TrackSig: reconstructing evolutionary trajectories of mutations in cancer

Yulia Rubanova[1,2,5], Ruian Shi[1,2], Roujia Li[2], Jeff Wintersinger[1,2,5], Nil Sahin[4,2,5], Amit Deshwar[3], Quaid Morris[1,2,3,4,5,*], PCAWG Evolution and Heterogeneity Working Group[6], and PCAWG network[6]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Canada
[3]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada
[4]Department of Molecular Genetics, University of Toronto, Toronto, Canada
[5]Vector Institute, Toronto, Canada
[6]Various affiliations
[*]quaid.morris@utoronto.ca

On behalf of the PCAWG Evolution and Heterogeneity Working Group and the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network.

ABSTRACT

We present a new method, TrackSig, to estimate the evolutionary trajectories of signatures of somatic mutational processes. TrackSig uses cancer cell fraction (CCF) corrected by copy number to infer an approximate order in which the somatic mutations accumulate. TrackSig segments mutation ordering by CCF and fits signature exposures (activities) as a piece-wise constant function of the mutation ordering. TrackSig uses optimal segmentation to find the points of change in signature activities.
We assess TrackSig's reconstruction accuracy using simulations. We find 2% median activity error on simulations with one to three change-points. The size and the direction of the signature change is consistent in 83% and 95% of cases respectively. There were an average of 0.02 missed change-points and 0.12 false positive change-points per sample. We provide a framework to estimate signature exposure trajectories across CCF scale as well as the way to determine active signatures. The code is available at https://github.com/YuliaRubanova/TrackSig.

1 Introduction

Somatic mutations accumulate throughout our lifetime, arising from external sources or from processes intrinsic to the cell[1,2]. Some sources generate characteristic patterns of mutations. For example, smoking is associated with G to T mutations; UV radiation is associated with C to T mutations[3–5]. Some processes provide a constant source of mutations[6] while others are sporadic[7].

Using mutational signature analysis, one can estimate the contribution of different mutation processes to the collection of somatic mutations present in a sample. In this type of analysis, single nucleotide variants (SNVs) are classified into 96 types based on the type of substitution and tri-nucleotide context (e.g., ACG to a ATG)[2]. Mutational signatures across the 96 types were derived by non-negative matrix factorization in the previous work by *Alexandrov et al.*[2]. Many of the signatures are strongly associated with known mutational processes including smoking[7], non-homologous double strand break repair[2], and ionizing radiation[8]. The activities of some signatures are correlated with patient age[6] and suggesting their use as a molecular clock[9]. Thus, signature analysis can identify the DNA damage repair pathways that are absent in the cancer, can predict prognosis[10] or guide treatment choice[11].

Formally, a *mutational signature* is a probability distribution over these 96 types, where each element is a

probability of generating mutations from the corresponding type[12]. Each signature is assigned an *activity* (also called *exposure*) which represents the proportion of mutations that the signature generates. These can be computed for pre-defined signatures from the total mutational spectrum of a sample by constrained regression[13,14].

Mutational sources can change over time. Mutations caused by carcinogen activity stop accumulating when the activity ends[7]. Mutations associated with defective DNA damage repair, such as BRCA1 loss[1,2] will begin to accumulate after that loss. Recent analyses have reported modest changes in signature activities between clonal and subclonal populations[9,15] based on groups of mutations identified by clustering their variant allele frequencies (VAFs). However, the accuracy of these methods relies heavily on the sensitivity and precision of this clustering, which is low for typical whole genome sequencing coverage[16].

In this paper we introduce TrackSig, a new method to reconstruct signature activities across time without VAF clustering. We use VAF to approximately order mutations based on their prevalence within the cancer cell population and then track changes in signature activity consistent with this ordering.

Using TrackSig, we have previously demonstrated that signature activities change often during the lifetime of a cancer and that these changes can help identify new subclonal lineages[17]. Here, we use bootstrap analyses and realistic simulations to help assess the accuracy of those reconstructions.

2 Methods

TrackSig has two stages. First, we sort single nucleotide variants (SNVs) by an estimated order of their accumulation. We compute this estimate using their variant allele frequencies (VAFs) and a copy number aberration (CNA) reconstruction of the samples. Next, we infer a trajectory of the mutational signature activities over the estimated ordering of the SNVs. We estimated activity trajectory for each signature is a piecewise constant function of the SNV ordering with a small number of change-points. These stages are described in detail below.

TrackSig is designed to be applied to VAF frequency data from a single, heterogeneous tumour sample. However, if an ordering of mutations is available through another sources, for example, a reconstruction of the cancer phylogeny, then this ordering can be used directly and the first stage of TrackSig can be omitted.

2.1 Estimating the order of acquisition of SNVs

We assume SNVs to be persistent and cumulative, meaning that mutations cannot be reverted to the reference state and no position is mutated twice. This is known as the *infinite sites assumption*[1,18].

Under this assumption, SNVs acquired earlier in the evolution of a cancer will generally be more prevalent in the population of tumour cells. In TrackSig, we sort mutations according to decreasing cancer cell fraction (CCF) thus assuming that mutations with higher CCF were acquired earlier. This assumption can be violated if multiple major subclones from different branches are represented in the sample. However, this situation occurs rarely[19], and may manifest as a characteristic oscillation in the reconstructed activities. See sec 4.3 for more details.

2.1.1 Estimating cancer cell fraction

Estimating a SNV's CCF requires both an estimate of its VAF and an estimate of the average number of alleles per cell at the locus where the SNV occurs. In TrackSig, we derive this estimate from a CNA reconstruction provided with the VAF inputs.

To account for uncertainty in a SNV's VAF due to the finite sampling, we model the posterior distribution over its VAF using a Beta distribution:

$$\mathrm{VAF} \sim \mathrm{Beta}(n_{\mathrm{var}}, n_{\mathrm{ref}}), \tag{1}$$

where n_{var} is the number of reads carrying a variant, and n_{ref} is the number of reference reads. To simplify the algorithm, and the subsequent sorting step, we sample an estimate of VAF_i (VAF of SNV i) from this

distribution. This gives us a single sampled ordering. With a large number of SNVs, we expect little variability in the estimated activity trajectory due to uncertainty in the VAFs of individual SNVs. With a smaller number of SNVs, multiple orderings can be sampled and the trajectories combined.

If no CNA reconstruction is available, TrackSig assumes that each SNV is in a region of normal copy number and TrackSig estimates CCFs in autosomal regions by setting:

$$\mathrm{CCF}_i = \frac{2 * \mathrm{VAF}_i}{\mathrm{Purity}}, \tag{2}$$

where Purity is the purity (i.e. proportion of cancerous cells) of the sample. If purity is not provided, TrackSig assumes Purity $= 1$.

If a CNA reconstruction is available, TrackSig uses it when converting from VAF to CCF. In regions of subclonal CNAs, making this conversion requires a phylogenetic reconstruction[16,20]. As such, we filter SNVs in these regions out when ordering SNVs in order to avoid this time consuming operation. However, TrackSig can make use of orderings of SNVs in regions of subclonal CNAs if provided by a phylogeny-aware method[16,21]. Also, TrackSig assumes there is a maximum of one variant allele per cell, and thus estimates CCF by setting:

$$\mathrm{CCF}_i = \frac{(2 + \mathrm{Purity} * (\mathrm{CN}_i - 2))}{\mathrm{Purity}} * \mathrm{VAF}_i, \tag{3}$$

where Purity is the purity of the sample, and CN_i is the clonal copy number of the locus. If the clonal CNA increases the number of variant alleles per cell, this will lead to CCFs larger than one. As such, these cases are easily detected and corrected.

