are more prone to variability than other signatures (e.g. 1, 2 and 13). Even though the method based on QP shows the best performance in terms of smallest decomposition error and runs much faster than other methods, it is still not clear how stable the optimal solution is and which signatures are credible. To answer these questions we analyzed the stability of the optimal solution in terms of variability in both input data and suboptimal solutions. This will help us to understand the origin of the observed discrepancies between the methods.

4.2 Confidence and stability analysis of signature contributions

The observed mutation frequencies in real biological samples may be contaminated by noise and the optimal solutions inferred based on such data do not always have meaningful and direct biological interpretation. In order to measure the confidence in the estimation of the

exposure intensities, we applied the bootstrap technique. Thus, we perturbed the original mutational catalog of each patient separately by randomized re-sampling with replacement and estimated signature contributions in each bootstrap sample using QP method. Figure 2A shows the distribution of exposure estimates in perturbed input catalogs for the same three examples shown in Figure 1A. The bootstrap estimates are, as expected, distributed around the optimal QP solutions in the original data, and their median values are reasonably close to the original contributions. However, the variability of the bootstrap solutions for some signatures and patients vary significantly from the original QP exposures. This variability seems to be related to specific signatures and independed of patients and exposure levels. For example, the contributions of signature 3 differ between the example patients in Figure 2A from 0 to over 0.2, but the variability of

