CloneSig: Joint inference of intra-tumor heterogeneity and signature deconvolution in tumor bulk sequencing data

Judith Abécassis[1,2,3], Fabien Reyal[1,4], Jean-Philippe Vert[5,2*]

[1] Institut Curie, PSL Research University, Translational Research Department, INSERM, U932 Immunity and Cancer, Residual Tumor & Response to Treatment Laboratory (RT2Lab), F-75248, Paris, France
[2] MINES ParisTech, PSL Research University, CBIO - Centre for Computational Biology, F-75006 Paris, France
[3] Institut Curie, PSL Research University, INSERM, U900, F-75005 Paris, France
[4] Department of Surgery, Institut Curie, F-75248 Paris, France
[5] Google Brain, F-75009 Paris, France.

Abstract

The possibility to sequence DNA in cancer samples has triggered much effort recently to identify the forces at the genomic level that shape tumorigenesis and cancer progression. It has resulted in novel understanding or clarification of two important aspects of cancer genomics: (i) intra-tumor heterogeneity (ITH), as captured by the variability in observed prevalences of somatic mutations within a tumor, and (ii) mutational processes, as revealed by the distribution of the types of somatic mutation and their immediate nucleotide context. These two aspects are not independent from each other, as different mutational processes can be involved in different subclones, but current computational approaches to study them largely ignore this dependency. In particular, sequential methods that first estimate subclones and then analyze the mutational processes active in each clone can easily miss changes in mutational processes if the clonal decomposition step fails, and conversely information regarding mutational signatures is overlooked during the subclonal reconstruction. To address current limitations, we present CloneSig, a new computational method to jointly infer ITH and mutational processes in a tumor from bulk-sequencing data, including whole-exome sequencing (WES) data, by leveraging their dependency. We show through an extensive benchmark on simulated samples that CloneSig is always as good as or better than state-of-the-art methods for ITH inference and detection of mutational processes. We then apply CloneSig to a large cohort of 8,954 tumors with WES data from the cancer genome atlas (TCGA), where we obtain results coherent with previous studies on whole-genome sequencing (WGS) data, as well as new promising findings. This validates the applicability of CloneSig to WES data, paving the way to its use in a clinical setting where WES is increasingly deployed nowadays.

1 Introduction

The advent and recent democratization of high-throughput sequencing technologies has triggered much effort recently to identify the genomic forces that shape tumorigenesis and cancer progression. In particular, they have begun to shed light on evolutionary principles happening during cancer progression, and responsible for intra-tumor heterogeneity (ITH). Indeed, as proposed by Nowell in the 1970s, cancer cells progressively accumulate somatic mutations during tumorigenesis and the progression of the disease, following similar evolutionary principles as any biological population able to acquire heritable transformations [1]. As new mutations appear in a tumor, either because they bring a selective advantage or simply through neutral evolution, some cancer cells may undergo clonal expansion until they represent the totality of the tumor or a

*Correspondance: jpvert@google.com

substantial part of it. This may result in a tumor composed of a mosaic of cell subpopulations with specific mutations. Better understanding these processes can provide valuable insights with implications in cancer detection and monitoring, patient stratification and therapeutic strategy [2, 3, 4, 5].

Bulk genome sequencing of a tumor sample allows us in particular to capture two important aspects of ITH. First, by providing an estimate of the proportion of cells harboring each single nucleotide variant (SNV), genome sequencing allows us to assess ITH in terms of presence and proportions of subclonal populations and, to some extent, to reconstruct the evolutionary history of the tumor [6, 7, 8, 9]. This estimation is challenging, both because a unique tumor sample may miss the full extent of the true tumor heterogeneity, and because the computational problem of deconvoluting a bulk sample into subclones is notoriously difficult due to noise and lack of identifiability [6, 10]. Second, beyond their frequency in the tumor, SNVs also record traces of the mutational processes active at the time of their occurrence through biases in the sequence patterns at which they arise, as characterized with the concept of mutational signature [11]. A mutational signature is a probability distribution over possible mutation types, defined by the nature of the substitution and its trinucleotide sequence context, and reflects exogenous or endogenous causes of mutations. Sixty-five such signatures have been outlined [12], and are referenced in the COSMIC database, with known or unknown aetiologies. Deciphering signature activities in a tumor sample, and their changes over time, can provide valuable insights about the causes of cancer, the dynamic of tumor evolution and driver events, and finally help us better estimate the patient prognosis and optimize the treatment strategy [2, 5]. A few computational methods have been proposed to estimate the activity of different signatures in a tumor sample from bulk genome sequencing [12, 13].

These two aspects of genome alterations during tumor development are not independent from each other. For example, if a mutation triggers subclonal expansion because it activates a particular mutational process, then new mutations in the corresponding subclone may carry the mark of this process, which may in turn be useful to identify the subclone and its associated mutations from bulk sequencing. Consequently, taking into account mutation types in addition to SNV frequencies may benefit ITH methods. Furthermore, identifying mutational processes specific to particular subclones, and in particular detecting changes in mutational processes during cancer progression, may be of clinical interest since prognosis and treatment options may differ in that case. However, current computational pipelines for ITH and mutational process analysis largely ignore the dependency between these two aspects, and typically treat them independently from each other or sequentially. In the sequential approach, as for example implemented in Palimpsest [14], subclones are first identified by an ITH analysis, and in a second step mutational signatures active in each subclone are investigated. In such a sequential analysis, however, we can not observe changes in mutational signature composition if the initial clonal decomposition step fails to detect correct subclones, and we ignore information regarding mutational signatures during ITH inference. Recently, TrackSig [15] was proposed to combine these two steps by performing an evolution-aware signature deconvolution, in order to better detect changes in signature activity along tumor evolution. However, while TrackSig overcomes the need to rely on a previously computed subclonal reconstruction, it does not leverage the possible association between mutation frequency and mutation type to jointly infer ITH and mutation processes active in the tumor. Furthermore, by design TrackSig can only work if a sufficient number of SNV is available, limiting currently its use to whole genome sequencing (WGS) data. This is an important limitation given the popularity of whole exome sequencing (WES) to characterize tumors, particularly in the clinical setting.

