JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 29, Number 1, 2022
Mary Ann Liebert, Inc.
Pp. 56–73
DOI: 10.1089/cmb.2021.0453

ScalpelSig Designs Targeted Genomic Panels from Data to Detect Activity of Mutational Signatures

NICHOLAS FRANZESE,[1–3,i] JASON FAN,[1,ii] RODED SHARAN,[4] and MARK D.M. LEISERSON[1,iii]

ABSTRACT

Over the past decade, a promising line of cancer research has utilized machine learning to mine statistical patterns of mutations in cancer genomes for information. Recent work shows that these statistical patterns, commonly referred to as "mutational signatures," have diverse therapeutic potential as biomarkers for cancer therapies. However, translating this potential into reality is hindered by limited access to sequencing in the clinic. Almost all methods for mutational signature analysis (MSA) rely on whole genome or whole exome sequencing data, while sequencing in the clinic is typically limited to small gene panels. To improve clinical access to MSA, we considered the question of whether targeted panels could be designed for the purpose of mutational signature detection. Here we present ScalpelSig, to our knowledge the first algorithm that automatically designs genomic panels optimized for detection of a given mutational signature. The algorithm learns from data to identify genome regions that are particularly indicative of signature activity. Using a cohort of breast cancer genomes as training data, we show that ScalpelSig panels substantially improve accuracy of signature detection compared to baselines. We find that some ScalpelSig panels even approach the performance of whole exome sequencing, which observes over 10 × as much genomic material. We test our algorithm under a variety of conditions, showing that its performance generalizes to another dataset of breast cancers, to smaller panel sizes, and to lesser amounts of training data.

Keywords: cancer genomics, combinatorial optimization, mutational signatures.

1. INTRODUCTION

Over the past few decades, research has revealed that cancer is a disease characterized by the accumulation of mutations (Stratton et al., 2009; Garraway and Lander, 2013; Kandoth et al., 2013; Gerstung et al., 2020). Over the lifetime of an organism, cells acquire mutations at random positions in the genome. Certain mutations disrupt the function of cellular systems, and if certain cellular systems are

[1]Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA.
[2]Department of Computer Science, Northwestern University, Evanston, Illinois, USA.
[3]National Institutes of Health, Bethesda, Maryland, USA.
[4]School of Computer Science, Tel Aviv University, Tel Aviv, Israel.
[i]ORCID ID (https://orcid.org/0000-0001-9773-8017).
[ii]ORCID ID (https://orcid.org/0000-0001-7617-4814).
[iii]ORCID ID (https://orcid.org/0000-0002-1034-4363).

disrupted simultaneously, a cell can lose its ability to regulate its rate of reproduction. Dysregulation of the reproductive cycle is one of a handful of ''hallmark'' qualities (Hanahan and Weinberg, 2011), which together transform a healthy cell into a cancer cell. These hallmark qualities are shared by all cancers, but they can be caused by random mutations in many respective genes. Furthermore, many such genes are mutated only at low frequency, implying that there are numerous combinations of mutated genes that can lead to cancer [e.g., see the analysis by Lawrence et al. (2014)]. The randomness involved in determining a cell's path to the disease state is responsible for part of the difficulty of treating cancer (Zugazagoitia et al., 2016). Since cancer genomes vary so widely, the set of therapeutic targets does also, necessitating a diverse set of treatment strategies and engendering difficult decisions about which one to use for a given patient.

One silver lining to this situation is that the random process by which mutations are acquired in the genome is not uniform—it is patterned (Helleday et al., 2014). For example, mutations acquired by smoking tend to produce a different pattern than those acquired from UV radiation (Nik-Zainal et al., 2015) or those acquired endogenously through errors made during genome replication (Tubbs and Nussenzweig, 2017), etc. Recent work demonstrates that identifying which of these mutational processes are active in a tumor genome provides a useful basis with which to categorize the diverse landscape of tumor phenotypes (Hoeck et al., 2019). Indeed, such activity can serve as an effective biomarker in the clinic. Past work has shown that certain mutational processes can indicate a tumor's vulnerability to particular therapies. One prominent example is mismatch repair deficiency, which can indicate the effectiveness of checkpoint inhibitor immunotherapy (Le et al., 2017). A second is homologous recombination repair (HR) deficiency, which can indicate effectiveness of poly ADP ribose polymerase (PARP) inhibitor therapy (Farmer et al., 2005; Lord and Ashworth, 2017).

Designing and applying methods for the detection of mutational processes is an ongoing effort in the computational cancer biology community. The most well known approach was pioneered in 2012 by Nik-Zainal et al. (2012), who utilized machine learning methods to extract the ''signature'' patterns of several mutational processes from aggregated tumor genome samples. Alexandrov et al. (2013b) formalize a *mutational signature* as a probability distribution over a set of mutation categories—that is, they suppose that a given mutational process can be identified by the frequency with which it causes each type of mutation. Their landmark article utilized distributions over 96 mutation categories, defined by all possible single base substitutions (SBS) with trinucleotide context. To infer the signatures from the data, they formulated a simple linear model of a tumor's mutations. In their model, a relatively small number of mutation signatures are shared across tumors, and the mutations of each tumor are given as a linear combination of these shared signatures plus some noise.

To realize this model, they apply non-negative matrix factorization (NMF) to genome-wide mutation counts on a large cohort of tumor samples. When applied to this problem, NMF infers a set of mutational signatures, as well as the activity of each signature on each genome in the cohort (often called the *exposure*). While other learning algorithms and schemes for categorizing mutations have been used, this approach remains the most common method of signature extraction in mutational signature analysis (MSA) studies (Alexandrov et al., 2013a, 2020; Kim et al., 2016). The current state of the art for SBS mutational signatures identified 49 signatures, extracted from an analysis of 23,829 tumors. Recent work has highlighted mutational signatures as a powerful diagnostic tool for clinical use, with several signatures implicated as potential therapeutic biomarkers (Davies et al., 2017; Van Hoeck et al., 2019).

While these results are indicative of great promise for the future of cancer treatment, current clinical treatments are largely unable to take advantage of these advances. This is primarily due to the outsized sequencing requirements of the computational methods used in MSA studies (Alexandrov et al., 2013b, 2020) relative to current clinical sequencing practices (Frampton et al., 2013; Cheng et al., 2015). The ''gold standard'' data source for MSA studies is whole-genome sequencing (WGS). This is because WGS gives a complete and unbiased picture of the mutations present in a tumor genome, which affords increased accuracy during signature extraction. When WGS data are not available or in short supply, MSA studies also commonly use whole-exome sequencing (WES) as a data source. WES takes a subset of the genome that is smaller but still sizable—it is thus cheaper to sequence but still provides utility for signature extraction.

Almost all MSA studies use WGS or WES to obtain mutation counts in tumor genomes. By contrast, WGS and WES are *unavailable* to most cancer patients—current clinical sequencing practices are commonly limited to *targeted sequencing*. This term refers to precisely-aimed therapeutic assays that sequence small pieces of the genome with known biological importance. This can provide valuable information about particular genes of interest, but provides a much more limited picture of the distribution of mutations on the

genome which MSA seeks to assess. Concretely, targeted sequencing assays identify $1000\times$ fewer mutations compared to WGS (Nik-Zainal et al., 2020) and $100\times$ fewer mutations than WES. Furthermore, while there have been calls to make large-scale sequencing assays clinically available, ramping up production of sequencing data in the clinic will take time and resources. Despite the decreasing costs of sequencing itself, providing large-scale genomic sequencing to patients requires infrastructure for analysis, interpretation, and storage of the resulting data. Thus, WGS in particular is unlikely to be offered as a routine clinical diagnostic for many years to come (Nik-Zainal et al., 2020). In the meantime, the therapeutic impact of MSA in the clinic is tied to the important open problem of inferring signature activity from targeted sequencing assays.

A few recent studies have designed methods to detect mutational signature activity from clinically accessible targeted sequencing assays. Campbell et al. (2017) found that in patients with hypermutation, panel data could be used to infer exposures using standard methods (Rosenthal et al., 2016). Gulhan et al. (2019) introduced SigMA to detect signature activity indicative of homologous recombination repair deficiency, with a specific focus and application to gene panel data. Sason et al. (2020) introduced Mix to infer mutational signatures and their exposures in clusters of patients for use on datasets with few mutations per patient. In general, these studies make methodological modifications to account for the limited sample of the genome afforded by targeted sequencing assays. Even still, these methods are constrained by *which* regions of the genome are sampled. In particular, these works have sought to detect signature activity from two existing targeted sequencing assays: the Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) (Cheng et al., 2015) and FoundationOne Panel Sequencing (Frampton et al., 2013).