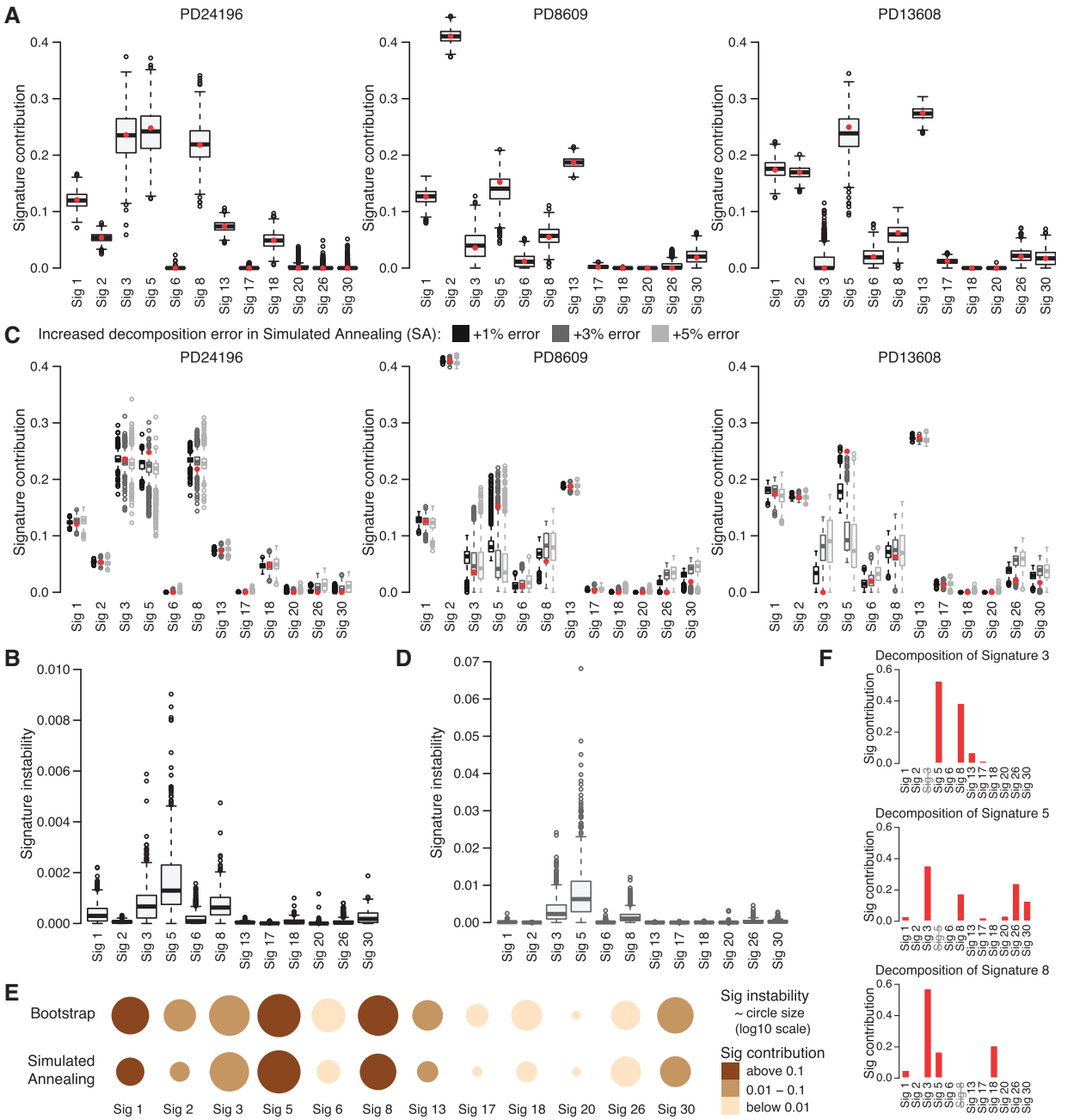

Fig. 2. Stability of signature contributions. (**A**) Distribution of signature contributions in 1000 bootstrap samples for the three selected patients (same as in Fig. 1A). Contributions in the original samples inferred by QP are shown as red dots. (**B**) Comparison of signature contribution instability between all signatures measured as the mean squared error between contributions in 1000 bootstrap samples and contributions in the original samples for each patient. Corresponding results for suboptimal solutions based on simulated annealing trials with increased decomposition error are shown in (**C**) and (**D**), respectively. Three error threshold increases (+1, +3 and +5%) relative to the optimal error from QP methods are compared (**C**), but only results of '+3% error' are further presented. (**E**) Summary of stability analysis based on bootstrap (top) and simulated annealing (bottom). The size of each circle represents median signature contribution instability (log10 scale) over all patients for each signature separately; the sizes were normalized by the most unstable signature. The color of each circle indicates median signature contribution. (**F**) Decomposition of a mutational signature (3, 5 and 8) into 11 remaining signatures using QP method (decomposition errors: 0.05, 0.03 and 0.06, respectively)

bootstrap solutions seems to be high in all three cases. On the other hand, the variability of signature 13 contributions in bootstrap samples is small across all examples and also independent of exposure levels. To assess the signature stability more broadly we investigated the divergence of their bootstrap contributions from the optimal QP solution in terms of the mean squared error for each patient separately (Fig. 2B). Signatures 3, 5 and 8 show large instability in the majority of patients; moderate instability can also be observed for signatures 1 and 30. On the contrary, signatures 2 and 13 are stable across all patients. Some of the other signatures (e.g. signature 17) seems to be stable as well, but they are rather infrequent signatures with median contribution below 0.01.

The COSMIC signatures were derived from a large but limited dataset containing different cancer types using a heuristic method based on NMF, so although the signatures are linearly independent, they should not be assumed to be 'orthogonal': weak dependencies between them might exist. As such, different linear combinations of signatures can lead to equivalent approximations of the optimal decomposition, i.e. approximations with the same decomposition error in close proximity of the optimal solution. Then, our confidence in the applicability of the optimal decomposition will be weakened. To check whether this problem applies to all signatures or only some of them, we randomly sampled the solution space of suboptimal decompositions for all patients using simulated annealing and stopping the simulations whenever the decomposition error was close to the error of optimal QP decomposition. We performed three runs of simulations with increasing thresholds of error for suboptimal solutions—optimal error times 1.01, 1.03 and 1.05, respectively. The results for our three exemplary patients are presented in Figure 2C. The summary of signature stability analysis in terms of MSE for 3% error increase is presented in Figure 2D, as an example since results for all three error levels are consistent. And again, we can observe that contributions of some signatures in suboptimal decompositions, like signatures 2 and 13, are close to the contributions inferred by QP method and that they are stable independently of patient, exposure level and increased error level. The contributions of signatures 1 and 30 seem to be more stable than in bootstrap analysis. However, the contributions of signatures 3, 5 and 8 in suboptimal decompositions still vary substantially across all patients and diverge from the optimal QP exposures (see for example patients PD8609 and PD13608). This divergence increases with the increased error of suboptimal decompositions and can quickly lead to over- or underestimation of contributions of some signatures like for example in patient PD13608 for signatures 3 and 5, respectively. This explains why the contributions of signatures 3 and 5 inferred by dS for this patient differed significantly from the contributions inferred by QP (Fig. 1A); dS found a suboptimal solution with decomposition error higher than in the optimal solution found by QP.