In this work, we propose CloneSig, the first method that leverages both the frequency and the mutation type of SNVs to jointly perform ITH reconstruction and decipher the activity of mutational signatures in each subclone. By exploiting the association between subclones and mutational processes to increase its statistical power, we show that CloneSig performs accurate estimations with fewer SNVs than competing methods, and in particular that it can be used with WES data. We show through extensive simulations that CloneSig reaches state-of-the-art performance in subclonal reconstruction and mutation deconvolution from WGS and WES data. We then provide a detailed CloneSig analysis of 8,954 pancancer WES samples from the Cancer Genome Atlas (TCGA), where we recover results coherent with a previous study on WGS [15] as well as novel promising findings of potential clinical relevance.

2 Results

2.1 Joint estimation of ITH and mutational processes with CloneSig

We propose CloneSig, a method to jointly infer ITH and estimate mutational processes active in different clones from bulk genome sequencing data of a tumor sample. The rationale behind CloneSig is illustrated in Figure 1, which shows a scatter-plot of all SNVs detected by WES in a sarcoma (TCGA patient TCGA-3B-A9HI) along two axis: horizontally, the mutation type of the SNV, and vertically, its cancer cell fraction (CCF) estimated from WES read counts. Following previous work on mutational processes [11, 12], we consider 96 possible mutation types, defined by the nature of the substitution involved and the two flanking nucleotides. Standard methods for ITH assessment and clonal deconvolution only exploit the distribution of CCF values in the sample, as captured by the histogram on the right panel of Figure 1, while standard methods for mutational signature analysis only exploit the mutation profiles capturing the distribution of mutation contexts, as represented by the histogram on the bottom panel. However, we clearly see in the scatter-plot that these two parameters are not independent, e.g., C>A mutations tend to occur frequently at low CCF, while C>T mutations occur more frequently at high CCF. CloneSig exploit this association by working directly at the 2D scatter-plot level, in order to jointly infer subclones and mutational processes involved in those subclones. Intuitively, working at this level increases the statistical power of subclone detection when subclones are better separated in the 2D scatter-plot than on each horizontal or vertical axis, i.e., when the activity of mutational processes varies between subclones.

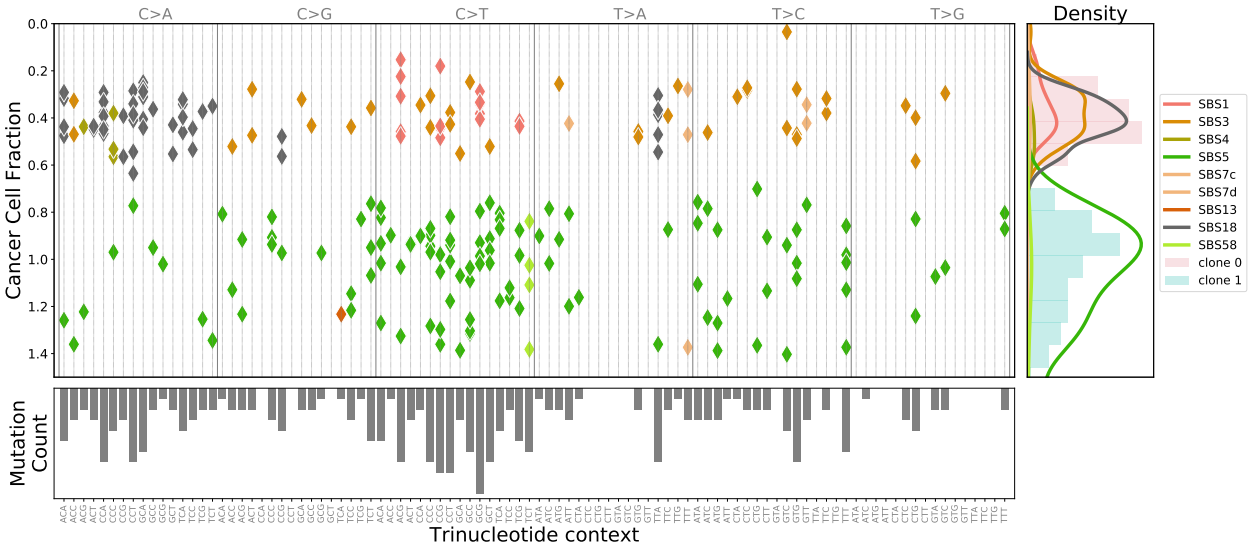


Figure 1: **CloneSig analysis of 246 SNVs obtained by WES of a sarcoma sample (patient TCGA-3B-A9HI)**. The main panel displays all SNVs in 2 dimensions: horizontally the mutation type, which describes the type of substitution together with the flanking nucleotides, and vertically the estimated CCF, as corrected by CloneSig with the estimated mutation multiplicity. From these data CloneSig infers the presence of 2 clones and a number of mutational signatures active in the different clones. Each mutation in the main panel is colored according to the most likely mutational signature according to CloneSig. On the right panel, the CCF histogram is represented and colored with estimated clones, and superimposed with mutational signature density. The bottom panel represents the total mutation type profile. The changing pattern of mutation types with CCF is clearly visible, illustrating the opportunity for CloneSig to perform joint estimation of ITH and signature activity, while most methods so far explore separately those data, considering solely the CCF histogram in the right panel for ITH analysis, or the mutation profile of the bottom panel to infer mutational processes.

More precisely, CloneSig is based on a probabilistic graphical model [16], summarized graphically in Figure 2, to model the distribution of allelic counts and trinucleotidic contexts of SNVs in a tumor. These observed variables are statistically associated through shared unobserved latent factors, including the num-

ber of clones in the tumor, the CCF of each clone, and the mutational processes active in each clone. CloneSig infers these latent factors for each tumor from the set of SNVs by maximum likelihood estimation, using standard machinery of probabilistic graphical models. Once the parameters of the model are inferred for a given tumor, we can read from them the estimated number of subclones together with their CCF, as well as the set of mutational processes active in each clone along with their strength. In addition, for each individual SNV, CloneSig allows us to estimate the clone and the signature that generated it, in a fully probabilistic manner; for example, in Figure 1, each SNV in the scatter-plot is colored according to the most likely mutational signature that generated it, according to CloneSig. Finally, we developed a likelihood ratio-based statistical test to assess whether mutational signatures significantly differ between subclones, in order to help characterize the evolutionary process involved in the life of the tumor. We refer the reader to the Material and Methods section for all technical details regarding CloneSig.