These panels are designed to identify specific *actionable* mutations, which are most commonly found in genes linked with cancer. This goal is fundamentally different from the goal of MSA, which seeks to analyze the *distribution* of mutations over the genome. Accordingly, such panels likely provide a biased view of the true mutation distribution. This is evidenced by a recent prominent study which showed that mutations in cancer genes (which are common locations of actionable mutations) are subject to powerful selective pressure due to their impact on the fitness of a tumor. The authors showed that this can result in distributions of mutations that are not representative of the underlying mutational signatures (Temko et al., 2018). Thus it is likely that other regions of the genome are more suitable for interrogating the genome-wide distribution of mutations.

The problem of designing new targeted sequencing assays tailored to the goals of MSA has been explored in a very limited capacity. One previous study contained a short analysis to identify a panel-sized set of genes, which could be used to detect one specific mutational signature (Polak et al., 2017). This analysis did not consider noncoding regions as candidates for the panel and generally was not the main focus of the article. Perner et al. (2020) recently introduced mutREAD, an assay for detecting mutational signature activity, but their approach concerns the method by which the DNA is sequenced rather than panel construction. To our knowledge, no existing study tackles the generalized problem of identifying a panel-sized set of genome regions that are suitable for the detection of signature activity. We take aim at this problem in the present study.

1.1. Contributions

In this study, we present ScalpelSig, an algorithm that learns from data to design *genomic panels* optimized for the detection of activity from an arbitrary mutational signature.[1] We use the term genomic panel to refer to a set of genome regions (including both coding and noncoding regions, in contrast to a *gene* panel) whose sequence can be used as a biomarker. We constructed ScalpelSig with the explicit goal of increasing clinical access to MSA, and as such all panels considered in the present study are small enough to be sequenced and analyzed using current clinical infrastructure. To our knowledge, the ScalpelSig method includes two novel contributions to the problem of panel design for mutational signature detection. First, ScalpelSig considers the whole genome (not just coding regions), and second it considers arbitrary mutational signatures rather than just those of homologous recombination repair.

We train ScalpelSig on a large cohort of breast cancer whole genome sequences and evaluate its performance on held-out data. Performance is evaluated by extracting signature exposures from the

[1]A preliminary version of this article was published in the proceedings of RECOMB 2021 [Franzese et al., (awaiting publication)].

discovered panel regions with standard methods and comparing the signature activity within panel regions to the ground truth of genome-wide signature activity. This performance is compared against two benchmarks: the MSK-IMPACT panel and a randomized baseline. Panels designed by ScalpelSig afford superior accuracy for signature detection in five out of six examined signatures. We additionally analyze the generalizability and robustness of our algorithm. We find that ScalpelSig's increased accuracy over baselines is maintained on an independent breast cancer dataset and for a wide variety of parameterizations.

2. PRELIMINARIES

For the purposes of this work, we shall refer to a mutational signature (or *signature* for ease of exposition) as is seminally defined by Nik-Zainal et al. (2012) and Alexandrov et al. (2013a,b). A signature is a multinomial distribution over a set of *mutation categories*. The most commonly studied mutation categories are SBS and their immediate 5′ and 3′ flanking bases (Alexandrov et al., 2013b), leading to 96 categories arising from six canonical SBS categories C>G, C>A, C>T, T>A, T>C, and T>G, and four possible 5′ and 3′ flanking bases. For example, the A[C>T]G mutational category refers to a C>T SBS immediately flanked by an A (5′) and a G (3′).

Each tumor or *sample* (belonging to a patient), whose genome is sequenced, can be summarily described by a 96-dimensional count vector of observed SBS belonging to each mutational category. MSA assumes that each observed mutation is emitted by a unique mutational signature, and it is generally assumed that only a few signatures are *active* in a tumor and in a cancer type. Briefly, MSA usually involves (1) deriving the distributions of the latent mutational signatures active in a cohort of samples; and/or, (2) determining the *exposure* each sample has to each signature—the proportion of mutations in each sample emitted by each signature.

In this work, we consider problems relating to the latter, more clinically relevant problem where a candidate set of potentially active mutational signatures in each sample is given and known. We perform analysis with respect to a canonical and widely adopted set of signatures determined by the Catalogue of Somatic Mutations in Cancer (COSMIC) ver. 2, and will always refer to (and index) COSMIC signatures by their numerical names (e.g., Signature 1) (Tate et al., 2019).

3. METHODS

A panel optimized for detecting signature activity should consist of a small set of genomic regions which, when sequenced, allow the accurate identification of *genome-wide* patterns of mutations. Initially, one may ask whether this is a feasible goal. For intuition, we refer to previous work which shows a strong relationship between cancer type and regional mutation density (Jiao et al., 2020), as well as other regionally varying chromatin features (Polak et al., 2015). Furthermore, many mutational signatures are tied to molecular mechanisms which vary in their activity over the genome (Haradhvala et al., 2016; Morganella et al., 2016). These findings suggest that some mutational processes may prefer to cause mutations in *specific* regions of the genome. Therefore, some regions of the genome may be far more informative than others for assaying the activity of these processes and their associated signatures.

Finding these informative regions poses a significant challenge, in part, due to the technical attributes of NMF, the most widely-used algorithm for signature extraction (Alexandrov et al., 2013a; Kim et al., 2016). Notably, NMF depends only on genome-wide mutation counts as input. Thus, NMF determines only the *counts* of mutations attributed to each signature and is agnostic to the *location* of mutations in the genome. This presents a barrier for understanding the regional distribution of signature activity.

Furthermore, NMF is computationally intensive, and its resource requirements are amplified within the present use case. Broadly, the outputs of NMF lack robust structure: solutions are nonunique, the loss surface is nonconvex, and several random initializations of the algorithm or specialized initialization approaches (Boutsidis and Gallopoulos, 2008) are required to obtain a reliable result. Thus, it is computationally intractable to utilize NMF to compute signature activity and explore the combinatorial space of possible panels.

One way to respond to these challenges is by simplifying the problem. Instead of finding regions that assay the activity of all signatures at once, we break the problem down into distinct binary classification tasks. We reason that in clinical applications, signature activity serves primarily as a biomarker to decide whether a patient should receive a particular tailored treatment. In this use case, it is more important for a panel to detect if one specific signature has *substantial* activity on a given genome, rather than an estimate of the absolute number of mutations (or exposure) attributed to a collection of signatures. This idea guides the construction of our algorithm.

We introduce ScalpelSig (Scalar projection Panels for mutational Signatures), a method for discovering genomic panels to detect mutational signature activity. Given a mutational signature, ScalpelSig designs a genomic panel optimized to distinguish samples where said signature is (substantially) *active* from those where it is *inactive*. ScalpelSig designs such a panel by selecting a set of discerning *windows* over the genome. The ScalpelSig algorithm can be briefly described in three steps.

1. ScalpelSig divides the genome into nonoverlapping windows of a fixed size.
2. ScalpelSig computes, with a *window scoring function*, a heuristic measure of mutational signature activity in each window across a given cohort of samples.
3. ScalpelSig combines the highest scoring windows to generate and output a genomic panel of a given fixed size.

3.1. ScalpelSig: scalar projection panels for mutational signatures

To efficiently optimize over the combinatorial space of possible panels, ScalpelSig uses *scalar projection* as a heuristic measure of mutational signature activity. Unlike typical methods for detecting signature activity, scalar projection has a closed form algebraic expression and can be computed deterministically in constant time. It also has robust structural guarantees—in particular it is a linear transformation. These properties make optimization tractable, and we show that scalar projection is indeed a reasonable measure of signature activity. For vectors u and v, we write the scalar projection of u onto v as: $\text{proj}_v(u) = \frac{1}{||v||}\langle v, u\rangle$.

Notably, the scalar projection of mutation category counts onto a mutational signature gives the magnitude of the least-loss signature activity vector (i.e., the vector that results from scaling a mutation signature by its exposure), if said signature was the only active signature in the sample. We formalize this result in Lemma 1.