TrackSig sorts SNVs in order of decreasing estimated CCF and use the rank of the SNV in this list as a "pseudo-time" estimate of its time of appearance. Note that this estimate will have a non-linear relationship to real time, if the overall mutation rate can vary during the tumour's development. If some of the SNVs can be interpreted as clock mutations, an SNV's rank can be converted into an estimate of real time (see, e.g,[9] for details).

2.1.2 Constructing a timeline

To derive an estimate of the activity trajectory, TrackSig converts the SNV ordering into a set of time points with non-overlapping subsets of the SNVs. TrackSig first partitions the ordered mutations into bins of 100 mutations and interpret each bin as one time point. The *timeline* of the cancer is the collection of the time points. TrackSig reports signature activity trajectories as a function of points in the timeline. Note that it does not use any information about subclones when partitioning the SNVs and that it is only using CCFs for the SNV from a single sample.

2.2 Computing activities to mutational signatures

To estimate activity trajectories, TrackSig partitions the timelines into sets containing one or more time points. Within each of these sets, it estimates signature activities using mixture of discrete distributions. Full details of the model are provided in the appendix A. In brief, TrackSig models each signature as a discrete distribution over the 96 types and it treats the mutation count vector over the 96 types as a set of independently and identically distributed samples from a mixture of the discrete distributions corresponding to each signature. The mixing coefficients of these distributions are interpreted as their activities for the mixture model that produced the set of mutations. TrackSig fits these activities using the Expectation-Maximization algorithm[22].

2.3 Detecting change-points

TrackSig identifies change-points in the timeline where there are discernible differences in the activity of mutations in the time points before and after the change-points. Specifically, the change-points delineate the partitions of the timeline into sets of mutations with approximately constant activities. TrackSig fits

activities for this set, as described above. This procedure generates piecewise constant activity trajectories for each signature. To select change-points, we adapt Pruned Exact Linear Time (PELT)[23], an optimal segmentation algorithm based on dynamic programming. We impose complexity penalty at each time point that is equivalent to optimizing the Bayesian Information Criteria (BIC) (see Supplement B for details). To reduce variance in our estimates of the signature activities, we do not allow partitions to be smaller than 4 time points (400 mutations).

We compute the BIC criteria the following way. Change-points split the timeline into $(\#\,\text{changepoints}+1)$ sections. In each section, TrackSig fits the signature activities, which have to sum to one. Therefore there are $(\#\,\text{signatures}-1)$ free parameters per section, or $(\#\,\text{changepoints}+1)\cdot(\#\,\text{signatures}-1)$ free parameters in total. As such, BIC objective takes the following form:

$$BIC = -2\ln\hat{\mathscr{L}} + (\#\,\text{changepoints}+1)\cdot(\#\,\text{signatures}-1)\cdot\ln(\#\,\text{timepoints}), \tag{4}$$

where $\hat{\mathscr{L}}$ is the likelihood of the current model.

2.4 Correcting cancer cell fraction greater than 1

If the number of variant alleles per cell is increased by a clonal copy number change, TrackSig's CCF estimates might be greater than 1. To correct for this, when displaying activity trajectories, it merges all the time points that have average CCF ≥ 1 into one time point. As such, the first time point can contain more than 100 mutations. To determine a signature activity at this new time point, TrackSig simply takes an average activity of all merged time points (those having CCF ≥ 1).

2.5 Bootstrapping to estimating activity uncertainty

TrackSig estimates uncertainty in the activity estimates by bootstrapping the mutations and refitting the activity trajectories. Specifically, it takes the random subset of N mutations by sampling uniformly with replacement from the N unfiltered SNVs in the sample under consideration. Using the pre-assigned CCF estimates, we sort the SNVs in decreasing order, as above, re-partition them into time points and recompute activity estimates. The trajectories obtained from bootstrapped mutation sets have the same number of time points, however the average CCF for each time point can change. We use these bootstrapped trajectories to compute uncertainty estimates for the sizes of activity changes.

3 Results

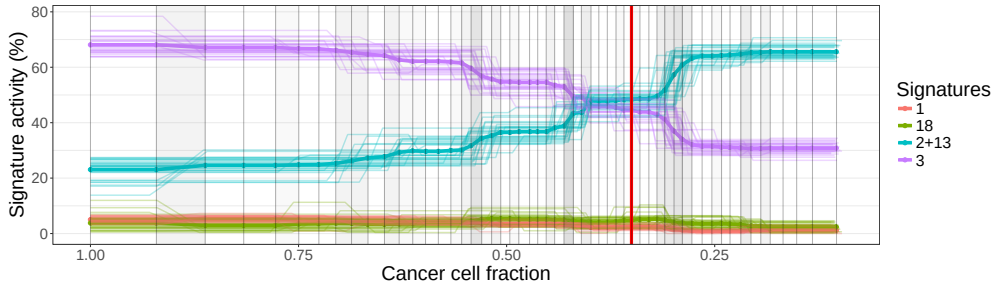
TrackSig was applied to the 2,552 whole-genome sequencing samples with more than 600 SNVs contained within the white and grey lists of the Pan-cancer Analysis of Whole Genomes (PCAWG) group. The results of these analyses are described elsewhere[17]. Here we describe the simulations establishing TrackSig's performance characteristics and provide some methodological details of TrackSig's use in PCAWG.

3.1 Choice of mutation signatures

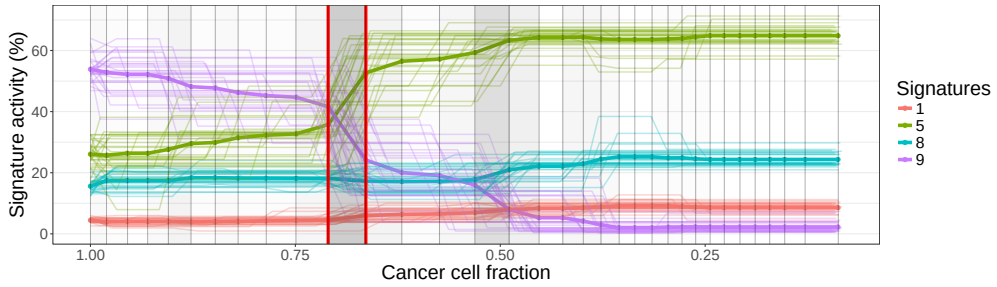
Following *Alexandrov et al.*[2], we classify mutations into 96 types based on their three-nucleotide context. Point mutations fall into 6 different mutation types (i.e., C -> [AGT] and T -> [ACG]) excluding complementary pairs. There are 16 (4*4) possible combinations of the 5' and 3' nucleotides. Thus, SNVs are separated into 96 ($K = 16*6 = 96$) types.