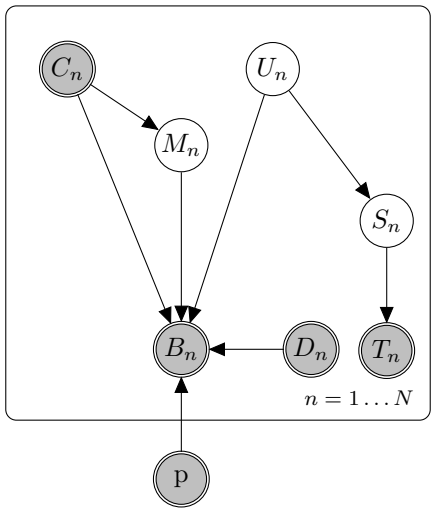


Figure 2: **Probabilistic graphical model for CloneSig**. This plot summarizes the structure of the probabilistic graphical model underlying CloneSig. Each node represents a random variable, shaded ones being observed, and edges between two nodes describe a statistical dependency encoded as conditional distribution in CloneSig. For a given tumor we observe p, the tumor purity of the sample, and for each SNV, B_n and D_n are respectively the variant and total read counts, C_n is the copy number state, and T_n is the trinucleotide context. Unobserved latent variable include U_n, the clone or subclone where the SNV occurs, S_n, the clone-dependent mutational process that generates the mutation, and M_n, the number of chromosomal copies harboring the mutation. See the main text for details about the distributions and parameters of the model.

2.2 Performance for subclonal reconstruction

We first assess the ability of CloneSig to correctly reconstruct the subclonal organization of a tumor on simulated data. To simulate data we used the probabilistic graphical model behind CloneSig with a variety of different parameters to investigate different scenarios, leading to a total of 6,300 simulations (see Materiel and Methods). For each simulation, we run CloneSig and other methods described below, and measure the correctness of the subclonal reconstruction using four different metrics adapted from [17] and described in details in the Material and Method section. Briefly, score1B measures how similar the true and the estimated number of clones are, score1C assesses in addition the correctness of frequency estimates for each subclone, score2A measures the adequacy between the true and predicted co-clustering matrices, and score2C the classification accuracy of clonal and subclonal mutations. We also assess the performance of five other state-of-the-art methods for ITH estimation and compare them to CloneSig. First we evaluate TrackSig [15], that reconstructs signature activity trajectory along tumor evolution by binning mutations in groups of 100 with decreasing CCFs, and for each group performs signature deconvolution using an

expectation-maximization (EM) algorithm. A segmentation algorithm is then applied to determine the number of breakpoints, from which we obtain subclones with different mutational processes. Because of this rationale, the authors recommend to have at least 600 observed mutations to apply TrackSig. For sake of completeness, however, we also apply TrackSig with fewer mutations in order to compare it with other methods in all settings. Second, we test Palimpsest [14], another method which associates mutational signatures and evolutionary history of a tumor. In Palimpsest, a statistical test based on the binomial distribution of variant and reference read counts for each mutation is performed, with correction for copy number, in order to classify mutations as clonal or subclonal. Then, for each of the two groups, signature deconvolution is performed using non-negative matrix factorization (NMF). This limitation to two populations can induce a bias in the metrics 1B, 1C and 2A that are inspired from [17], so we introduce the metric 2C to account for the specificity of Palimpsest. Finally, we test three popular methods for ITH reconstruction which do not model mutational processes: PyClone [7], a Bayesian clustering model optimized with a Markov Chain Monte Carlo (MCMC) algorithm, Ccube [8], another Bayesian clustering model, optimized with a variational inference method, and SciClone [18], also a Bayesian clustering model, optimized with a variational inference method, that only focuses on mutation in copy-number neutral regions.

Figures 3 summarize the performance of the different methods according to the different metrics, and under different scenarios, where we vary respectively the number of clones in the simulation (more clones should be more challenging), the number of mutations available (more mutations should help), and the percentage of diploid genome (a higher percentage should be easier). In addition, we provide in Supplementary Note S2 a more complete benchmark of the different methods when we vary as well the type of mutational signatures used as prior knowledge.

Regarding the estimation of the number of clones (score1B), CloneSig is the best method in all settings, except in the presence of 6 clones. It is in particular the only method achieving a perfect accuracy in identifying samples with one or two clones, and exhibits the best performance for score1B up to 5 clones. Both CloneSig and TrackSig see their performance decrease with the number of clones, as expected, while surprisingly Ccube has the opposite behavior and achieves better results when the number of clones is large. During the experiments we noticed that PyClone tends to find large numbers of clones with only one mutation, so we ignore these clones when we compute score1B in order not to excessively penalize PyClone for this problematic behavior. PyClone, SciClone and Palimpsest have overall a stable performance with varying numbers of clones. Regarding the impact of the number of mutations on score1B, we see that CloneSig outperforms all other methods in all settings. As expected, both CloneSig and TrackSig improve when the number of SNV increases, and we confirm that TrackSig requires at least 1,000 SNVs to be competitive with other methods in this experiment, while CloneSig reaches the best performance of TrackSig with as few as 100 SNVs. A surprising result is that for PyClone, SciClone and Ccube, score1B decreases with the number of observed mutations, which may suggest a bad calibration of the clone number estimate for large numbers of SNV; for CloneSig we designed a specific, adaptive estimator for the number of clones since we observed that standard statistical approaches for model selection perform poorly in this setting (see Material and Methods and Supplementary Section S1.2) . The percentage of diploid genome has no visible impact on the performance of any method. Regarding score1C, which focuses not on the number of clones estimated but on their ability to correctly recapitulate the distribution of CCF values, we also see that CloneSig outperforms all other methods in all settings, while PyClone and Ccube are not far behind. TrackSig performs slightly worse, especially as the number of clones increases, but this may be explained by its poor performance when the number of mutations is too low, as performance matches the other methods for 5,000 mutations. Palimpsest has comparatively a relatively poor performance, and seems particularly impacted when the proportion of diploid regions decreases. Indeed, the number of mutated copies in Palimpsest is made under the assumption that the CCF for the mutation is 1, which may jeopardize the correct detection of subclonal mutations. Finally, SciClone is clearly the worse method for score1C, particularly with 1 to 3 clones.