Lemma 1. *Given any mutation count vector c and a signature vector q, the magnitude of the least-loss exposure vector in the direction of q is given by the scalar projection of c onto q, written as* $\text{proj}_q(c) = \frac{1}{||q||}\langle q, c\rangle$.

Proof. We seek the exposure a such that the residual between c and aq (the exposure *vector*) is minimized. Writing $||\ \ ||$ to mean the 2-norm, we wish to solve $\min_a ||c - aq||$, or equivalently:

$$\min_a ||c - aq||^2. \tag{1}$$

Writing Equation (1) as an inner product, yields:

$$||c - aq||^2 = \langle c - aq, c - aq\rangle = a^2||q||^2 - 2a\langle q, c\rangle + ||c||^2. \tag{2}$$

Equation (2) is quadratic in a. Thus, we solve for a by taking the derivative with respect to a and setting it to be 0, which gives $0 = 2a||q||^2 - 2\langle q, c\rangle$. Rearranging, we find the solution, $a = \frac{\langle q, c\rangle}{||q||^2}$. Notice that a is indeed the least-loss exposure and is non-negative since q and c are non-negative.

The exposure *vector* aq is given precisely by the projection of c onto q. Writing the unit vector in the direction of q as $\hat{q} = \frac{q}{||q||}$, we get,

$$aq = \frac{\langle q, c\rangle}{||q||^2}q = \frac{\langle q, c\rangle}{||q||}\frac{q}{||q||} = \text{proj}_q(c)\hat{q}. \tag{3}$$

Thus, the *magnitude*, $\text{proj}_q(c)$, of the least-loss exposure *vector*, aq, relates to the *value* of the least-loss exposure, a, by a constant factor, $\frac{1}{||q||}$, with $a = \frac{1}{||q||}\text{proj}_q(c)$. □

Given a mutational signature q, a training set S of samples with subset A that has signature q active, and a maximum panel size N, ScalpelSig designs and outputs a panel, $\mathbf{P}$, that distinguishes samples with substantial signature activity from those without. ScalpelSig designs such a panel by selecting an optimally-scoring subset from the set of all nonoverlapping genome windows of a given fixed width.

Let a mutational signature q be a vector representing a multinomial distribution over the standard 96 mutation categories described in Alexandrov et al. (2013b), and let c_w^i represent the 96-dimensional vector of mutation category counts that fall within genome window w for patient i. We define a panel $\mathbf{P}$ as a set of genome windows. Then, we denote the mutation category counts for patient i over panel $\mathbf{P}$, to be $c_{\mathbf{P}}^i = \sum_{w \in \mathbf{P}} c_w^i$.

Using scalar projection as a heuristic for signature activity, ScalpelSig seeks to find a panel where the estimated activity of q is high only in samples where q is known to have substantial activity on the whole genome (i.e., the samples in A). Formally, ScalpelSig solves the following optimization problem to find a genomic panel $\mathbf{P}$ that best detects the activity of signature q:

$$\begin{aligned} &\underset{\mathbf{P}}{\text{maximize}} && \sum_{i \in A} \text{proj}_q\left(c_{\mathbf{P}}^i\right. - \sum_{j \in S \setminus A} \text{proj}_q\left(c_{\mathbf{P}}^j\right. , \\ &\text{subject to} && |\mathbf{P}| = N. \end{aligned} \tag{4}$$

By the linearity of scalar projection, we have, for each sample i,

$$\text{proj}_q\left(c_{\mathbf{P}}^i\right. = \text{proj}_q\left(\sum_{w \in \mathbf{P}} c_w^i\right) = \sum_{w \in \mathbf{P}} \text{proj}_q\left(c_w^i\right. . \tag{5}$$

Thus, the objective [Eq. (4)] can be rewritten,

$$\begin{aligned} &\underset{\mathbf{P}}{\text{maximize}} && \sum_{w \in \mathbf{P}} \left\{ \sum_{i \in A} \text{proj}_q\left(c_w^i\right. - \sum_{j \in S \setminus A} \text{proj}_q\left(c_w^j\right. \right\}, \\ &\text{subject to} && |\mathbf{P}| = N. \end{aligned} \tag{6}$$

This formulation of the problem allows us to optimize without exploring the large combinatorial space of possible panels. Instead, ScalpelSig determines an optimal panel by simply selecting the *top scoring N* windows for the window scoring function,

$$h(w) = \sum_{i \in A} \text{proj}_q\left(c_w^i\right. - \sum_{j \in S \setminus A} \text{proj}_q\left(c_w^j\right. . \tag{7}$$

Notably, the contrastive definition of this window scoring function naturally mitigates the impact of background noise, that is, enrichment of signature-associated mutation types for spurious reasons such as underlying nucleotide composition. To see this, observe that if a window contains spurious enrichment of a set of mutation types, we would expect inactive samples to have about as many mutations of these types as active samples. As a result, we would expect the second term in the equation (the sum of scores from inactive samples) to be high, thus discouraging the selection of such a window.

Given that tumors are known to have widely varying mutation rates, it is important to ensure that the panel is not simply tailored to patients and windows with the highest numbers of total mutations. To this end, we introduce a parameter $\alpha \in (0, 1]$ and reparameterize the window scoring function to be:

$$h_\alpha(w) = \sum_{i \in A} \text{proj}_q\left(c_w^i\right.^{\alpha} - \sum_{j \in S \setminus A} \text{proj}_q\left(c_w^j\right.^{\alpha}. \tag{8}$$

Setting $\alpha = 1$ is equivalent to Equation (7), whereas setting $\alpha = 0.5$ applies square roots to each of the summed terms. With $\alpha = 0.5$, the scoring function downweights the contribution from samples that have anomalously high projections. This yields a window scoring function that favors windows with high activity in *most* active samples and not windows with high activity in *only a few* samples in the training set. While arbitrary values of α could be used, in this article we only consider $\alpha = 1$ and $\alpha = 0.5$.

We graphically illustrate the ScalpelSig algorithm in Figure 1.

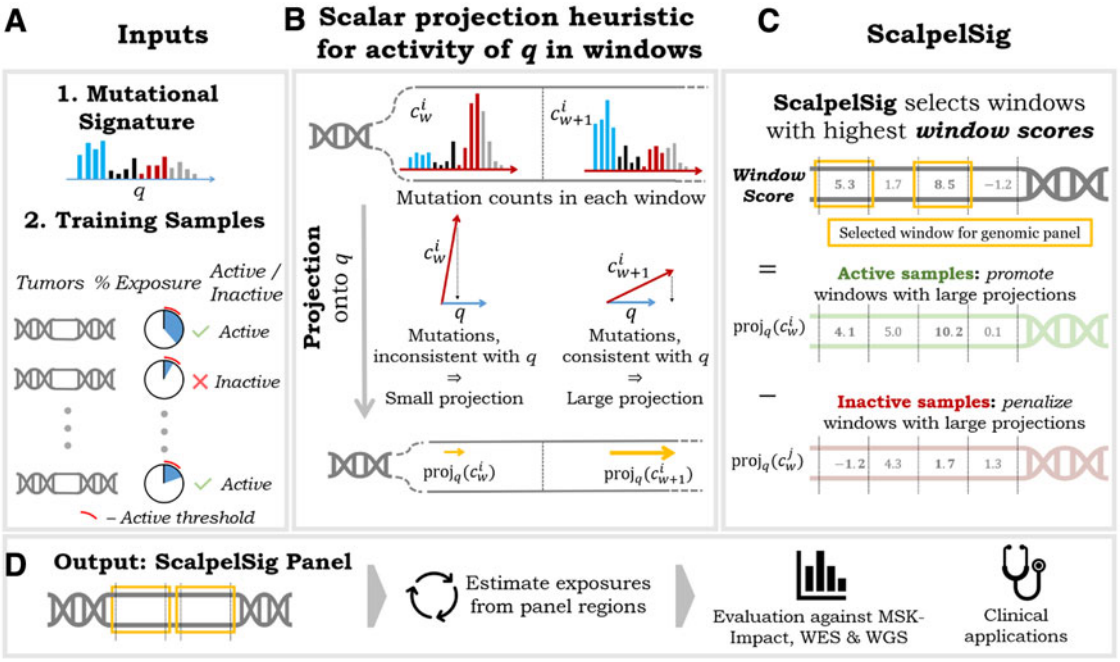

FIG. 1. ScalpelSig designs a genomic panel to detect the activity of a given mutational signature. **(A)** ScalpelSig takes as input a mutational signature q and a set S of training samples with WGS data. Signature exposures are estimated for training samples. Samples which have >5% of mutations attributed to signature q are labeled active and inactive otherwise. **(B)** Projection of each genome window's mutation count vector onto q yields a heuristic measure of signature exposure. The value of the scalar projection is highest when the given window has a high number of mutations with a distribution of mutation categories similar to q. **(C)** ScalpelSig evaluates windows with a contrastive window scoring function [see Eq. (8)] that encourages the selection of windows with large projections in active samples and small projections in inactive samples. Given a panel size parameter N, ScalpelSig combines the N top scoring windows into a panel. **(D)** ScalpelSig panels detect signature activity with improved accuracy and require considerably less genomic material than WES or WGS. Thus ScalpelSig panels offer an improved, clinically accessible assay of signature activity. WES, whole-exome sequencing; WGS, whole-genome sequencing.