Within the context of PCAWG, we use the set of 48 signatures developed by PCAWG-Signature group. The first 30 of those signatures are slightly modified versions of original signatures defined by *Alexandrov et al.*[2,12] and have the same numbering and interpretation. The original 30 signatures are described at COSMIC[1]. Signature analysis methods, including TrackSig, fit activities for only a subset of the signatures. These signatures are called the *active* signatures. The activities for the non-active signatures are clamped to

[1] http://cancer.sanger.ac.uk/cosmic/signatures



(a) Breast cancer



(b) Chronic lymphocytic leukemia

Figure 1. Activity trajectories for two samples. Each plot is constructed from VAF data from one tumour sample. Each line is an activity trajectory that depicts inferred activities for a single signature (y-axis) as a function of decreasing CCF (x-axis). Signatures are indicated by colour. The thin lines are trajectories from each of 30 bootstrap runs. The bold line depicts the mean activities across bootstraps. The vertical lines indicate time points in the original dataset, and are placed at the average CCF of their associated mutations. As such, the distance between time points in the plot is inversely proportional to the density of the mutations near that CCF. Changes in activity trajectories are not necessarily aligned with vertical bars because mean CCFs of time points change across bootstraps. Frequency of changepoints between two vertical bars is indicated by shade, the darker shades indicate higher density of change-points. Subclonal boundaries are shown in red vertical lines. Subclonal information is not used in trajectory calculation and is only shown for comparison. **(a) Breast cancer sample** In clonal signatures remain constant with dominating signature 3 (associated with BRCA1 mutations). In the subclone activity to signature 3 decreases and is replaced by SNVs associated with APOBEC/AID (signatures 2 and 13). **(b) Chronic lymphocytic leukemia sample** Signature 9 (somatic hypermutation) dominates during clonal expansion and drops from 55% activity to almost zero in the subclone. Signature 5 compensates for this change.

zero. For example, signature 7 is associated with ultraviolet light has been detected almost exclusively in skin cancers[2]. As such, it is only assigned active status in skin cancers. In our analysis, we use the active signatures reported by PCAWG-Signature group. For analyses based on COSMIC signatures, one can use active signatures per cancer type is provided on COSMIC website. TrackSig can also be used to automatically select active signatures, as described in a later section.

3.2 Signatures with most changing activities

In this section we analyze the variation of signature activities on PCAWG data across time and across samples. We compute the maximum change of the signatures in each sample, which is simply the difference between maximum and minimum activity of the signature. To assess whether a signature change is statistically significant, we perform the following procedure. We permute the mutations in each sample and run the trajectory estimation on the permuted set. Since permuted mutations are not sorted in time, we expect no

change in the activity trajectories over time. The maximum activity change that we observe on permuted set of mutations does not exceed 5% in any sample. Therefore, we will consider signature changes below 5% to be insignificant (Fig. 2).

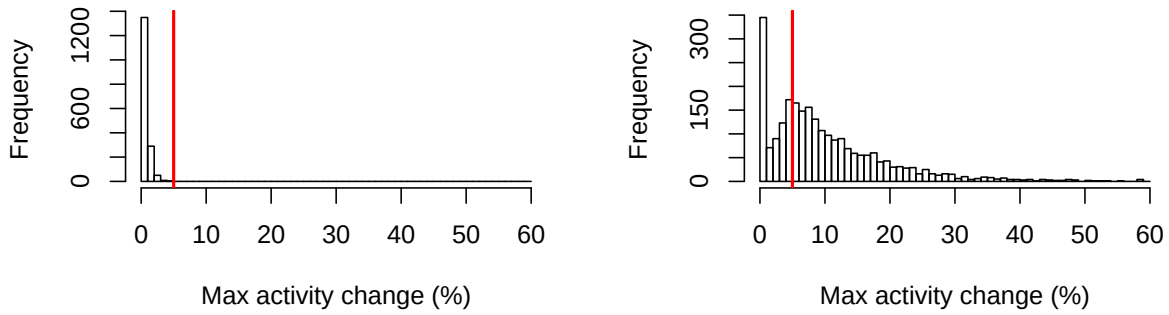


Figure 2. Distribution of maximum signature activity changes across 2486 PCAWG samples. The red line shows the threshold of 5%, below which we consider changes to be insignificant. **(a)** Changes on random orders of mutations where we don't expect to see change in activities. **(b)** activity changes in TrackSig trajectories across all samples (on mutations sorted by CCF). Frequency axis shows the number of samples where we observe the certain activity change.

As shown by Figure 3, samples typically have only two or three signatures with high activities. These signatures are usually the most variable (up to 87.2% max change, 12% on average). Other signature have low activity and remain constant. On average 3.6% of overall activity is explained by low-activity signatures (with activity <5%). Low-activity signatures most likely appear due to the uncertainty of our signature activity estimates. As mentioned in section 3.4, a mean standard deviation of signature activities of 2.9%, thus, we remove signatures with activity less than 5% as they within two standard deviations of 0%.

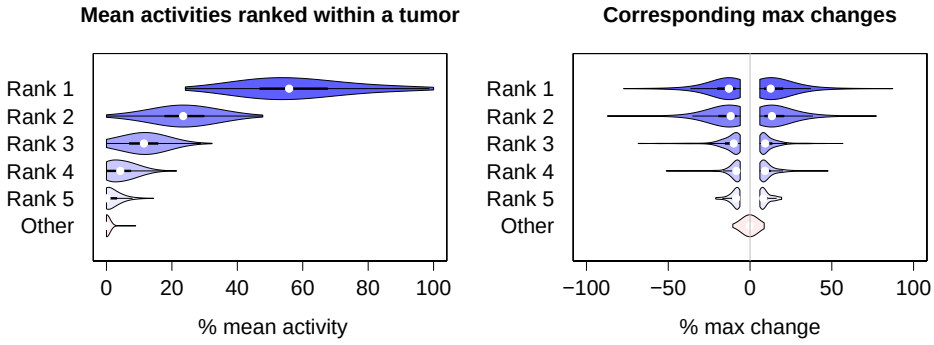


Figure 3. Left: Mean signature activities ranked from the largest to the smallest within each sample. Only the top 5 signatures with the highest activities in a sample are shown. **Right:** Maximum changes of signature activities for the corresponding signatures on the left plot. The changes below 5% are omitted as insignificant.

3.3 Trends in signature change per cancer type

The majority of samples have a signature change: 76.1% of samples have a max change >5% in at least one signature; 48.4% of samples have change >10%. However, the number of signature changes correlates to

some extent with the number of mutations in the sample. Out of samples with less than 10 timepoints (tp) only 26.3% of samples have a change >5% compared to 80.4% across the rest of the samples (see distribution on fig. 4).

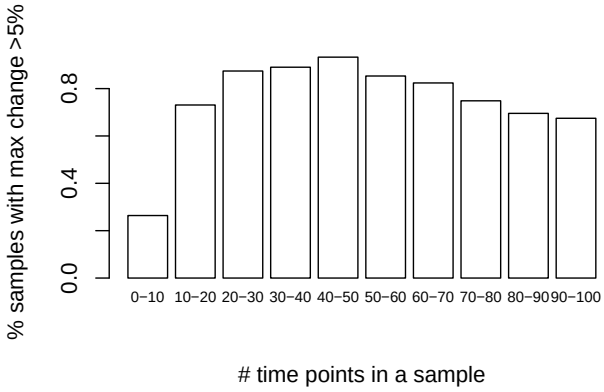


Figure 4. Proportion of tumours that have a significant change greater than 5% activity depending on the number of time points in a sample. Each bar corresponds to the range of number of time points in a sample.