Besides the ability of different methods to reconstruct the correct number of subclones and their CCF, as assessed by score1B and score1C, we measure with score2A their ability to correctly assign individual mutations to their clones, an important step for downstream analysis of mutations in each subclone. According to score2A, CloneSig outperforms all other methods in all scenarios, illustrating the improved accuracy of accounting for both CCF and mutational signatures when achieving ITH reconstruction. For all methods, score2A decreases when the number of clones increases and when the percentage of diploid genomes decreases, as expected, but the relative order of methods does not change, with CloneSig followed by a group

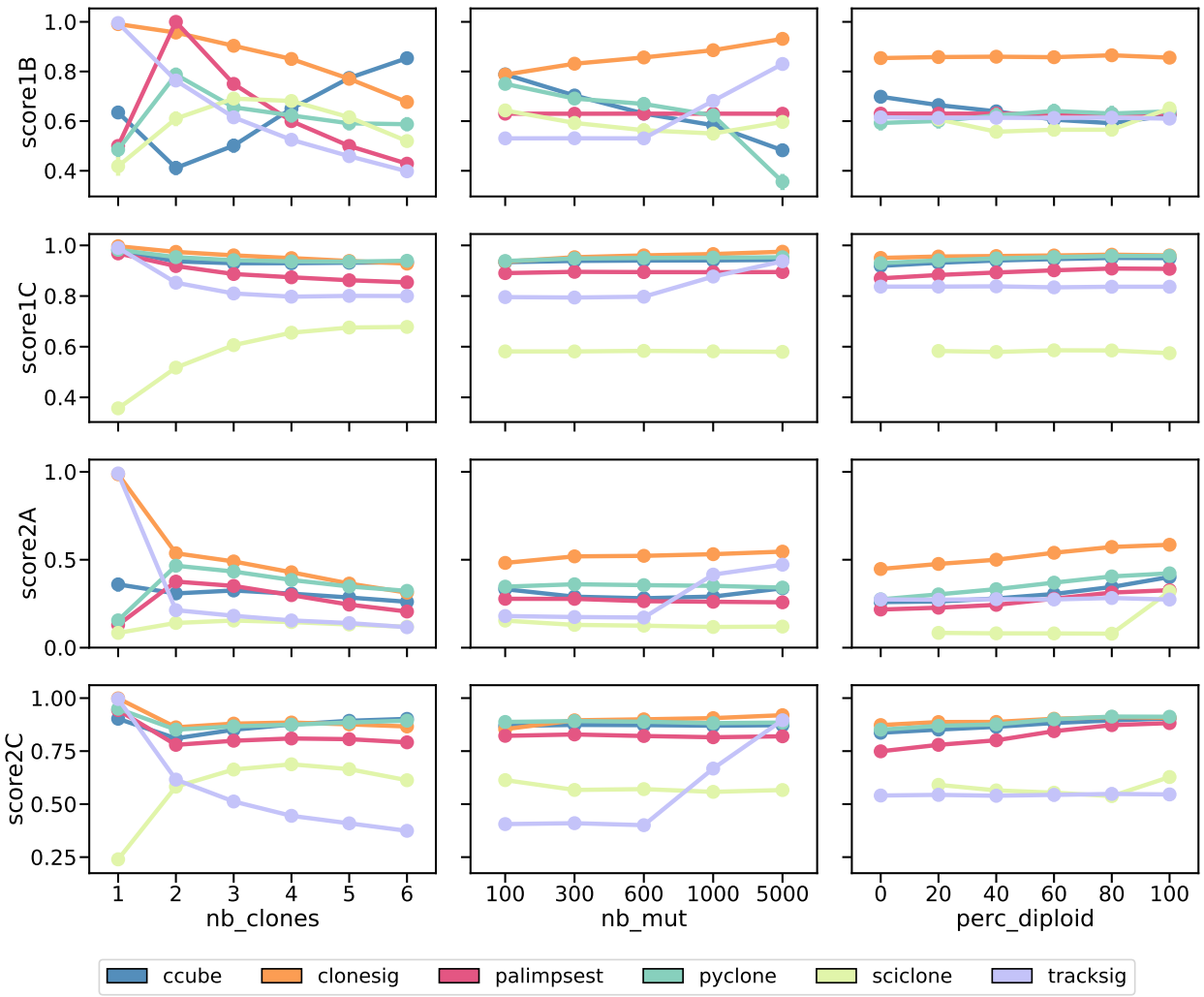


Figure 3: **Comparison of CloneSig, TrackSig, Palimpsest, PyClone, SciClone and Ccube for subclonal reconstruction.** Each row corresponds to one score, as detailed in the main text. All scores are normalized between 0 and 1, with 1 being the best and 0 the worst. Each column corresponds to a setting where one parameter in the simulation varies: the true number of clones (left), the observed number of mutations (middle), and the diploid proportion of the genome (right). Each point represents the average of the score over all available simulated samples. Bootstrap sampling of the scores was used to compute 95% confidence intervals.

of three methods with similar performances: PyClone, Ccube and Palimpsest. SciClone performs poorly except when the genome is fully diploid, in which case it gets competitive with Palimpsets but still below CloneSig, PyClone and Ccube. The number of mutations has a limited impact on the performance of all methods except for TrackSig, which only becomes competitive after 1,000 mutations. CloneSig with 100 mutations still outperforms TrackSig with 1,000 mutations, though. Finally, when we assess the capacity of each method to simply discriminate clonal from subclonal mutations using score2C, a measure meant not to penalize Palimpsets which only performs that task, we see again that CloneSig is the best in all scenarios, closely followed by Ccube and PyClone, as well as TrackSig with 5,000 mutations. Palimpsest is a bit below these methods, while SciClone and TrackSig with 1,000 mutations or less are clearly not competitive for this metric.