3.2. Mutational signatures and breast cancer cohorts

In this work we primarily analyze a publicly available cohort of 560 breast cancer genomes (Nik-Zainal et al., 2016). We chose this dataset because of its large number of samples, and because many different mutational signatures are typically active in breast cancer (Tate et al., 2019), allowing for a varying set of test cases for our framework. Furthermore, the motivation of using mutational signatures as clinical biomarkers is particularly relevant to breast cancer, in which HR deficiency (and its associated mutational signatures) is a promising biomarker for PARP inhibitor therapy (Davies et al., 2017). WGS was performed on each sample. The dataset contains 3.5 million total mutations categorized into the standard 96 categories used in Alexandrov et al. (2013b).

To further assess the generalizability of ScalpelSig, we also evaluate genomic panels learned from the above cohort (Nik-Zainal et al., 2016) on a completely held-out cohort of 237 WGS samples from Staaf et al. (2019).

3.2.1. Computing mutational signature exposures. For the estimation of mutational signature exposures on whole genomes, as well as all ScalpelSig panels and other baselines, we use an in-house, open-source Python implementation of the SignatureEstimation framework (Huang et al., 2018; https://github.com/lrgr/signature-estimation-py).

For each sample in each cohort, we compute exposures to mutational signatures taken from COSMIC (ver. 2), a curated mutational signature repository (Tate et al., 2019). In this article, we will always refer to (and index) COSMIC signatures by their numerical names (e.g., Signature 1). Only 13 of the 30 COSMIC signatures are known to be active in breast cancer: Signatures 1, 2, 3, 5, 6, 8, 10, 13, 17, 18, 20, 26, 30 (Tate

et al., 2019). Accordingly, from each sample's genome-wide mutation counts, we compute exposures solely to these signatures. From these exposures, we define a sample as *active* for a signature if the exposure of that signature is responsible for 5% or more of the total mutations (compared to all other signatures), and *inactive* otherwise.

3.3. Evaluation of panels

For each signature of interest, we evaluate a SCALPELSIG panel by how well it detects mutation signature activity in held-out samples. As our primary measure of performance, we ask the following question: given a sample, can we determine whether a signature is active on the *whole genome* by looking solely at panel regions obtained from SCALPELSIG?

We formalize this question with the following classification task: for each sample in the test set predict whether a signature of interest is *active* on the whole genome, given the signature exposure estimated from mutations that fall within a SCALPELSIG panel.

3.3.1. Comparison of SCALPELSIG against MSK-IMPACT and other baselines. We primarily analyze the performance of SCALPELSIG on the cohort of 560 samples from Nik-Zainal et al. (2016). Our experiments use stratified sampling to split the data into training and testing sets. In each experiment we report the mean performance across 15 random test/train splits. For each signature and unless otherwise noted, SCALPELSIG uses 90% of samples as a training set to design a panel. Afterward, signature exposures are extracted from panel regions on the remaining samples. We measure how well these panel exposures distinguish active from inactive samples by computing area under the precision-recall curve (AUPR).

To further demonstrate the effectiveness of SCALPELSIG panels, we also use Spearman's rank correlation to measure the strength of the relationship between exposures computed solely from mutations that fall within panel regions and exposures computed from whole genome mutation counts. Notably, this metric is independent of the formalism of active and inactive samples. We additionally tested multiple values for the threshold at which samples were considered active (see Section 3.3.3 for details). Our primary set of experiments are performed on 2.5 Mb SCALPELSIG panels. This size was chosen to match the size of the MSK-IMPACT panel (Cheng et al., 2015). We evaluate SCALPELSIG with $\alpha \in \{0.5, 1.0\}$ and set the window size to 10 Kb for all experiments.

We perform experiments only for mutational signatures that are active in >5% of samples in the breast cancer cohort (Alexandrov et al., 2013a). We also omit Signatures 1 and 5, since these signatures are known to be endogenous and ''clock-like'' and are expected to be present in *all* samples (Alexandrov et al., 2015). In sum, we perform experiments to evaluate SCALPELSIG panels optimized for the detection of each of the six remaining signatures (2, 3, 8, 13, 18, and 30). We report the number of active samples for each signature in Table 1.

We compared SCALPELSIG panels against three benchmarks, each an alternative approach that sequences fewer bases than WGS: the MSK-IMPACT panel (Cheng et al., 2015), whole exome sequencing (WES),

TABLE 1. THE NUMBER OF *ACTIVE* SAMPLES IN THE 560 BREAST CANCER SAMPLES FOR THE MUTATIONAL SIGNATURES KNOWN TO BE PRESENT IN BREAST CANCER

Signature	*Active samples*	*Active samples*
2	232	41.4%
3	278	49.6%
8	494	88.2%
13	262	46.8%
18	64	11.4%
30	135	24.1%

A signature is active if that signature is responsible for 5% or more of the total mutations in the sample. Signatures 6, 10, 17, 20, and 26 are active in fewer than 5% of samples and, thus, are not considered in this study. Signatures 1 and 5 are also not considered because they are expected to be active in *every* sample (Alexandrov et al., 2015).

and a randomized baseline. To compare against the MSK-IMPACT panel (Cheng et al., 2015), we identified the subset of mutations which fell within panel regions. We obtained genomic coordinates, including ''off-target'' positions in noncoding regions, of the MSK-IMPACT panel from Gulhan et al. (2019). To compare against the baseline of WES, we took the subset of mutations from the dataset which fell within exonic regions as identified by the GENCODE project (Coffey et al., 2011).

We also compare ScalpelSig panels to a randomized baseline—the mean performance of 1000 random panels each 2.5 Mb in size. Each random panel is generated by sampling 250 unique windows from 10 Kb nonoverlapping windows across the genome. Windows with no mutations in the cohort were removed before sampling. The performance of each baseline is evaluated on the same test set as the ScalpelSig panels in each experiment.

3.3.2. Evaluation on held-out breast cancer cohort. To establish that the panels discovered by ScalpelSig are potentially applicable in the clinic, we further assess the generalizability of ScalpelSig panels learned from one study to new samples in another. That is, we also evaluate ScalpelSig on a completely held-out cohort of samples from Staaf et al. (2019). In this study, we use the entire cohort of 560 samples from Nik-Zainal et al. (2016) to train ScalpelSig panels and evaluate these panels in an unseen cohort of 273 samples from Staaf et al. (2019). For this experiment, we only perform analysis with respect to signatures that resulted in panels that outperform baselines in Nik-Zainal et al. (2020) (Signatures 2, 3, 8, 13, and 18).

3.3.3. Assessment of robustness. We performed experiments to test the robustness of ScalpelSig's performance under varied conditions. First, we vary the amount of genomic material included in the panels. In these experiments we construct ScalpelSig panels of sizes 0.1, 0.5, 1.0, 1.5, 2.0, and 2.5 Mb and evaluate their performance. Smaller panels are produced by running the ScalpelSig algorithm exactly as described above, except during the evaluation step, mutations are counted from, for example, the top-scoring 10 windows (for 0.1 Mb panels) instead of the top 250 (for 2.5 Mb panels).

Next, we vary the amount of data available for training the algorithm. In these experiments we construct ScalpelSig panels using 20%, 40%, 60%, and 80%, in addition to the default setting of 90%, of the 560 genome breast cancer cohort and evaluate their performance. For each experiment, training sets were obtained using stratified sampling as described above. Evaluation was conducted using all of the held-out samples as a test set.