3.4 Bootstrapping

We assess the variability in activity trajectories by performing bootstrap on the PCAWG data. We sample mutations with replacement from the original set and re-calculate their activities and change-points. We perform 30 bootstrap runs for each sample. Fig. 1 shows examples of bootstrapped trajectories from two samples (breast cancer and leukemia).

Signature trajectories calculated on bootstrap data are stable. The mean standard deviation of activity values calculated at each time point is 2.9%. We also evaluate the consistency of signature changes across the entire activity trajectory: size of signature change and location of the change-point. The mean standard deviation of the *signature change* is 5.3% across the bootstraps. This standard deviation does not exceed 5% in 55.8% of samples (does not exceed 10% in 94.3% of samples, fig. C.1).

In TrackSig the number of change-points is calculated during activity fitting does vary across bootstrap samples. We observe 1.02 standard deviation in the number of change-points. To assess the variability in the location of the change-points, we matched nearby change-points between bootstrap samples and measured their average distance in CCF. Because the number of change-points can change between samples, as a reference, we randomly choose one of the samples that has a number of change-points equal to the median number of change-points among all samples. Then, in all other bootstrap runs, we match each change-point to the closest run in the reference. We found that location of the change-points is consistent across bootstraps: on average, change-points are located 0.093 CCF apart from the closest reference change-point.

3.5 Simulations

To test TrackSig's ability to reconstruct the activity trajectories, we generate a set of simulated samples with known ground truth. Simulations have 50 time points (average number of time points in PCAWG samples). Each simulation has four active signatures. Two of those signatures are 1 and 5, which are nearly always active in the PCAWG samples. For the remaining two signatures, we test all possible combinations of the other 46 signatures. Thus, we have 1035 (= 46 choose 2) different signature combinations.

We generate simulations with 0 to 3 change-points that are placed randomly on the timeline. For each segment on the timeline, we sample signature activities from a symmetric Dirichlet distribution with all

parameters $\alpha_i = 1$, in other words, all activity vectors are equally likely. Finally, we sample 100 mutations from the discrete distribution derived using the sampled activities as mixing coefficients for the four signatures.

The simulations mimic the input from the real data (100 mutations per time point). In earlier simulations, we evaluated bin sizes up to 500, and found that 100 mutation bins provided an excellent balance between accuracy and sensitivity (data not shown).

Next, we run TrackSig on the simulated data and compare the reconstructed activity trajectories to the ground truth. We remove change-points with small change, that is, where activities of all signatures change by less than 5% in reconstructed trajectories. This threshold is derived in section 3.2 from permutation analysis.

We computed the absolute difference between predicted activities and the ground truth at each time point and take the median across all time points and all four signatures. We called this the median per simulation difference. On the simulations with no change-points, the median of these median per simulation differences is 0.7% On simulations with 1 to 3 change-points, this median increases slightly to 2%. The cumulation distribution of the median per simulation differences is shown in fig. 5.

For the PCAWG data, we report the maximum activity change (MAC) across activity trajectory[17]. The maximum change is the difference between maximum and minimum activity across all time points in a sample. We also report the direction of change (down if maximum occurs before minimum and up otherwise). Here, we evaluate TrackSig's accuracy in these estimates on the simulated data. The MAC discrepancies between the estimated and ground-truth trajectories is less than 5% in 83.2% of cases across all signatures in all simulations (fig. 5b).

To compare the direction of the activity change, we divide signatures into three categories: with decreasing activity, increasing activity and no activity change (if max change is less than 5%). The direction of maximum change is consistent in 95.2% of all signatures across all simulations.

To compute number of false positives and false negatives, we use modified criteria that accounts for the sliding-window smoothing. Specifically, we count a true positive detection if at least one of predicted change-points occur with three time points of an actual one. A false negative is when no predicted change-points are within three time points of an actual change. This criteria is identical to the one we use to evaluate whether a change-point supports a subclonal boundary[17]. We deem a predicted change-point a false positive if it occurs more than three time points away from the closest actual change-point.

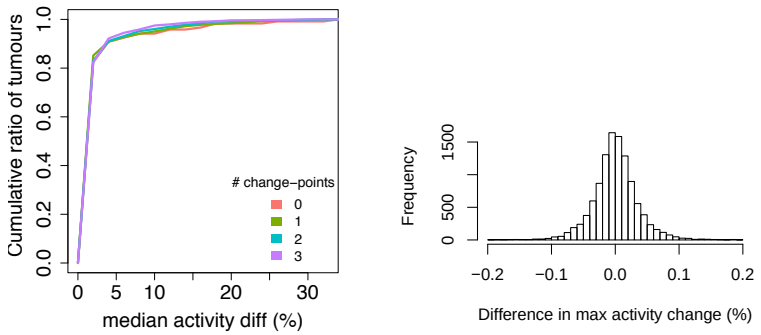
Table 1 shows the percentage of simulations where we observe the certain number false positives. On average, we observe 0.12 false positives per simulation. We detect 0.02 false negatives on average per simulation.

3.6 Choosing active signatures

Only a subset of signatures are active in a particular sample, and this subset is largely determined by a cancer type. For the analyses reported above, we use a set of active signatures provided by PCAWG, which contains a list of active signatures per sample (on average, four per sample). For COSMIC signatures the list of active signature per cancer type is available on the website. However, such data is frequently unavailable. Here we explore different ways to select the active signatures, comparing them to those selected by PCAWG-Signature group. Note that is all the approaches described below it is sufficient to fit the signatures to overall mutation counts without separating mutations into time points. Once active signatures are selected, they can be used to compute the activity trajectories across time.

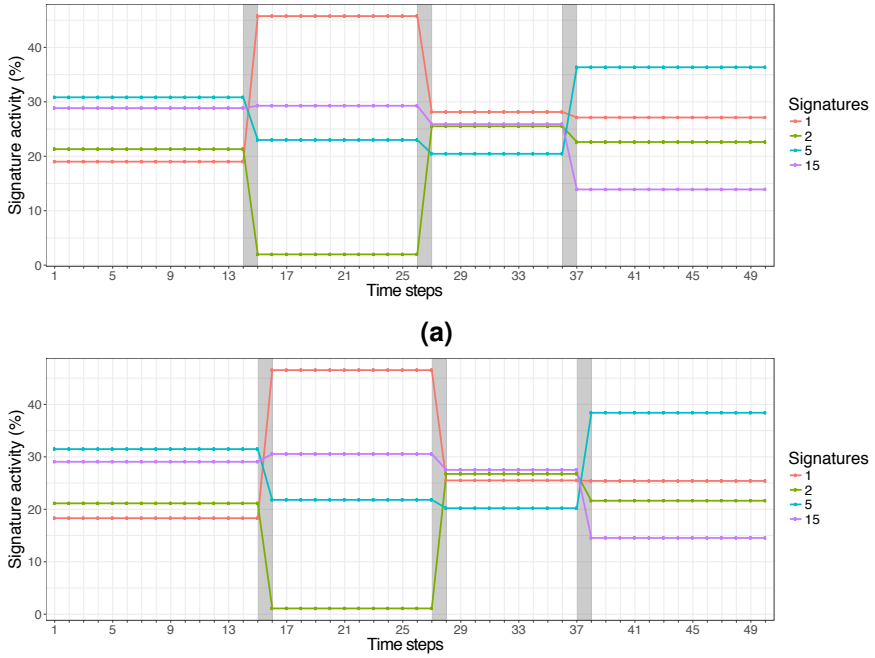
TrackSig supports three ways to determine active signatures. The first is to simply use the full set of signatures to fit the data. The second is to use all signatures reported as active in the cancer type under consideration. The final is to fit activities using one of the two previous methods, and use as active signatures only those signatures with activities greater than a threshold (by default, 5%) in the initial fit. We evaluated each strategy by compared the active signatures selected by TrackSig with those reported by PCAWG-Signature group.