Presented applications of both bootstrap and simulated annealing approaches provide different but complementary views on the stability of contributions of mutational signatures inferred from patient's mutational profile. Contributions of signatures 3, 5 and 8 are unstable from both perspectives (Fig. 2E) and their variability seems to be interrelated, especially in simulated annealing analysis (Fig. 2C). To check whether any signature can be easily replaced by a combination of other signatures, we decomposed each signatures into the remaining signatures using the QP approach (i.e. treating a signature as a patient's mutational profile and decomposing it using the set of the remaining signatures). The only signatures that can be decomposed with relatively small error (within decomposition errors of real BRCA patients, <0.1) are signatures 3, 5 and 8 (Fig. 2F). Each of these signatures requires considerable presence of the two other signatures and limited contribution from some of the remaining signatures. This helps to explain the exposure instability of these three signatures. Moreover, both analyses show that contributions of signatures 2 and 13, both related to the activity of AID/APOBEC, are stable across all patients and that they cannot be replaced by a combination of other signatures (not shown). This suggests that these two signatures are distinctively defined and well separated from other signatures. Similar properties apply, to some extent, to signatures 1 and 30, which offers insight into their reduced instability in simulated annealing relative to that of the bootstrap analysis.

Table 1. The number of breast cancer patients exposed to each of 12 mutational signatures (columns) at different minimal contribution levels (rows) with *P*-value of 0.01 as assessed by bootstrap analysis of the perturbed patient's mutational catalogs

Exposure	Signatures											
threshold	1	2	3	5	6	8	13	17	18	20	26	30
>0.01	512	366	220	505	16	479	402	30	107	5	25	115
>0.05	411	173	187	459	7	374	218	8	41	4	11	20
>0.10	342	110	164	368	6	225	118	4	16	3	8	2
>0.15	258	71	144	301	5	121	78	3	6	1	7	1
>0.20	159	50	128	219	4	65	60	2	3	1	7	1
>0.25	75	35	118	144	1	27	45	1	1	1	5	1

In both analyses, we ran 1000 simulations and computed relevant decompositions for each patient (e.g. three patients in Fig. 2A and C). Based on these computations, one can estimate confidence intervals for each signature contribution in each patient to better assess which signatures are present in a patient's mutational profile and what their contribution levels are. Alternatively, one can ask what is the probability that a patient's exposure to a signature is above a certain level. We counted the number of patients with different exposure levels of all BRCA signatures (Table 1). As we can observe, a huge majority of patients (above 85%) is exposed to signatures 1, 5 and 8 at a minimal exposure level of at least 0.01 (with *P*-value 0.01), and these signatures are relatively abundant in a number of patients. Other signatures such as 6, 17, 20 and 26 are only present in a limited number of patients, and only few patients are exposed to these signatures at levels above 0.15. Such analysis assesses which signatures are present in a single sample at a minimal exposure level, so further analysis could be focused on essential signatures only and signatures with little contribution could be disregarded from analysis of a particular patient.