Finally, we vary the threshold at which samples are considered to be ''active'' versus ''inactive.'' We performed experiments where samples were considered to be ''active'' if the signature of interest was responsible for 1%, 5%, 10%, and 20% of the total mutations in the sample, respectively. Cases where either the active or inactive class comprised fewer than 5% of samples were discarded (consistent with the protocol for our primary set of experiments) since these cases reduce the possible variation in the test set. Since class balance varied at each different setting of the activity threshold, distinct random test/train splits were sampled for each of the settings.

3.3.4. Construction of a preliminary combined-signature panel. We performed an exploratory set of experiments to determine whether regions from individual-signature ScalpelSig panels could be combined to detect multiple signatures simultaneously. Using ScalpelSig's window scoring function at the $\alpha = 0.5$ setting, we took 0.5 MB of the highest scoring regions of panels designed for detection of Signatures 2, 3, 8, 13, and 18, respectively (regions for the detection of Signature 30 did not outperform baselines in the primary set of experiments and, thus, were not included). We combined these regions to form a single 2.5 MB panel and evaluated its performance for detection of those five signatures. Note that when testing for the activity of multiple signatures simultaneously, samples can no longer be neatly categorized into binary classes (i.e., a sample that is active for some signatures may be inactive for others). As a result, we could not use the stratified sampling protocol that was used to split the data into test and train sets in the individual signature panel experiments. We instead used uniform random sampling to split the data into test and train sets. As in previous experiments, 90% of the breast cancer genome cohort was used as a training set, and the remaining 10% was held out for evaluation.

3.4. Software

We implemented ScalpelSig in R (ver. 3.6.3) (R Core Team, 2017). Our code makes use of the R packages PRROC (Keilwagen et al., 2014) and pROC (Robin et al., 2011). We additionally incorporate an

in-house, open-source Python implementation of the SignatureEstimation package (Huang et al., 2018). The source code is publicly available at https://github.com/lrgr/scalpelsig

4. RESULTS

In the desired use case for a mutational signature panel, a clinician would sequence panel regions and use them to make a judgment about the mutational signatures that have acted on the patient's genome. Accordingly, a good panel is one where signature activity within panel regions is predictive of signature activity genome wide. Our experiments measure how well a panel makes this prediction with two complementary metrics: (1) we compute AUPR for a binary classification task which asks whether a signature is substantially active in a sample given the exposure of that signature in panel regions; and (2) we compute the Spearman correlation between signature exposure in panel regions and signature exposure genome wide. In the interest of clinical accessibility, we focus on SCALPELSIG panels that are roughly equivalent in size to the MSK-IMPACT panel, since this size is evidently practical for clinical use. However, to further characterize SCALPELSIG's performance we analyze panels at various smaller sizes. We contextualize our results by comparing SCALPELSIG's performance with that of the MSK-IMPACT panel, a random panel, and whole exome sequencing. We show that SCALPELSIG improves panel accuracy substantially. In multiple cases, SCALPELSIG even outperformed baselines using panels that were less than $\frac{2}{3}$ the size of said baselines.

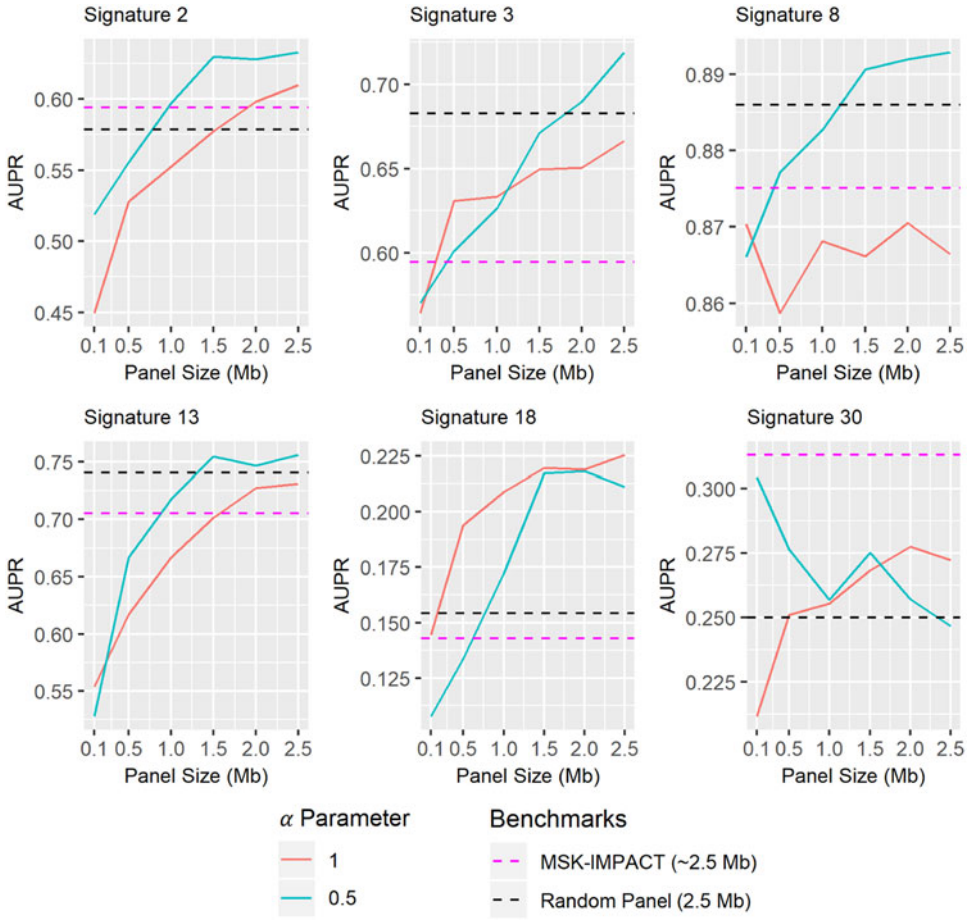

FIG. 2. Assessment of genomic panels constructed with our framework for their ability to predict mutational signature activity at various panel sizes. Values shown are mean AUPR across 15 randomized test and train sets. Each plot tests distinct panels optimized for the detection of a particular mutational signature using two different settings of the α parameter: $\alpha = 1$ (*orange*) and $\alpha = 0.5$ (*blue*). Values are compared against a randomized baseline panel (*dotted black line;* see Section 3 for details), and mean AUPR of the MSK-IMPACT panel (*dotted magenta line*), for the same test sets. AUPR, area under the precision-recall curve.

Table 2. Spearman Correlation Between Exposures Computed Only from Panel Regions and Exposures Computed from Whole Genome Mutation Counts

Signature	*ScalpelSig (2.5 Mb)*	*MSK-IMPACT (2.5 Mb)*	*Random panel (2.5 Mb)*
2	**0.3819**	0.2695	0.2068
3	**0.3883**	0.2123*	0.3138
8	**0.3895**	0.0275	0.1276
13	**0.5749**	0.4517	0.4580
18	0.1482*	0.0309*	0.0215
30	0.0569*	0.0528*	0.0091

Values shown are mean Spearman correlation coefficients across 15 randomized test and train sets. ScalpelSig is run with $\alpha = 0.5$ for all signatures. In ScalpelSig and MSK-IMPACT columns, values where fewer than half of the trials yielded p-value <0.05 are marked with an asterisk. The highest value in each row is bolded—except when most of the trials for that value are not significant.

MSK-IMPACT, Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Target.

4.1. ScalpelSig outperforms baselines

Panels discovered by ScalpelSig outperform the MSK-IMPACT panel and randomized baseline on five out of the six signatures examined (Signatures 2, 3, 8, 13, and 18) in terms of both AUPR of activity/inactivity classification (Fig. 2) and Spearman correlation between panel exposures and genome-wide exposures (Table 2). The 2.5 Mb panels constructed using the $\alpha = 1$ parameterization of the window scoring function outperformed the MSK-IMPACT panel on four out of six signatures (Signatures 2, 3, 13, and 18). The 2.5 Mb panels constructed with $\alpha = 0.5$ outperform the $\alpha = 1$ setting in almost all cases, obtaining better AUPR than both the MSK-IMPACT panel and randomized baseline—with the exception of Signature 30.