For the first strategy, we used all 48 signatures and we found on average, 44.7% of overall activity



(a) **(b)**

Figure 5. Simulation results. **(a)** Median activity difference between the reconstructed trajectories and the ground truth. Lines correspond to the simulations with 0, 1, 2 or 3 change-points. The median in computed across all signatures and time points in the sample. **(b)** Distribution of MAC discrepancies between between estimated activities and ground truth.



(a)

(b)

Figure 6. Example of the simulation with three change-points. **(a)** Reconstructed sample **(b)** Ground truth

assigned by TrackSig is assigned to the active signatures selected by PCAWG-Signature group. Each incorrect signatures gets 1.3% of activity on average. In other words, the incorrect activity is widely distributed among the signatures. Therefore, we recommend constraining the number of signatures by one of the approaches described below.

Fitting only the cancer-specific signatures improves the correspondence to 68.7% of the total activity on average. Using sample-specific sets reduced the initial set of potentially active signatures from 48 down to 12 on average (ranging from 4 signatures in Lower Grade Glioma to 24 signatures in Liver Cancer). In this case, we observe that signature 5 and 40 are the most prevalent among the incorrect signatures, having the average

Table 1. Simulation results. False positives and false negatives versus change-points in the ground truth. Each cell shows the percentage of simulations that have certain number of false positives/negatives (normalized within the column). See main text for definition of positive and negative time points. The last row of the table shows the average number of false positives per simulation.

		# true change-points			
		0	1	2	3
# false positive change-points (FP)	0	0.909	0.896	0.9	0.889
	1	0.06	0.083	0.087	0.106
	2	0.024	0.019	0.011	0.005
	3	0.005	0.002	0.002	0
	4	0.002	0	0.001	0
Avg # of FP per simulation		0.130	0.128	0.118	0.116

# false negative change-points (FP)	0	1	0.992	0.962	0.947
	1	0	0.008	0.038	0.049
	2	0	0	0	0.003
Avg # of FN per simulation		0.0	0.008	0.038	0.058

activity of 14% and 12.6% respectively in the samples where they are supposed to be inactive.

Finally, we fit activity for all 48 signatures and then re-fit only those with the high activity (for instance, >5%), we exactly recover the active signatures reported by PCAWG-Signature group.

Fitting either per-cancer or per-sample signatures results in more activity mass to be on the correct signatures and speeds up the computations. Therefore, we recommend choosing per-cancer or per-sample signatures instead of using activities from the full set.

4 Summary and Conclusions

TrackSig reconstructs the evolutionary trajectories of mutation signature activities by sorting point mutations according to their inferred CCF and then partitioning this sorted list into groups of mutations with constant signature activities. TrackSig estimates uncertainty in the location of the change-points using bootstrap. TrackSig is designed to be applied to VAF data on SNVs from a single sample, however, it can be applied to sorted lists of point mutations derived from subclonal reconstruction algorithms.

Change-points often correspond to boundaries between subclones[17]. By reconstructing changes in mutation activities, TrackSig can potentially help identify DNA damage repair processes disrupted in the cell and, in doing so, help inform treatment[11].

4.1 Relationship to previous work

Previous approaches estimate signature activities for a group of mutations without considering their timing (e.g. deconstructSigs[13]). Therefore, the attempts to compare activity changes across evolutionary history have relied on pre-specified groups of mutations, such as those occurring before or after whole genome duplications[7,9,19,24] or those classified as clonal or subclonal[1,9]. The approaches mentioned above are limited to 1) the samples where the certain events have occurred or to 2) the ability of other methods to reconstruct subclonal structure of the tumour. The number of time point bins remains restricted to the number of subclones (only 2.6% of our samples have more than 2 subclones).

TrackSig uses the distributions of mutation types to group mutations. Compared to previous approaches, TrackSig allows to look at the timeline tumour development at greater resolution, where the number of time points increases with the number of mutations. We have shown that this leads to more sensitive detection of changes in signature activity. In particular, TrackSig can detect new subclones that are missed by VAF clustering methods[17]. We also provide a way to infer active signatures instead of fitting all signatures that are available.

Another important innovation of TrackSig is the used of CCF as a surrogate for evolutionary timing. Similar ideas have been used in human population genetics, where variant allele frequency to get relative order of mutations along the ancestral lineage[25]. In population genetics, allele frequency is calculated across individuals, while we calculate VAF across cell population within a single sample. In TrackSig we introduce a way to calculate cancer cell fraction (CCF) using VAF and use it as a timing estimate instead of VAF. We further improve the estimates by correcting them for CNAs.

4.2 Applicability to other mutation types

In TrackSig, the number of mutation types is provided as a parameter and is not fixed to 96 types. Because of this, it is straightforward to generalize TrackSig to reconstruct the activities of different mutation signatures or different mutations, so long as these mutations can be approximately ordered by their evolutionary time and each mutation can be classified into one of a fixed number of categories. In this paper, we ordered SNVs by decreasing CCF. This same strategy could be naturally extended to indels for which the infinite sites assumption is also valid. The infinite sites assumption should also be valid for structural variants (SVs) associated with well-defined breakpoints, thus permitting TrackSig to be used to track activities to recently defined SV signatures[24]. The CCFs of SVs can be estimated using the VAFs of split-reads mapping to their breakpoints[26]. Because they cover larger genomic regions, infinite sites is less valid for CNAs, although it is possible to approximately order clonal CNAs based on the inferred multiplicity of SNVs affected by them[9].

TrackSig also requires a pre-defined set of mutation signatures, each of which is a probability distribution over the mutation types. However, if these signatures are unavailable, they can be defined by non-negative matrix factorization, or Latent Dirichlet Allocation[27], if counts across mutation types are available from multiple cancer samples.

The alternative way to obtain more comprehensive view of tumour development is to recover evolution tree of the tumour and investigate mutations separately within each node of the tree. The root node corresponds to clonal expansion, while each of the child nodes denote the subclones. This structure can be reconstructed using a variety of algorithms including PhyloWGS[16,21,28], which builds a subclone tree based on mutation cancer cell fraction and copy number. These methods assign mutations to one of the nodes of subclonal hierarchy, allowing to analyze mutational signatures independently for each subclone.