4.3 Re-analysis of mutational signatures presence in breast cancer

Nik-Zainal *et al.* (2016) analyzed whole-genome sequences of 560 breast cancers and extracted 12 base substitution mutational signatures from patient's catalogs of somatic mutations using their own framework based on NMF (Alexandrov *et al.*, 2013a). The method focuses only on the essential mutational signatures that contribute large numbers of mutations to the inferred mutational profile of each sample. Using the tools we presented in our paper, we reanalyzed the data using the full set of 30 COSMIC mutational signatures and determined which signatures, beyond the 12 previously identified signatures, are present in breast cancers with high confidence. Figure 3A shows the distribution of sample exposures to 30 signatures over all patients as inferred by QP method. Signatures 9, 12 and 16, formerly detected in other cancer types, are likely to be

present in a number of patients as their contribution levels are comparable to, or even exceed, the contributions of some of the 12 signatures. The results of bootstrap simulations confirmed these observations, see Table 2; and simulated annealing analysis of suboptimal solutions showed consistent results (not shown). As we can observe, each of these novel signatures is present in at least a few patients with high contribution above 0.15 and empirical *P*-value of 0.01; the same is true for the 12 known breast cancer signatures. The remaining signatures do not show the same presence in breast cancer patients, so they were excluded from further analysis. We