Classifications based on signatures extracted from WES are consistently more accurate than the other measurements, but this is to be expected since the exome (30 Mb) covers over ten times as much genetic material as the panels ($\leq$2.5 Mb). As such, we posit that the performance gap between the random baseline and WES gives a reasonable notion of how much the classification performance can be improved by simply observing more genomic material in a naive manner (the performance gap between MSK-IMPACT and WES could serve a similar function, but MSK-IMPACT performs worse than the random baseline in most cases). Thus the efficacy of ScalpelSig panels is demonstrated by their ability to partially bridge this performance gap (Table 3).

By this metric, ScalpelSig's increase in performance over the baseline panels ranged from moderate to sizable, bridging at least 7.9% and at most 38.3% of the performance gap between the randomized panel and whole exome sequencing (Table 3). If we were to assume, solely for the purpose of ballpark estimation, that performance scales linearly with addition of exomic regions to the panel, these improvements would represent an increase in performance proportional to the addition of between 2.1 and 10.5 Mb of exomic material to the 2.5 Mb baseline. As noted in the Section 3, the $\alpha = 0.5$ parameterization is designed to obtain results that better generalize outside of the training set by lowering the impact of windows where just a few

Table 3. Comparison of ScalpelSig ($\alpha = 0.5$; 2.5 Mb) Panels to Whole Exome (30 Mb) and Random Panel (2.5 Mb) Benchmarks for Five Out of Six Examined Signatures

	AUPR			
Signature	*Whole exome*	*Random panel*	*ScalpelSig*	*Gap bridged*
2	0.7959	0.5785	0.6330	25.1%
3	0.8727	0.6826	0.7191	19.1%
8	0.9338	0.8860	0.8929	14.5%
13	0.9370	0.7408	0.7562	7.9%
18	0.3027	0.1542	0.2111	38.3%

The right-most column gives the percentage of the performance gap between WES and the random panel that is bridged by our method.

AUPR, area under the precision-recall curve.

active samples have very high signature activity and favoring windows where a large number of active samples have moderately high activity. The improved performance of $\alpha = 0.5$ over $\alpha = 1$ suggests the necessity and effectiveness of this parameterization.

As an aside, it is worth noting that the MSK-IMPACT panel performs worse than the random baseline on all examined signatures except Signatures 2 and 30. The MSK-IMPACT panel was designed for the detection of common driver mutations, so it follows reasonably that the mutations it captures have a different distribution than the distribution of mutations over the whole genome (and consequently, the signatures obtained from its captured mutations may be inconsistent with genome-wide signatures). We additionally note that the random panel benchmark is in actuality given by the mean performance of 1000 panels with randomly selected windows (see Section 3)—thus while it is a useful point of comparison, it does not represent a clinically actionable assay, as the performance of any *individual* random panel is highly variable.

Finally, to assess the generalizability of the panels designed by ScalpelSig, we evaluated its performance on a completely held-out dataset. To identify a single panel per signature, we trained ScalpelSig on the whole cohort of 560 breast cancer genomes described above (Nik-Zainal et al., 2016) and evaluated the resulting panels using a new cohort of 237 breast cancer genomes as a test set from Staaf et al. (2019). We found that ScalpelSig continued to outperform the MSK-IMPACT panel in this setting, both in terms of Spearman correlation and AUPR (Table 4). This provides preliminary evidence that the improved performance of ScalpelSig panels generalizes beyond our initial dataset.

4.2. Robustness of performance

In this section, we present several experiments demonstrating that the ScalpelSig algorithm maintains a strong level of performance, even outside of the canonical conditions featured in the study thus far. We vary the amount of genomic material that is included in the panel, the amount of data available for training the algorithm, and we also vary the threshold at which samples are considered to be ''active'' versus ''inactive.'' The robust performance of ScalpelSig in these various settings is evidence that the algorithm may find effective application outside of the controlled setting of our study, which is in line with our goal of expanding clinical access to MSA.

4.2.1. Performance at smaller panel sizes. We investigated how performance of designed panels changed at different panel sizes, moving beyond our initial focus on the 2.5 Mb panel size used by MSK-IMPACT. One would naturally expect that the more genomic material is sampled, the better performance should become, since the number of observed mutations approaches the whole genome mutation count as the panel gets larger. This sensible intuition holds in most cases, but there are a few exceptions worth discussing. First, we note that panel performance is *not* a monotonically increasing function of panel size (Fig. 2). This phenomenon is unintuitive, but in fact observing additional genome regions is not guaranteed to increase accuracy and may even decrease it. To see this, it is important to understand that individual genome windows may have mutation distributions which differ starkly from the genome-wide distribution, due to variation in nucleotide composition and other factors. Adding such windows to a panel could

Table 4. Evaluation of ScalpelSig ($\alpha = 0.5$) and MSK-IMPACT on a Completely Held-Out Dataset of 237 Breast Cancer Genomes from Staaf et al. (2019)

	Spearman correlation		*AUPR*	
Signature	*ScalpelSig*	*MSK-IMPACT*	*ScalpelSig*	*MSK-IMPACT*
2	**0.3437**	0.1804	**0.3844**	0.2900
3	**0.5048**	0.4749	**0.9321**	0.9137
8	**0.4275**	0.2004	**0.9406**	0.9195
13	**0.5365**	0.3276	**0.7318**	0.6837
18	—	—	**0.1458**	0.0298

Each row reports the results of a single ScalpelSig panel, trained on all 560 samples from the previous dataset. Spearman correlations with p-value ≥ 0.05 are not shown. The highest values in each row for each of the two evaluation metrics (Spearman and AUPR) are bolded.

produce a mutation distribution that misrepresents the genome-wide distribution, despite containing a greater number of mutations. This misleading distribution would result in a false impression of signature activity. Thus it is reasonably possible for performance to decline with the addition of certain windows.

We further observe that panels constructed with $\alpha = 0.5$ for Signatures 2, 13, and 18 all appear to plateau in their performance before the 2.5 Mb panel size is reached—this plateau seems to occur at 1.5 Mb for Signatures 2 and 13 and at 2.0 Mb for Signature 18. This indicates that in some cases, panels significantly smaller than those presently in clinical use may be sufficient to achieve the full performance boost provided by our framework. If one recalls that the panel is formed from the highest-scoring windows, this behavior makes sense: as the panel gets bigger, progressively lower-scoring windows are added, so it follows that the performance increase might plateau.

4.2.2. Performance with limited training data. To investigate the amount of data required to effectively train ScalpelSig, we performed experiments in which we varied the amount of training data. While using 90% of the cohort (504 samples) as training data were the default setting, we additionally ran ScalpelSig using 80%, 60%, 40%, and 20% of the cohort (448, 336, 224, and 112 samples, respectively) as a training set and observed only slight declines in performance (Table 5). Indeed, even at the lowest setting (using 20% of the cohort as training data rather than the usual 90%) ScalpelSig continues to outperform MSK-IMPACT in terms of both Spearman correlation and AUPR for Signatures 2, 3, 8, and 13 (see Table 2 and Fig. 2 for the performance of MSK-IMPACT). These results demonstrate the stability of the performance demonstrated in the previous sections, but more importantly they suggest that ScalpelSig may be effective even when applied to cancer types that have substantially fewer whole-genome samples available. This broadens the potential use cases of ScalpelSig, furthering the goal of expanding clinical access to MSA.

4.2.3. Performance across varied activity thresholds. A critical component of the ScalpelSig algorithm is separating genomes into binary ''active'' and ''inactive'' categories for each signature. The threshold of signature activity for this categorization is a parameterization choice. By default, we set the activity threshold to be 5% (i.e., if a signature contributes 5% or more of the mutations in a genome, we consider that genome to be ''active'' for that signature). However, to show the robustness of the algorithm to different parameterization choices, we experimented with other thresholds.

We tested multiple values for the activity threshold (1%, 5%, 10%, 20%) and show these results in Table 6. We found that ScalpelSig maintains a strong performance across these varied settings—ScalpelSig outperforms MSK-IMPACT across the board for Signatures 2, 3, 8, and 13. Of note, the default activity threshold of 5% generally performed the best, but this was not always the case. For example, on Signature 2 the stricter thresholds of 10% and 20% achieved a moderate increase in performance compared to the default. These results are interesting as they suggest that the algorithm could obtain even higher performance by tuning the activity threshold parameter for specific use cases, but we leave a detailed exploration of this idea for future work. An additional detail of Table 6 is that different test and train sets were sampled in each row. This was a necessary step since the class balance changes as the activity threshold is varied (see the Active Samples column), and we use stratified sampling to guarantee that the class balance of test sets is equivalent to the overall cohort. As a consequence, the values for the MSK-IMPACT panel vary across experiments despite the fact that the observed genome regions remain the same.