4.3 Sensitivity to misorderings of the SNVs

TrackSig assumes that ordering SNVs by CCF recovers the order in which they accumulated in the genomes of ancestral cells, thus, our conclusions are sensitive to the correctness of this assumption. With a large number of SNVs, we do not expect large deviation in activity trajectories due to a small amount of uncertainty in CCF. Indeed, TrackSig's activity trajectory varied little in bootstrap samples. For this same reason, we do not expect activity trajectories to be impacted if a small fractions of SNVs violate the infinite sites assumption due to high, regional mutation rates.

However, these trajectories can be impacted by incorrect ordering of a large numbers of SNVs. These can occur in two ways. First, misordering can occur if a CNA changes the number of SNV allele's per cell. For example, daughter cells can fail to inherit SNVs in their mother cells due to a loss of heterozygosity (LOH). If a CNA reconstruction is available, TrackSig will correct for any detected clonal LOH when ordering SNVs, and will not attempt to order SNVs in regions affected by subclonal CNAs, thereby resolving this difficulty. However, if a CNA reconstruction is not available, or it is inaccurate, the accuracy of the activity trajectories

can suffer.

Second, SNV ordering can be incorrect when a single sample contains SNVs from subclones from different branches of the cancer phylogeny. In these circumstances, there is not a single linear order for the activities, and furthermore late occurring subclones on a different branch can have higher CCF than earlier ones occurring in the sample. However, in lung cancer, for example, few biopsies contain SNVs from branching subclones[19].

Note that a subclone can only be misordered if its CCF is less than 50% due to the Pigeonhole Principle[1], so the ordering by CCFs in guaranteed to be correct up until 50% CCF. Furthermore, if there is a change in signatures in the misordered subclone that is not reflected in the minor branch, misordering due to branching could be diagnosed by the presence of oscillations in the activity trajectories. To address this issue, when assessing overall change in signature activity, we computed the difference between the lowest and highest activities for each signature. This difference will be consistent regardless of ordering. If a phylogeny was available, one could use the phylogeny rather than CCF to order the mutations and run TrackSig separately on each branch.

4.4 Accounting for overall mutation rate

The timelines reconstructed by TrackSig are computed with a fixed number of mutations in each bin. If overall rate of generating mutations in tumour was constant, our timeline would correspond to the real time. However, tumour mutation rate often accelerates throughout development[29,30]. Although the changing rate does not affect our analysis, the estimates of the pseudo-time might not be linearly related to real time.

Estimating changes in overall mutation rate is difficult. A possible way to correct for this is to adjust the time line based on activities of signatures 1 and 5. It was suggested that signatures 1 and 5 operate as cell "clock" as the number of mutations contributed by these signatures is be proportional to the age of the individual[6]. However, it requires additional data, such as tumour samples from the same patient at different time steps and the medical history when these samples were taken. Determining the association between our pseudo-time estimates and real time is left for further investigation.

Our method TrackSig provides further insight how signature profile changes throughout tumour development. We show that through signatures analysis we can detect major events in tumour evolution, notably, transitions to a new subclone. Mutational signatures provide a unique way to recover tumour evolution path, track activities of mutational processes, adjust the treatment strategy and detect changes in therapy response.

Acknowledgments

We thank Pan-cancer Analysis of Whole Genomes (PCAWG) network, and in particular the PCAWG Evolution and Heterogeneity working group, for providing data, analysis and valuable input on this project. We would in particular like to highlight Peter Van Loo, Clemency Jolly, Stefan Dentro, David Wedge, Paul Boutros, Lydia Liu, and Moritz Gerstung who provided valuable feedback during the development of the TrackSig methodology. We would like to acknowledge SciNet as part of Compute Canada for providing computational resources. This research was partially supported by an NSERC operating grant to QDM and is part of the University of Toronto's Medicine by Design initiative, which receives funding from the Canada First Research Excellence Fund (CFREF).

References

1. Nik-Zainal, S. *et al.* The Life History of 21 Breast Cancers. *Cell* **149**, 994–1007 (2012). URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412005272. DOI 10.1016/j.cell.2012.04.023.

2. Alexandrov, L. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3776390/. DOI 10.1038/nature12477.

3. Hainaut, P. & Pfeifer, G. P. Patterns of p53 G–>T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. *Carcinogenesis* **22**, 367–374 (2001).

4. Pfeifer, G. P., You, Y.-H. & Besaratinia, A. Mutations induced by ultraviolet light. *Mutation Research* **571**, 19–31 (2005). DOI 10.1016/j.mrfmmm.2004.06.057.

5. Pfeifer, G. P. *et al.* Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* **21**, 7435–7451 (2002). DOI 10.1038/sj.onc.1205803.

6. Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nature Genetics* **47**, 1402–1407 (2015). URL http://www.nature.com/ng/journal/v47/n12/full/ng.3441.html. DOI 10.1038/ng.3441.

7. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016). URL http://science.sciencemag.org/content/354/6312/618. DOI 10.1126/science.aag0299.

8. Behjati, S. *et al.* Mutational signatures of ionizing radiation in second malignancies. *Nature Communications* **7**, 12605 (2016). URL http://www.nature.com/ncomms/2016/160907/ncomms12605/full/ncomms12605.html. DOI 10.1038/ncomms12605.

9. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *bioRxiv* (2017). URL https://www.biorxiv.org/content/early/2017/07/11/161562. DOI 10.1101/161562. For the Evolution and Heterogeneity Working Group of the Pan-Cancer Analysis of Whole Genomes Initiative (PCAWG-11), https://www.biorxiv.org/content/early/2017/07/11/161562.full.pdf.

10. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics* **48**, 758–767 (2016). URL http://www.nature.com/ng/journal/v48/n7/full/ng.3573.html. DOI 10.1038/ng.3573.

11. Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nature Medicine* **23**, 517–525 (2017). DOI 10.1038/nm.4292.

12. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Reports* **3**, 246–259 (2013). URL http://www.cell.com/cell-reports/abstract/S2211-1247(12)00433-0. DOI 10.1016/j.celrep.2012.12.008.

13. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology* **17**, 31 (2016). URL https://doi.org/10.1186/s13059-016-0893-4. DOI 10.1186/s13059-016-0893-4.

14. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLOS Genetics* **11**, e1005657 (2015). URL http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005657. DOI 10.1371/journal.pgen.1005657.

15. Yates, L. R. *et al.* Genomic Evolution of Breast Cancer Metastasis and Relapse. *Cancer Cell* **32**, 169–184.e7 (2017). DOI 10.1016/j.ccell.2017.07.005.

16. Deshwar, A. G. *et al.* PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology* **16**, 35 (2015). URL http://dx.doi.org/10.1186/s13059-015-0602-8. DOI 10.1186/s13059-015-0602-8.

17. Dentro, S. C. *et al.* Pervasive intra-tumour heterogeneity and subclonal selection across cancer types. *To be submitted* (2017). For the Evolution and Heterogeneity Working Group of the Pan-Cancer Analysis of Whole Genomes Initiative (PCAWG-11).

18. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3922638/. DOI 10.1186/1471-2105-15-35.

19. Jamal-Hanjani, M. *et al.* Tracking the evolution of non-small-cell lung cancer. *New England Journal of Medicine* **376**, 2109–2121 (2017). DOI 10.1056/NEJMoa1616288. PMID: 28445112.

20. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature Methods* **11**, 396–398 (2014). URL http://www.nature.com/nmeth/journal/v11/n4/full/nmeth.2883.html?foxtrotcallback=true. DOI 10.1038/nmeth.2883.