Overall, these experiments show that CloneSig performs as well as or better than the state-of-the-art according to all metrics considered and in all simulated scenarios, confirming that accounting for the mutation

type for each mutation, in addition to its CCF, improves the accuracy of subclonal reconstruction. We also confirm that TrackSig, the only existing method that combines CCF and mutational signature information to detect subclones, requires at least 1,000 mutations to obtain results competitive with other methods in our benchmark, while CloneSig reaches good accuracy in all scores with as few as 100 mutations.

CloneSig, like TrackSig, benefits from situations where mutational processes are not similarly active in different subclones to better detect them and assign individual mutations to them. As expected, we observe for example that the improvement of CloneSig over other methods in terms of score2A fades when there is no difference of signature activity between clones, with CloneSig performing as well as PyClone and Ccube in this situation (Supplementary Figure S12). To further illustrate the interplay between signature change and ability to detect clones, we now test CloneSig on simulations with exactly two clones, and where we vary how the clones differ in terms of CCF, on the one hand, and in terms of mutational processes, on the other hand (quantified in terms of cosine distance between the two profiles of mutation type). Figure 4 shows the accuracy of the number of clones detected by CloneSig as a function of these two parameters. We see an increased number of cases where the two clones are correctly distinguished by CloneSig as the distance between the mutation type profiles increases, for a constant CCF difference. For example, when two clones have similar signatures (small cosine distance), they can be detected with a 50% accuracy when the difference between their CCF is around 0.3; when their signatures are very different (large cosine distance), they can be detected with the same accuracy when their CCF only differ by 0.1. We show in Supplementary Figure S58 how other parameters (number of mutations, sequencing depth, diploid proportion of the genome) also impact the performance of CloneSig in this setting.

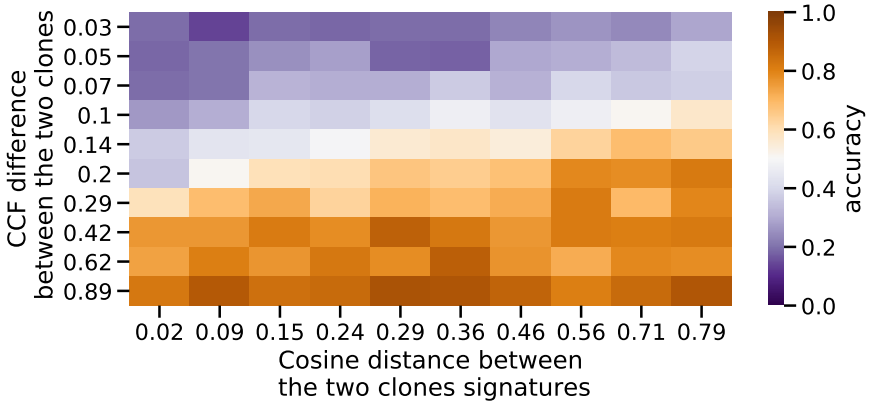


Figure 4: Accuracy of correctly estimating the presence of two clones by CloneSig as a function of the difference in the CCF between the two clones (vertical axis), and of the cosine distance between their mutational profiles. The accuracy denotes the proportion of runs where CloneSig rightfully identifies two clones.

2.3 Performance for signature deconvolution

In addition to ITH inference in terms of subclones, CloneSig estimates the mutational processes involved in the tumor and in the different subclones. We now assess the accuracy of this estimation on simulated data, using six performance scores detailed in the Material and Methods section. In short, score_sig_1A is the Euclidean distance between the normalized mutation type counts and the reconstructed profile (activity-weighted sum of all signatures); score_sig_1B is the Euclidean distance between the true and the reconstructed profile; score_sig_1C measures the identification of the true signatures; score_sig_1D is the proportion of signatures for which the true causal signature is correctly identified; and score_sig_1E reports the median of the distribution of the cosine distance between the true and the predicted mutation type profile that generated each mutation. We compare CloneSig to the two other methods that perform both ITH and mutational process estimation, namely, TrackSig and Palimpsest, and add also deconstructSigs [13] in the benchmark, a method that optimizes the mixture of mutational signature of a sample through multiple linear regressions without performing subclonal reconstruction.

Figure 5 shows the performance of the different methods according to the different metrics. For Score_sig_1A and Score_sig_1B, all methods exhibit overall similar performances, with a small advantage for CloneSig and TrackSig over Palimpsets and deconstructSigs in several scenarios. For Score_sig_1C, CloneSig and TrackSig exhibit the best AUC to detect present signatures. It may be related to a better sensitivity as CloneSig and TrackSig perform signature deconvolution in smaller subsets of mutations. All methods perform similarly with respect to Score_sig_1D, with CloneSig slightly better than all methods in all settings. The median cosine distance (Score_sig_1E) is also slightly better for CloneSig than for other methods in all settings. Surprisingly, the performance for TrackSig is worse with 5000 mutations; we observed on a few examples that this may be due to the fact that TrackSig tends to find several change points for a single clone change, due to the gradual change in activities along CCF in the overlap zone between two clones.

Overall, as for ITH inference, we conclude that CloneSig is as good as or better than all other methods in all scenarios tested. Further results where we vary other parameters in each methods, notably the set of mutations used as inputs or the set of signatures used as prior knowledge, can be found in Supplementary Note S2; they confirm the good performance of CloneSig in all settings tested.

2.4 Pan-cancer overview of signature changes

We now use CloneSig on real data, to analyze ITH and mutational process changes in a large cohort of 8,954 tumor WES samples from the TCGA spanning 31 cancer types. An overview of the main characteristics of the cohort is presented in Table S3.