4.3. Preliminary result for a combined-signature panel

In the present study we break down the problem of mutational signature detection from small amounts of genomic material by designing a distinct panel for each *individual* signature of interest. However, it would be ideal to detect the activity of multiple signatures simultaneously from the same genomic regions. We performed an exploratory set of experiments to determine whether regions from individual-signature ScalpelSig panels could be combined to detect multiple signatures simultaneously. We took 0.5 MB of the highest scoring regions of panels designed for detection of Signatures 2, 3, 8, 13, and 18, respectively (see Section 3 for details). We combined these regions to form a single 2.5 MB panel and observed its performance for detecting the same five signatures. The results are reported in Table 7. Overall, the combined panel tended to outperform MSK-IMPACT, but by a less impressive margin than was seen in the individual panel experiments. This is an interesting result, as it indicates that the individual panels are

Table 5. Spearman Correlation Between Exposures Computed Using Panel Regions and Exposures Computed from Whole Genome Mutation Counts, As Well As AUPR for Signature Activity Prediction, for Panels Constructed by ScalpelSig Using Various Amounts of Training Data

Signature	*% Training data*	*Spearman*	*AUPR*
2	90	0.3819	0.6330
	80	0.4006	0.6462
	60	0.3850	0.6199
	40	0.3975	0.6302
	20	0.3639	0.6197
3	90	0.3883	0.7191
	80	0.3830	0.6984
	60	0.3635	0.6888
	40	0.3572	0.7002
	20	0.3722	0.7025
8	90	0.3895	0.8929
	80	0.3479	0.8834
	60	0.3600	0.891
	40	0.3434	0.8891
	20	0.3400	0.8978
13	90	0.5749	0.7562
	80	0.5467	0.7433
	60	0.5393	0.7405
	40	0.5596	0.7372
	20	0.5416	0.7409
18	90	0.1482*	0.2111
	80	0.0792*	0.158
	60	0.0749*	0.1612
	40	0.0792*	0.141
	20	0.0965*	0.163
30	90	0.0569*	0.2499
	80	0.1032*	0.2493
	60	0.0787*	0.2567
	40	0.06884*	0.2479
	20	0.0888	0.254

For each row, 15 randomized training sets were obtained using stratified sampling as detailed in the Section 3. The size of the training sets for each row is indicated in the Training Data column. The ScalpelSig algorithm was run separately on each of the training sets and evaluated using all of the held-out samples. Values shown are mean Spearman correlation coefficients and mean AUPR across the randomized trials. Spearman values where fewer than half of the trials yielded p-value <0.05 are marked with an asterisk.

indeed specialized for the detection of their respective signatures. That is, the regions detected in the individual-panel experiments are not interchangeable—mixing and matching incurs a substantial penalty to performance. This suggests that the genome regions detected by our algorithm may have biologically meaningful differences in terms of their active mutational processes.

5. DISCUSSION

While a growing body of literature attests to the efficacy of mutational signature activity as a predictive biomarker for targeted cancer therapies, the sequencing demands of almost all methods for MSA are not met by current clinical infrastructure. In this study we seek to address this problem with ScalpelSig, an algorithm that designs genomic panels optimized for the detection of mutational signature activity. ScalpelSig takes as input a mutational signature and a set of training tumor genomes and learns genome regions in which mutations are highly indicative of signature activity.

Table 6. Spearman Correlation Between Exposures Computed Using Panel Regions and Exposures Computed from Whole Genome Mutation Counts, for Panels Constructed Using Various Thresholds for Activity Classification

Signature	*Activity threshold*	*Active samples*	*ScalpelSig ($\alpha = 0.5$)*	*MSK-IMPACT*
2	1	88.4	0.3965	0.2688
	5	41.4	0.3819	0.2695
	10	23.6	0.4171	0.3193
	20	10.7	0.4532	0.2903
3	1	66.4	0.3832	0.1597*
	5	49.6	0.3883	0.2123*
	10	38.4	0.3775	0.2733
	20	28.4	0.4020	0.2273*
8	1	95.5	—	—
	5	88.2	0.3895	0.0275
	10	73.6	0.2815	0.0447*
	20	27.0	0.3035	0.1000*
13	1	92.0	0.5378	0.4234
	5	46.8	0.5749	0.4517
	10	73.6	0.4905	0.4636
	20	27.0	0.5125	0.4671
18	1	32.9	0.0803*	0.0213*
	5	11.4	0.1482*	0.0309*
	10	4.6	—	—
	20	0.5	—	—
30	1	63.0	0.1255*	0.1163*
	5	24.1	0.0569*	0.0528*
	10	3.5	—	—
	20	0.2	—	—

The Activity Threshold column gives the required percentage of mutations contributed by a signature for a sample to be considered ''active.'' That is, in the first row samples were classified as ''active'' if at least 1% of their mutations were contributed by Signature 2; in the second row samples were considered ''active'' if at least 5% (the default for this study) came from Signature 2, etc. For the ScalpelSig columns, and the MSK-IMPACT column, values shown are mean Spearman correlation coefficients across 15 randomized test and train sets. Different random test/train splits were sampled for each row. Cases where either the active or inactive class comprised fewer than 5% of the samples were discarded, since these cases reduce the possible variation in the test set. Spearman values where fewer than half of the trials yielded p-value <0.05 are marked with an asterisk.

Table 7. Preliminary Results from Assessing Genomic Panels Designed to Detect Multiple Signatures Simultaneously

	Spearman correlation		*AUPR*	
Signature	*ScalpelSig*	*MSK-IMPACT*	*ScalpelSig*	*MSK-IMPACT*
2	0.3865	0.3162	0.6232	0.6091
3	0.3700	0.1698	0.6894	0.5790
8	0.2948	0.0507	0.8918	0.8900
13	0.4796	0.4401	0.7270	0.6962
18	0.1102	0.0675	0.1723	0.1193

The panels were constructed by combining windows from ScalpelSig ($\alpha = 0.5$) panels for individual signature detection. No windows for detection of Signature 30 were incorporated into the panel, and we did not evaluate the combined panel's performance at detecting Signature 30, since the ScalpelSig panels that were individually optimized for Signature 30 detection did not outperform baselines. Values shown are median AUPR and Spearman correlation across 15 randomized test and train sets. We used uniform random sampling to split the data into test and train sets (see Section 3 for details). This means that there was more variance in test sets of the combined panel experiments, which likely accounts for the differences in the MSK-IMPACT scores in comparison to the individual signature panel experiments. MSK-IMPACT, Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Target.

In the present study, we train ScalpelSig on breast cancer data to obtain panels optimized for the detection of six respective mutational signatures. We find that in five out of six examined signatures, ScalpelSig panels outperform the commonly studied MSK-IMPACT panel and a random baseline. The increased accuracy of our panels is substantial even compared to whole exome sequencing (Table 3), which uses over 10× as much genomic material as targeted panel sequencing. We show that ScalpelSig panels bridge significant portions of the performance disparity between the baseline panels and the whole exome benchmark. In the most favorable case, this performance increase is roughly proportional to the addition of 10.5 Mb of exonic material to the baseline panels. We find that the performance of ScalpelSig is robust under a variety of parameterizations. For four of six examined signatures, panels smaller than 2.5 Mb are sufficient to obtain all or most of the demonstrated performance increase. Furthermore, our panels maintain strong performance even when the amount of training data is significantly reduced. These results suggest that the performance increase afforded by our method may generalize beyond the conditions of our study—for example, to other cancer types that have less training data available.