21. Jiang, Y., Qiu, Y., Minn, A. J. & Zhang, N. R. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **113**, E5528–5537 (2016). DOI 10.1073/pnas.1522203113.

22. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1–38 (1977). URL http://www.jstor.org/stable/2984875.

23. Killick, R., Fearnhead, P. & Eckley, I. A. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association* **107**, 1590–1598 (2012). DOI 10.1080/01621459.2012.737745.

24. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016). URL http://www.nature.com/nature/journal/v534/n7605/full/nature17676.html. DOI 10.1038/nature17676.

25. Harris, K. & Pritchard, J. K. Rapid evolution of the human mutation spectrum. *eLife* **6** (2017). URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5435464/. DOI 10.7554/eLife.24284.

26. Cmero, M. *et al.* Svclone: inferring structural variant cancer cell fraction. *bioRxiv* (2017). URL https://www.biorxiv.org/content/early/2017/08/04/172486. DOI 10.1101/172486. https://www.biorxiv.org/content/early/2017/08/04/172486.full.pdf.

27. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003). URL http://dl.acm.org/citation.cfm?id=944919.944937.

28. Satas, G. & Raphael, B. J. Tumor phylogeny inference using tree-constrained importance sampling. *Bioinformatics* **33**, i152–i160 (2017). URL +http://dx.doi.org/10.1093/bioinformatics/btx270. DOI 10.1093/bioinformatics/btx270. /oup/backfile/content_public/journal/bioinformatics/33/14/10.1093_bioinformatics_btx270/4/btx270.pdf.

29. Alberts, B. *Essential Cell Biology*, vol. Unit 5.5 (Garland Science, 2010). URL https://books.google.com/books?id=RAwvAQAAIAAJ.

30. Wodarz, D., Newell, A. & Komarova, N. Passenger mutations can accelerate tumor suppressor gene inactivation in cancer evolution. *bioRxiv* (2017). URL https://www.biorxiv.org/

content/early/2017/10/13/202531. DOI 10.1101/202531. https://www.biorxiv.org/content/early/2017/10/13/202531.full.pdf.

31. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84 (2012). URL http://doi.acm.org/10.1145/2133806.2133826. DOI 10.1145/2133806.2133826.

A Computing activity to mutational signatures

We apply topic modeling[31] to infer signature activities.Withing the time point, we separate mutation separate into K mutation types. Mutation types relate to vocabulary in topic modeling. The types used in TrackSig are described in section 3.1. Then we use mixture of discrete distributions to infer signature activities. We describe this model below.

We represent each mutation as a K-dimensional binary vector – "one-hot-encoding" of a mutation type. "One-hot-encoding" of a mutation of type k is a binary vector where k-th component is equal to 1, and other components are zeros.We will denote $\mathbf{x}^{(n)}$ to be the "one-hot-encoding of mutation n. A sample containing N mutations is represented as a $N \times K$ binary matrix $\mathbf{X}$, where each column corresponds one mutation.

$$\mathbf{x}^{(n)} = \begin{bmatrix} 0 \\ 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{bmatrix}; \quad \mathbf{x}_k^{(n)} = \begin{cases} 1, & \text{mutation } n \text{ belongs to type } k \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

A mutation process is represented as a distribution over mutation types, known as a "*mutation signature*". We will denote signature multinomials as K-dimensional probability vectors $\boldsymbol{\mu}_i$, where $i = \{1..M\}$ is an index over signatures. Signatures are fixed and are not updated during the training.

We aim to estimate signature *activities* $\boldsymbol{\pi}$ – the proportion of mutations generated by each signature.

We will use the following notation:

K – number of mutation types
M – number of signatures
N – number of mutations
$\mathbf{x}^{(n)}$ – K-dimensional binary vector of mutation n
$x_k^{(n)}$ – k-th component of vector $\mathbf{x}^{(n)}$
$\boldsymbol{\mu_i}$ – i-th signature (K-dimensional vector)
μ_{ik} – k-th component of vector $\boldsymbol{\mu_i}$
$\boldsymbol{\pi}$ – signature activities (mixture coefficients, M-dimensional vector)
π_i – i-th component of $\boldsymbol{\pi}$ (signature activity of signature i)
z_n – signature assignment for mutation n

We represent mutation matrix $\mathbf{X}$ as a mixture of signature multinomials $\boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}$ with mixture coefficients $\boldsymbol{\pi}$:

$$\mathbf{X} \sim \text{Multinomial}(N; \sum_{i=1}^{M} \pi_i \, \boldsymbol{\mu_i}) \tag{6}$$

We denote z_n to be the signature assignment of mutation n. The probabilities of mutation n to be assigned to i-th signature are equal to the mixing coefficients:

$$p(z_n = i|\boldsymbol{\pi}) = \pi_i; \quad i \in \{1..M\} \tag{7}$$

The probability of a mutation n to be generated by signature i is given by:

$$p(\mathbf{x}^{(n)}|z_n = i, \boldsymbol{\pi}, \boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}) = \prod_{k=1}^{K} \mu_{ik}^{x_k^{(n)}}; \quad i \in \{1..M\}; \quad n \in \{1..N\} \tag{8}$$

Then log likelihood of the collection of mutations in a sample:

$$\log \mathscr{L}(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}) = \sum_{n=1}^{N} \log p(\mathbf{x}^{(n)}|\boldsymbol{\pi}, \boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}) = \sum_{n=1}^{N} \log \sum_{i=1}^{M} p(\mathbf{x}^{(n)}|z_n = i, \boldsymbol{\pi}, \boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}) p(z_n = i|\boldsymbol{\pi}) \tag{9}$$

To estimate the activities, we fit mixing coefficients $\boldsymbol{\pi}$ in each bin using Expectation-Maximization (EM) algorithm[22]. The EM algorithm iterates between updating a posterior distribution over z_n and updating an estimate of the mixing coefficients $\boldsymbol{\pi}$

We start with initializing EM algorithm with uniform mixing coefficients:

$$\pi_i^{(0)} = \frac{1}{M}; \quad i \in \{1..M\} \tag{10}$$

Then, we repeat the following E-step and M-step until the algorithm converges.

In E-step, at the t-th iteration, the posterior probabilities of mutation assignments to signatures are estimated as such:

$$p(z_n = i|\mathbf{x}^{(n)}, \boldsymbol{\pi}^{(t-1)}, \boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}) = \pi_i^{(t-1)} \prod_{k=1}^{K} \mu_{ik}^{x_k^{(n)}}; \quad i \in \{1..M\}; \quad n \in \{1..N\} \tag{11}$$

In M-step we update the estimates of the mixing coefficients:

$$\pi_i^{(t)} = \frac{1}{N} \sum_{n=1}^{N} p(z_i|\mathbf{x}^{(n)}, \boldsymbol{\pi}^{(t-1)}, \boldsymbol{\mu_1}, ..\boldsymbol{\mu_K}); \quad i \in \{1..M\} \tag{12}$$

The algorithm has converged when the value of π is updated by less than 0.001 between iterations. The resulting mixture coefficients as the activities of the mutational signatures. We show the activities as percentage for the convenience of interpretation.