For each sample in the cohort, we estimate with CloneSig the number of subclones present in the tumor, the signatures active in each subclone, and test for the presence of a signature change between clones. Figure 6 shows a global summary of the signature changes found in the cohort. For each cancer type, it shows the proportion of samples where a signature change is found, and a visual summary of the proportion of samples where each individual signature is found to increase or to decrease in the largest subclone, compared to the clonal mutations. The thickness of each bar, in addition, indicates the median change of each signature. Overall, CloneSig detects a significant change in signature activity from the protected set of mutations in 32% of all samples, and in 11% when it is trained on the public set of mutations, although these proportions vary between cancer types. In terms of signature changes, we recover patterns already observed in other cohorts, usually using WGS, which confirms that CloneSig is able to detect patterns of ITH and signature activity change using WES data. For example, similarly to the cohort of 2,778 WGS tumors analyzed by the International Cancer Genome Consortium's Pan-Cancer Analysis of Whole Genomes (PCAWG) initiative which represents the largest dataset of cancer WGS data to date [2], we observe that signature 5, of unknown aetiology, varies in almost all cancer types, and can be both increasing or decreasing. Lifestyle-associated signatures associated with tobacco-smoking (signature 4) and UV light exposure (signature 7) decrease systematically in lung tumors and oral cancers and skin melanoma respectively.

More precisely, patterns of change detected by CloneSig on the TCGA are similar to what was described on the PCAWG cohort for cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), glioblastoma multiforme (GBM), uterine carcinosarcoma (UCS) and uterine corpus endometrial carcinoma (UCEC), kidney chromophobe (KICH), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), skin cutaneous melanoma (SKCM) and stomach adenocarcinoma (STAD). In addition, CloneSig detects several new patterns of variations. In bladder carcinoma (BLCA), signature 3, related to defective homologous recombination-based DNA damage repair is found increasing. In breast cancer (BRCA), CloneSig detects three new signature variation patterns: signature 8 is increasing, and signatures 26 and 30 are varying in both directions, while signatures 1 (deamination of 5-methylcytosine to thymine) and 18 (possibly damage by reactive oxygen species) tend to be preferentially decreasing and increasing respectively, instead of varying in both directions according to [2]. In prostate adenocarcinoma (PRAD), CloneSig finds signature 3 to be varying in both direction, contrary to solely increasing in [2], but similarly to the findings of [19]. Signature 37 is found to vary in both directions instead of decreasing. Additionally to changes identified in [2], but already described in [19], CloneSig detects variations in signatures 8, 9 and 16. A new signature seems to exhibit variations along tumor evolution: signature 15 (defective DNA mismatch repair), which was not previously described in PRAD to the best of our knowledge. In lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), we observe the important increase in signature 17 as in [2], but no variation of signature 9 (mutations induced during replication by polymerase η), and an undescribed increase in signa-