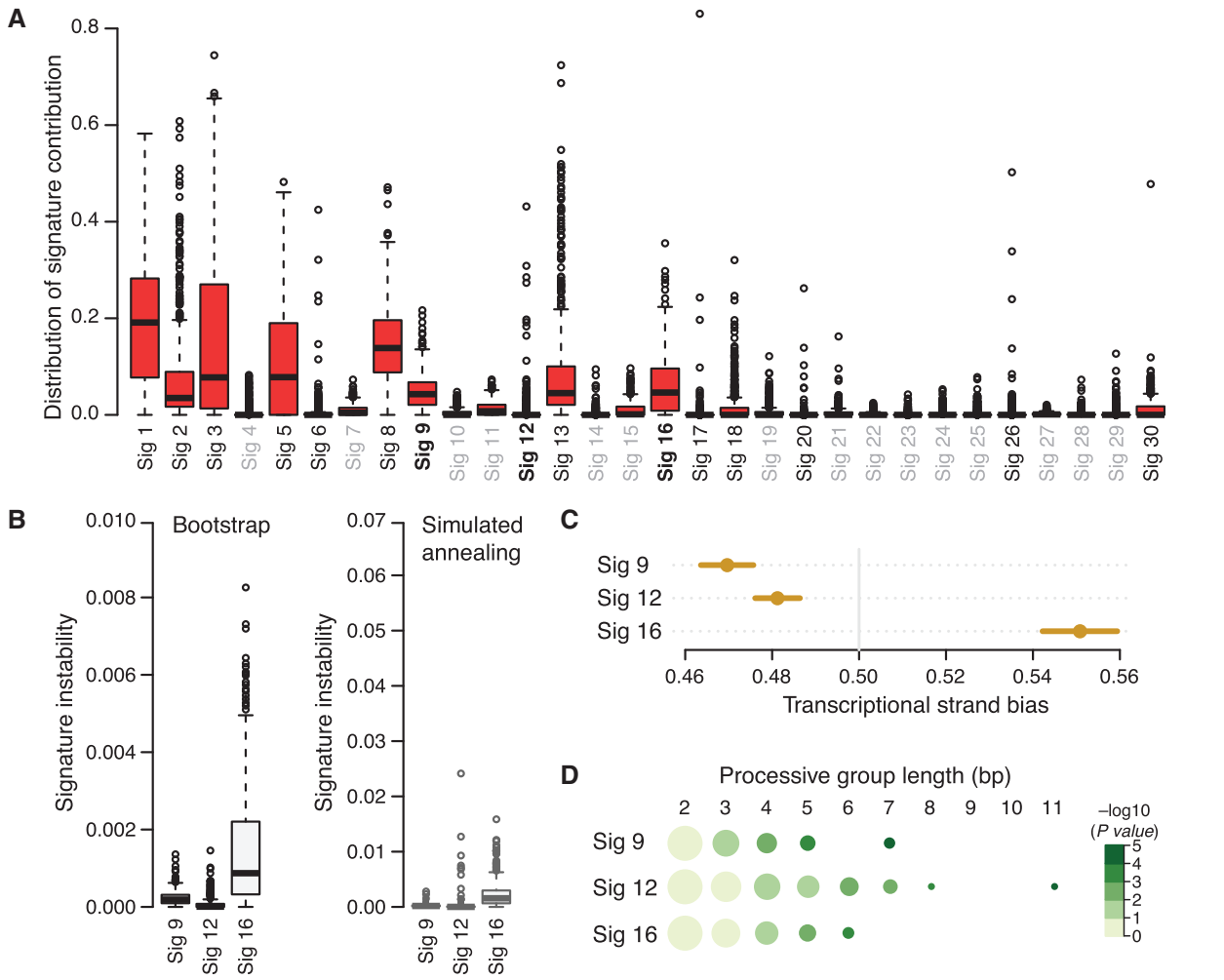

Fig. 3. (**A**) Distribution of contributions of 30 COSMIC signatures over 560 breast cancer samples inferred using QP method. Three signatures (9, 12 and 16) suggested to be novel in breast cancers are marked bold; signatures not present in breast cancer are gray. The signatures 9, 12 and 16 were evaluated in terms of their contribution stability in the bootstrap and simulated annealing analyses (**B**) and genomic features known to be exhibited by some signatures such as the transcriptional strand bias (**C**) and length of processive groups (**D**). Observed transcriptional strand bias is shown as a circle with 95% confidence intervals against expected bias of 0.5. Processive groups of different lengths (columns) for each signature (rows) are represented as circles whose size corresponds to the number of groups (log10 scale) and color to the *P*-value of detecting a processive group of a defined length (-log10 scale). The numbers of groups were normalized by the number of groups of length 2 (Color version of this figure is available at *Bioinformatics* online.)

Table 2. The number of breast cancer patients exposed to each of 30 COSMIC mutational signatures (columns) at different minimal contribution levels (rows) with *P*-value of 0.01 as assessed by bootstrap analysis of the perturbed patient's mutational catalogs

Exposure	Signatures																												
threshold	*1*	*2*	*3*	4	*5*	*6*	7	*8*	**9**	10	11	**12**	*13*	14	15	**16**	*17*	*18*	19	*20*	21	22	23	24	*25*	*26*	27	28	*29*	*30*
>0.01	523	335	250	7	91	8	10	434	246	13	18	15	386	5	12	117	9	66	3	5	6	0	0	0	8	7	0	4	2	8
>0.05	423	165	210	1	62	6	0	312	65	0	0	9	210	2	1	32	4	25	0	3	3	0	0	0	0	5	0	1	0	1
>0.10	347	104	175	0	33	6	0	174	12	0	0	8	117	0	0	8	3	9	0	2	1	0	0	0	0	4	0	0	0	1
>0.15	265	69	154	0	12	4	0	91	2	0	0	7	78	0	0	2	3	5	0	1	0	0	0	0	0	3	0	0	0	1
>0.20	177	49	136	0	4	2	0	36	0	0	0	4	61	0	0	1	2	1	0	1	0	0	0	0	0	2	0	0	0	1
>0.25	96	35	120	0	1	2	0	16	0	0	0	3	45	0	0	0	1	1	0	0	0	0	0	0	0	2	0	0	0	1

Note: The 12 known breast cancer signatures are marked italic and the three signatures (9, 12 and 16) suggested to also be present in breast cancers—bold.

recomputed the data decomposition into 15 signatures—12 known plus 3 novel, for all patients.

To further evaluate the presence of novel signatures 9, 12 and 16 in breast cancers we assessed their stability and tested if these signatures exhibit known transcriptional features previously shown for some of the 12 signatures by Morganella et al. (Morganella *et al.*, 2016). The novel signatures show similar ranges of stability as the 12 signatures, from stable signature 12 to unstable signature 16 (Fig. 3B; compare with Fig. 2B and D). The transcriptional strand bias of the novel signatures is highly significant (P-value < e-12) and much stronger than for any other signature (Fig. 3C). Signatures 9 and 12 show bias towards the non-transcribed strand, while signature 16 shows extremely strong bias towards the transcribed strand, mostly with T > C mutations at ATN context. The bias of signatures 12 and 16 was formerly observed in other cancer types, while the bias of signature 9 was not observed before and seems to be associated with the function of DNA polymerase η, which plays an essential role in replicating damaged DNA. Some mutational processes cause long stretches of successive mutations, called processive groups, to occur on the same DNA strand. We found significantly long processive groups containing at least 5-6 mutations associated with the novel signatures, especially for signature 12, which had the longest processive group of 11 substitutions (Fig. 3D). These results shows that the novel signatures, 9, 12 and 16, carry genomic features previously observed for breast cancer signatures or signatures present in other cancer types and independently support the outcome of our tools for decomposition of patient's mutational profile into predefined mutation signatures with confidence.