B Pruned Exact Linear Time (PELT) Algorithm

We adapt Pruned Linear Exact Time (PELT)[23] algorithm to detect change points in activity trajectories given cost function (likelihood) and BIC penalty. PELT is based on dynamic programming and uses heuristics to prune the set potential change-points, thus reducing the computational time.

In this section, we will use the following notation:

T – number of time points

P – number of change-points

M – number of signatures

B.1 Locating change points

As described in 2.1.2, we separate mutations into bins 100 mutations, each of which represents one time point. Our input is the set of mutation counts across 96 types for each time point: $y_{1:T} = (y_1, \ldots, y_T)$. We aim to find P change-points, or in other words, $P+1$ segments. We denote $\tau_{1:P} = (\tau_1, \ldots, \tau_P)$ to be the boundaries for our segments, meaning each segment will contain the data points $y_{\tau_{i-1}}..y_{\tau_i}$.

Given a set of change-points we can compute the likelihood of the data the following way. We fit mutational signatures within each segment (treating all mutations within each segment as one bin) and compute the likelihood $\hat{\mathscr{L}}(y_{\tau_{i-1}}..y_{\tau_i})$ as described in A. The total likelihood is the sum of likelihoods in each segment:

$$\hat{\mathscr{L}} = \sum_{i=1}^{P+1} \mathscr{L}(y_{(\tau_{i-1}+1):\tau_i})$$

We aim to minimize the Bayesian Information Criterion (BIC):

$$BIC = -2\ln\hat{\mathscr{L}} + k \cdot \ln(T)$$

where k is the number of parameters in our model and T is the number of time points. In our case $k = (P+1)\cdot(M-1)$ as we fit $(M-1)$ signature activities in $(P+1)$ segments (recall that signature activities sum to 1).

We adapt PELT objective to minimize the BIC criterion. PELT aims to minimize sum of cost functions at each time point, while using a penalty β for each placed change-point

$$\text{minimize} \sum_{i=1}^{P+1} \mathscr{C}(y_{(\tau_{i-1}+1):\tau_i}) + \beta(P+1)$$

Intuitively, we are trying to select changepoints which result in the lowest cost (or highest likelihood) while reducing the penalty associated with adding changepoints. We set the parameters as follows to make the PELT equivalent to BIC:

$$\mathscr{C} = -2\hat{\mathscr{L}}; \quad \beta = (M-1)\ln(T)$$

TrackSig-PELT algorithm find the change-points as follows. The algorithm starts with finding partial solution in a subset of the timeline and then increases the search space until change-points are the whole timeline are located. An algorithm keeps track of the time points $R_{\tau*}$ that satisfy the pruning condition and which will be considered as potential change-points at further iterations. At each iteration τ^*, the algorithm considers adding a new change-point out of the set of available time points $R_{\tau*}$. To score a potential new change-point, the algorithm refits the activities in bins formed by a potential change-point. It finds a time point τ' with the smallest likelihood and adds it to the list of change-points cp. Then the list of available time points $R_{\tau*}$ is updated: the potential change-points are removed from further consideration if the increase in likelihood associated with this change-point does not exceed the complexity penalty β.

B.2 Pruning

PELT provides an improvement in runtime by pruning certain change-points from consideration. We prune time point t if for all $t < s < T$:

$$\mathscr{C}(y_{(t+1):s}) + \mathscr{C}(y_{(s+1):T}) + \beta \leq \mathscr{C}(y_{(t+1):T}) \tag{13}$$

Intuitively, the cost placing the last changepoint prior to T at t will always be higher than cost of placing the last changepoint prior to T at s. Given this result, we can eliminate t as a potential changepoint for all iterations of the dynamic programming algorithm as it will never be optimal going forwards.

Algorithm 1 TrackSig PELT Method (Killack and Eckley 2012)

Input: Mutation counts at each time step $(y_1, y_2, ..y_T)$

Initialize: Set $\beta = (M-1)\ln(T)$; $F(0) = 0$; $cp = \{\}$; $R_1 = \{0\}$
for $\tau^* = 1, ..., T$ **do**
    Calculate $F(\tau*) = \min_{\tau \in R_{\tau^*}}[F(\tau) + \mathscr{C}(y_{(\tau+1):\tau^*}) + \beta]$, where $\mathscr{C}(y_{(\tau+1):\tau*}) = -2\hat{\mathscr{L}}(y_{(\tau+1):\tau*})$
    Let $\tau' = \arg\min_{\tau \in R_{\tau^*}}[F(\tau) + \mathscr{C}(y_{(\tau+1):\tau^*}) + \beta]$
    Append τ' to cp
    Set $R_{\tau^*+1} = \{\tau \in R_{\tau^*} \cup \{\tau*\} : F(\tau) + \mathscr{C}(y_{\tau+1:\tau^*}) + \beta \leq F(\tau*)\}$
end for
return cp – a set of change-points

C Supplementary figures and tables

Table 2. Simulation results. Predicted change-points versus change-points in the ground truth. Each cell shows the percentage of simulations which have certain number of estimated change-points (normalized within a column). Note that due to smoothing, there might be several predicted change-points that correspond to the same change-point in the ground truth. In this case, predicted change-points have to be located no more than 3 time points away from the ground truth.

		# true change-points			
		0	1	2	3
# predicted change-points	0	0.91	0.004	0	0
	1	0.061	0.9	0.019	0.001
	2	0.024	0.078	0.898	0.037
	3	0.006	0.02	0.075	0.861
	4	0.002	0.001	0.009	0.091
	5	0	0.001	0.001	0.002

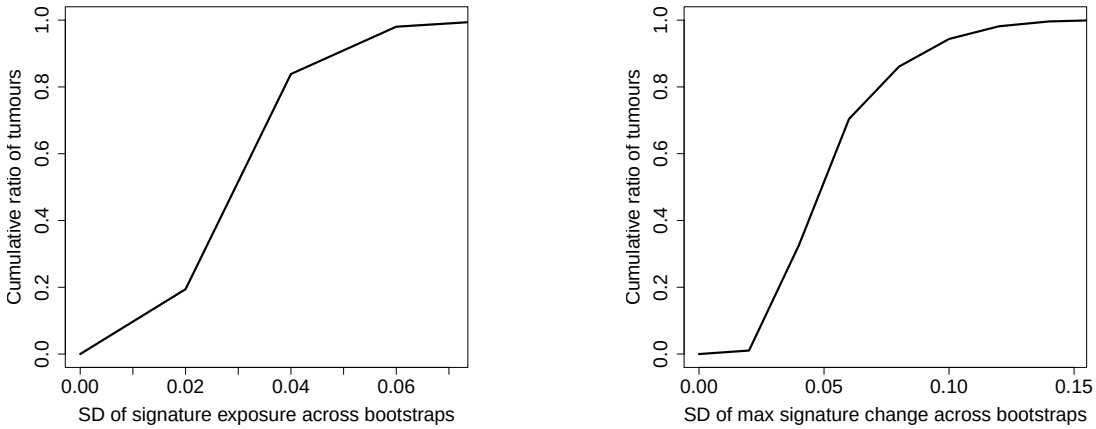


Figure C.1. Discrepancies in signature activities on bootstrap data. **(a)** Standard deviations of signature activities at each time point for each signature across bootstraps. **(b)** Standard deviations of activity *change* at each time point for each signature across bootstraps.