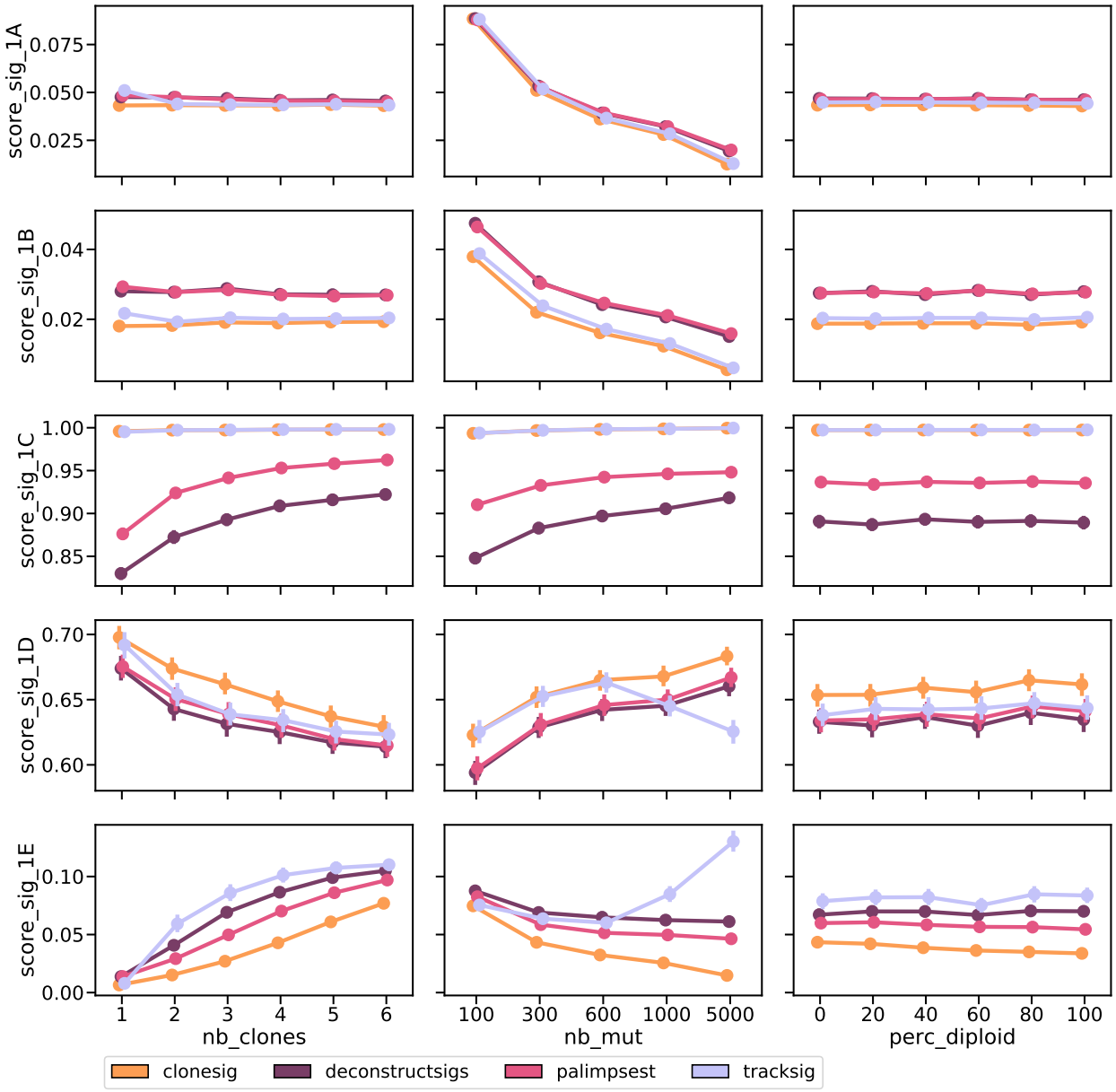


Figure 5: **Comparison of CloneSig, TrackSig, Palimpsest, and deconstructSigs for signature deconvolution.** Several metrics have been implemented, and are detailed in the main text. Scores 1A, 1B and 1E (respectively first, second and fifth rows) are distance and are better when close to 0, while scores 1C and 1D (respectively third and fourth rows) are normalized between 0 and 1 and are better when close to 1. The results are presented depending on several relevant covariates: the true number of clones (left), the number of mutations (middle), and the diploid proportion of the genome (right). Each point represents the average of the score over all available simulated samples. Bootstrap sampling of the scores was used to compute 95% confidence intervals.

tures 18 and 6 (defective DNA mismatch repair). In esophageal carcinoma (ESCA), we do not observe the important decrease of signature 17 [2], however, we describe an increase of signature 18 and a variation of signature 16 in both directions. For head-neck squamous cell carcinoma (HNSC), we observe similar patterns for signatures 5, 2 and 13 (related with APOBEC enzymes activity), and 18, but an undescribed increase of signature 3 [2], and a decrease of signature 4 (related to tobacco smoking), probably in relation to the

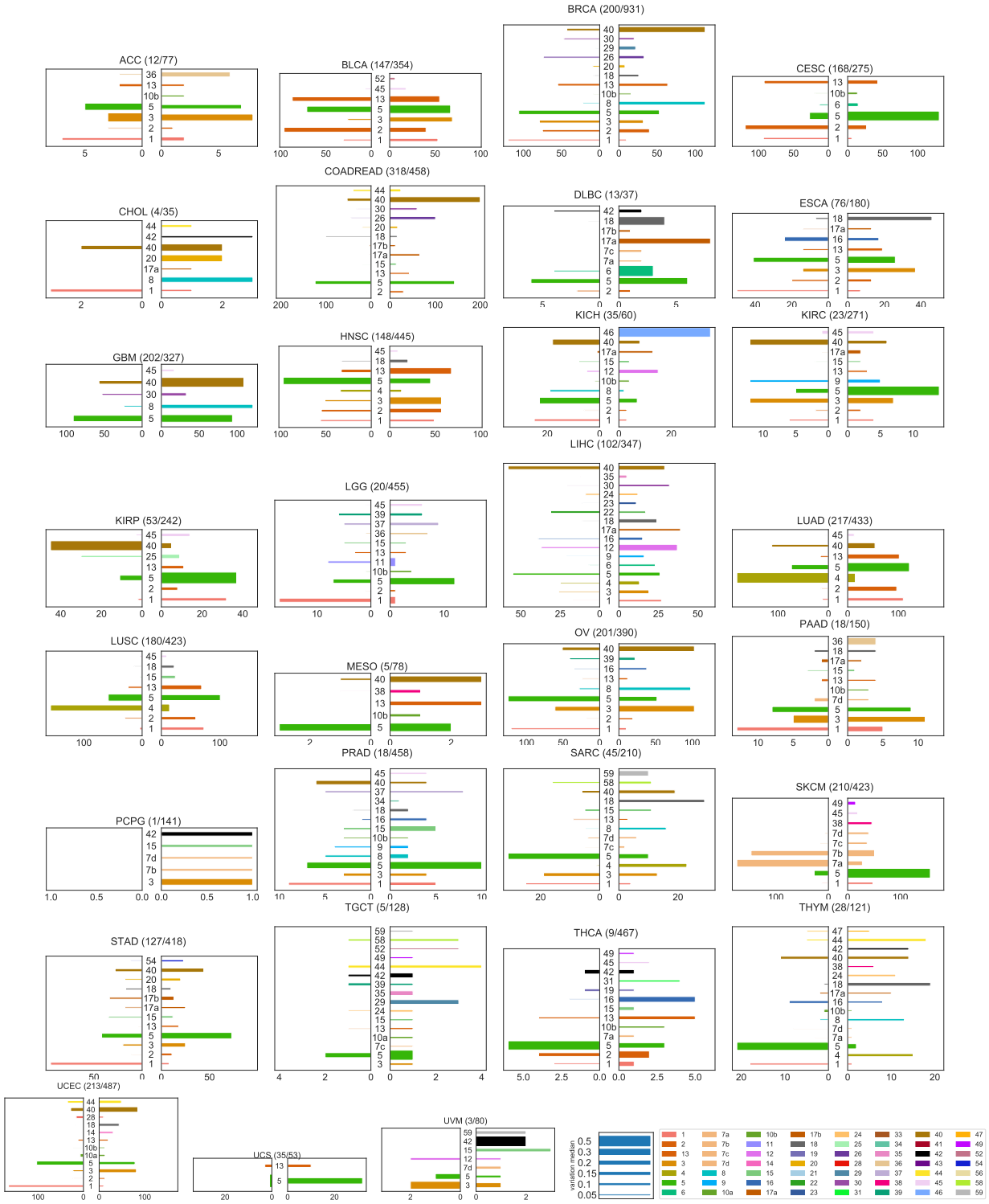


Figure 6: **Mutational signature changes in the TCGA cohort.** Each plot corresponds to one cancer type, indicates the number of samples with a significant signature change compared to the total number of samples, and shows on the right panel an increase of a signature in the largest subclone, compared to clonal mutations, and on the left panel a decrease. The length of each bar corresponds to the number of patients with such changes, and the thickness to the median observed change.

fact that this cohort includes oral tumors. In ovary tumors (OV), increase of signature 40 and decrease of signature 5 are coherent with the findings of [2], however, CloneSig finds an important number of samples with an increase of signature 8, while a decrease of this signature was reported in [2]. For thyroid carcinoma (THCA), the variations of signatures found are different, however the number of samples with a significant change of signature activity is small. In liver hepatocellular carcinoma (LIHC), and pancreatic adenocarcinoma (PAAD), we report important differences between patterns, in particular with signature 12 reported to decrease systematically in LIHC [2] while we observe an increasing trend, and no variation of signature 40 in PAAD. In colorectal cancer (COADREAD), we observe as described in [2] an strong increase in signatures 40 and 17, and a decrease in signature 18, a variation of signature 5 in both direction, and not only an increase, and no variation of signature 1. We also observe an increase in signature 26, observed in one of the three samples analyzed with single cells in [20], and an increase in signature 30 that was not previously reported.