5 Conclusion

With the steadily decreasing cost of sequencing, we can now catalog cancer somatic mutations for an ever increasing number of individual patients. It is important to be able to use this information optimally. In particular, key information that we can obtain from such mutation data includes the mutagenic processes shaping the mutation catalog of each individual patient. Such information can point to specific defects in the replication mechanism, enzymatic activities, etc. Until now methods that decomposed patient's mutational catalog into mutations associated with specific signatures made no attempt to assess confidence in such decomposition. We showed that the results provided by existing methods can differ significantly not only in estimating exposure levels to each mutagenic process, but also in assessing whether or not a given signature is actually present. In particular, we provided statistical evidence and biological support for the presence of three additional signatures in breast cancers. However, many of the differences are not merely a reflection of differences in the accuracy of decomposition methods, but are, at least in part, related to the inherent instability of some signatures.

In this work, we used two complementary approaches to assess the confidence and stability of the resulting decomposition. Our analysis showed that some mutagenic signatures, such as the signatures related to APOBEC activity, are very stable while others, especially noisy are signatures 3, 5 and 8, are not. This instability can be explained, in part, by the fact that these signatures can be decomposed into a linear combination of other signatures with a very small error. This problem can only deepen with the increasing number of novel mutational signatures that are being discovered through the steadily increasing number of cancer datasets. We note that some of the 30 COSMIC signatures are already highly correlated (Supplementary Fig. S2). Importantly, various cancer related events such as differences in methylation, chromatin status, or similar events might influence inferred mutational signatures even when the mutagenic processes remain unchanged. Thus the analysis of mutational signatures can potentially provide a wealth of additional information while at the same time can lead a set of signatures that are relatively similar. Our results emphasize the importance of analyzing the confidence and stability of inferred signature contributions from the perspective of input data perturbation and approximate suboptimal solutions; the evaluation methods and software developed for this study can aid such analyses.

Funding

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of Interest: none declared.

References

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

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

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

Alexandrov,L.B. *et al.* (2016) Mutational signatures associated with tobacco smoking in human cancer. *Science*, **354**, 618–622.

Burns,M.B. *et al.* (2013) Apobec3b is an enzymatic source of mutation in breast cancer. *Nature*, **494**, 366–370.

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

Goldfarb,D. and Idnani,A. (1983) A numerically stable dual method for solving strictly convex quadratic programs. *Math. Program.*, **27**, 1–33.

Goncearenco,A. *et al.* (2017) Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Res.*, **45**, W514–W522.

Haradhvala,N.J. *et al.* (2016) Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell*, **164**, 538–549.

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

Henderson,S. *et al.* (2014) Apobec-mediated cytosine deamination links pik3ca helical domain mutations to human papillomavirus-driven tumor development. *Cell Rep.*, **7**, 1833–1841.

Kanu,N. *et al.* (2016) Dna replication stress mediates apobec3 family mutagenesis in breast cancer. *Genome Biol.*, **17**, 185.

Kim,Y.A. *et al.* (2017) Wesme: uncovering mutual exclusivity of cancer drivers and beyond. *Bioinformatics*, **33**, 814–821.

Morganella,S. *et al.* (2016) The topography of mutational processes in breast cancer genomes. *Nat. Commun.*, **7**, 11383.

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

Nik-Zainal,S. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.

Roberts,S.A. *et al.* (2013) An apobec cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.*, **45**, 970–976.

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