Of the signatures investigated in this article, Signature 3 is arguably the most exciting as a clinically actionable biomarker. Signature 3 has been found to be correlated with biomarkers of homologous recombination repair deficiency (Nik-Zainal et al., 2016; Davies et al., 2017; Polak et al., 2017; Gulhan et al., 2019), which is a biomarker for PARP inhibitor therapy (Farmer et al., 2005). Recently, Gulhan et al. (2019) developed an algorithm for improved detection of Signature 3 from MSK-IMPACT panel data. Our results demonstrate that ScalpelSig panels confer a substantial increase in accuracy for detection of Signature 3 compared to MSK-IMPACT when using standard methods for signature detection. Concretely, relative to MSK-IMPACT, ScalpelSig panels give an 83% increase in Spearman correlation coefficient between Signature 3 activity in panel regions and genome-wide Signature 3 activity. We note however that when the panels were evaluated on a held-out dataset, the improvement for Signature 3 was much more modest, indicating that the impact of distributional changes across data sets merits further investigation. Even still, this begs the question of whether the algorithm used in Gulhan et al (2019) could be trained using ScalpelSig panel data to achieve a synergistic boost in accuracy for detection of Signature 3. If so, combining the two approaches could be a boon for detecting HR deficiency in the clinic. We point to this idea as an important direction for future work.

For other future work, we plan to investigate other methods for inferring signature exposures and active signatures, including Huang et al. (2018)'s method for automatically detecting which signatures are active in each sample with confidence. While we show that ScalpelSig's good performance is stable on a variety of parameterizations, additional work toward finding optimal parameterizations (e.g., hyperparameter tuning using cross-validation) may be beneficial. We leave this fine-grained optimization for future work. In terms of exploring clinical applications of ScalpelSig, one obvious avenue would be to train the algorithm for other cancer types and signatures and validate it with completely held-out datasets. Ultimately, we anticipate that such a validated genomic panel for detecting multiple signatures across multiple cancer types will be the most straightforward path to clinical impact. To that point, we performed an exploratory analysis that points to the challenge of creating a multiple signature panel. Further improvement of the accuracy of such a panel will likely require a more sophisticated algorithm—one that can score genome regions for how well they assay the activity of multiple signatures simultaneously.

ACKNOWLEDGMENT

The authors thank Marina Knittel for her helpful suggestions during the design of the window scoring function.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no conflicting financial interests.

FUNDING INFORMATION

This research was supported by a grant from the United States—Israel Binational Science Foundation (BSF), Jerusalem, Israel. This project has been made possible, in part, by National Science Foundation

Graduate Research Fellowship Grants No. DGE-1842165 to N.F. and DGE-1840340 J.F. This work was also funded, in part, by NSF award DGE-1632976 to J.F. This work was supported, in part, by the Intramural Research Program of the National Cancer Institute, Center for Cancer Research, NIH.

REFERENCES

Alexandrov, L.B., Jones, P.H., Wedge, D.C., et al. 2015. Clock-like mutational processes in human somatic cells. *Nat. Genet.* 47, 1402–1407.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., et al. 2020. The repertoire of mutational signatures in human cancer. *Nature* 578, 94–101.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., et al. 2013a. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* 3, 246–259.

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., et al. 2013b. Signatures of mutational processes in human cancer. *Nature* 500, 415–421.

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

Campbell, B.B., Light, N., Fabrizio, D., et al. 2017. Comprehensive analysis of hypermutation in human cancer. *Cell* 171, 1042–1056.

Cheng, D.T., Mitchell, T.N., Zehir, A., et al. 2015. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. *J. Mol. Diagn.* 17, 251–264.

Coffey, A.J., Kokocinski, F., Calafato, M.S., et al. 2011. The GENCODE exome: Sequencing the complete human exome. *Eur. J. Hum. Genet.* 19, 827–831.

Davies, H., Glodzik, D., Morganella, S., et al. 2017. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.* 23, 517–525.

Farmer, H., McCabe, N., Lord, C.J., et al. 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* 434, 917–921.

Frampton, G.M., Fichtenholtz, A., Otto, G.A., et al. 2013. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* 31, 1023–1031.

Franzese, N., Fan, J., Sharan, R. and Leiserson, M.D. (awaiting publication), Scalpelsig: Automated design of genomic panels to expand clinical access to mutational signature analysis. *In Proceedings of RECOMB 2021.*

Garraway, L.A., and Lander, E.S. 2013. Lessons from the cancer genome. *Cell* 153, 17–37.

Gerstung, M., Jolly, C., Leshchiner, I., et al. 2020. The evolutionary history of 2,658 cancers. *Nature* 578, 122–128.

Gulhan, D.C., Lee, J.J., Melloni, G.E.M., et al. 2019. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.* 51, 912–919.

Hanahan, D., and Weinberg, R.A. 2011. Hallmarks of cancer: The next generation. *Cell* 144, 646–674.

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

Helleday, T., Eshtad, S., and Nik-Zainal, S. 2014. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15, 585–598.

Hoeck, A., Tjoonk, N.H., Boxtel, R.V., et al. 2019. Portrait of a cancer: Mutational signature analyses for cancer diagnostics. *BMC Cancer* 19, 457.

Huang, X., Wojtowicz, D., and Przytycka, T.M. 2018. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* 34, 330–337.

Jiao, W., Atwal, G., Polak, P., et al. 2020. A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nat. Commun.* 11, 728.

Kandoth, C., McLellan, M.D., Vandin, F., et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature* 502, 333–339.

Keilwagen, J., Grosse, I., and Grau, J. 2014. Area under precision-recall curves for weighted and unweighted data. *PLoS One* 9:e92209. https://doi.org/10.1371/journal.pone.0092209.

Kim, J., Mouw, K.W., Polak, P., et al. 2016. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* 48, 600–606.

Lawrence, M.S., Stojanov, P., Mermel, C.H., et al. 2014. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501.

Le, D.T., Durham, J.N., Smith, K.N., et al. 2017. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science* 357, 409–413.

Lord, C.J., and Ashworth, A. 2017. PARP inhibitors: Synthetic lethality in the clinic. *Science* 355, 1152–1158.

Morganella, S., Alexandrov, L.B., Glodzik, D., et al. 2016. The topography of mutational processes in breast cancer genomes. *Nat. Commun.* 7, 11383.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., et al. 2012. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993.

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

Nik-Zainal, S., Kucab, J.E., Morganella, S., et al. 2015. The genome as a record of environmental exposure. *Mutagenesis* 30, 763–770.

Nik-Zainal, S., Memari, Y., Davies, H.R. 2020. Holistic cancer genome profiling for every patient. *Swiss Medical Weekly* 150, w20158.

Perner, J., Abbas, S., Nowicki-Osuch, K., et al. 2020. The mutREAD method detects mutational signatures from low quantities of cancer DNA. *Nat. Commun.* 11, 3166.

Polak, P., Karlić, R., Koren, A., et al. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 518, 360–364.

Polak, P., Kim, J., Braunstein, L.Z., et al. 2017. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.* 49, 1476–1486.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robin, X., Turck, N., Hainard, A., et al. 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12, 77.

Rosenthal, R., McGranahan, N., Herrero, J., et al. 2016. deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31.

Sason, I., Chen, Y., Leiserson, M.D.M., et al. 2020. A mixture model for signature discovery from sparse mutation data, 271–272. *In Proceedings of RECOMB 2020.* Springer, Padua, Italy.

Staaf, J., Glodzik, D., Bosch, A., et al. 2019. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.* 25, 1526–1533.

Stratton, M.R., Campbell, P.J., and Futreal, A.P. 2009. The cancer genome. *Nature* 458, 719.

Tate, J.G., Bamford, S., Jubb, H.C., et al. 2019. COSMIC: The catalogue of somatic mutations in cancer. *Nucleic Acids Res.* 47(D1), D941–D947.

Temko, D., Tomlinson, I.P., Severini, S., et al. 2018. The effects of mutational processes and selection on driver mutations across cancer types. *Nat. Commun.* 9, 1857.

Tubbs, A., and Nussenzweig, A. 2017. Endogenous DNA damage as a source of genomic instability in cancer. *Cell* 168, 644–656.

Van Hoeck, A., Tjoonk, N.H., van Boxtel, R., et al. 2019. Portrait of a cancer: Mutational signature analyses for cancer diagnostics. *BMC Cancer* 19, 457.

Zugazagoitia, J., Guedes, C., Ponce, S., et al. 2016. Current challenges in cancer treatment. *Clin. Ther.* 38, 1551–1566.

Address correspondence to:
Dr. Mark D.M. Leiserson
Department of Computer Science and
Center for Bioinformatics and Computational Biology
University of Maryland
College Park, MD 20742
USA

E-mail: mdml@umd.edu