In addition, CloneSig detects changes in signature activity in cancer types where they have not yet been characterized to the best of our knowledge, though the number of samples is too low in some cases to detect a strong trend. In adrenocortical carcinoma (ACC), we observe an increase in signature 36 (associated to defective base excision repair) and variations in signature 3. In kidney renal papillary cell carcinoma (KIRP) and kidney renal clear cell carcinoma (KIRC), signature 40 is strongly decreasing, and signature 5 increasing. Additionally CloneSig uncovers variations in signature 3 in most samples with a signature change in KIRC; activity of signature 3 in KIRC was previously outlined in [21].

2.5 Clinical relevance of ITH and signature changes

We now explore relations between the ITH detected by CloneSig and the potentially associated changes in signature activity and relevant clinical features. Looking first at the pan-cancer scale, we assess whether ITH measured either through the number of detected subclones or the presence of mutational signature changes is associated to overall survival. For that purpose, we split all TCGA samples in three groups using two different strategies, based on CloneSig's output on the protected input mutation set. In the first strategy, the three groups are based on the number of (sub-)clonal populations only (1, 2 or 3+ clones). A multivariate Cox model fitted to the data indicates for 2 clones a hazard ratio (HR) of 1.25 (95% confidence interval (CI): $[1.14, 1.37]$, $p = 2.27\text{e} - 6$), and for 3 clones a HR of 1.41 (CI= $[1.26, 1.58]$, $p = 2.03\text{e} - 9$). A univariate Cox model fitted to compare the populations with 2 or 3+ clones indicates a HR of 1.12 for 3+ clones (CI=$[1.02, 1.23]$, $p = 0.022$). This confirms that the presence of subclones as estimated by CloneSig is associated to survival, but that the difference between 2 and 3+ clones is limited in terms of survival. In the second strategy, we still keep the group of samples with only a single clone, but split the other samples (with 2 or more clonal populations) into two groups based on whether or not CloneSig detects a change in mutational signatures. The Cox results shows a HR of 1.14 without signature change (CI= $[1.04, 1.26]$, $p = 7.11\text{e} - 3$), and 1.51 with signature change (CI= $[1.37, 1.67]$, $p = 3.30\text{e} - 16$). With a focus on heterogeneous tumors only, the hazard ratio with a signature change compared to those without signature change is 1.33 (CI= $[1.22, 1.44]$, $p = 5.22\text{e}-11$). As with the first strategy, we observe a significant difference in survival between patients with homogeneous and heterogeneous tumors. However, the presence of a significant change in signature activity (second strategy) is more strongly associated to survival among heterogeneous tumors, compared to the case when we split the heterogeneous tumors based on the number of clones (Figure 7). We get similar results when using the public input mutation set (Supplementary Figure S59), illustrating CloneSig's robustness to the input signatures, and ability to detect ITH and signature activity changes with a very small number of observed mutations.

When considering the same survival analysis for each cancer type separately, we find no significant difference in survival between the different groups (homogeneous and heterogeneous tumors) after correcting for multiple tests. This may be due both to a lack of statistical power in the cancer-specific analysis because of the smaller number of samples available when we split them per cancer types, and to a confounding effect of cancer types where, for example, cancer types with a bad prognosis are enriched in heterogeneous tumors with a significant change in signature activity. Indeed, as shown in Figure 8, the proportion of tumors harboring ITH and changes in mutational processes varies a lot between cancer types. Finally, we also investigate whether patient stratification based on CloneSig output, in particular ITH and patterns of signature changes, is correlated with other clinical characteristics such as sex, age, tumor size or grade, but find overall no significant association; for sake of completeness we present detailed results of this analysis in

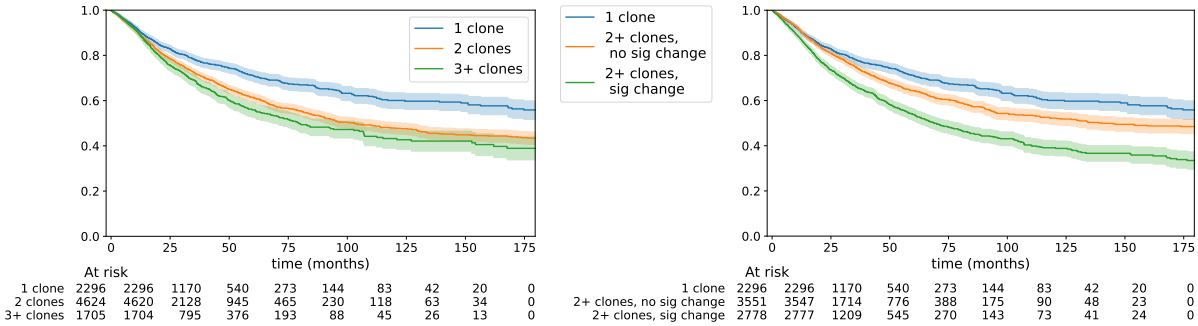


Figure 7: Kaplan-Meier curves for all TCGA samples (8,954 patients with available survival data) distinguishing tumors only along the number of clones (left) or along the number of clones and the presence of a significant change in signatures along tumor evolution (right) using the protected input mutation sets. A multivariate Cox model was fitted in both cases, and indicates for 2 clones, hazard ratio (HR) of 1.25 (95% confidence interval (CI): $[1.14, 1.37]$, $p = 2.27\text{e}-6$), and 3 clones (HR= 1.41, CI= $[1.26, 1.58]$, $p = 2.03\text{e}-9$). Considering only heterogeneous tumors, the Cox model results in a HR of 1.12 (CI=$[1.02, 1.23]$, $p = 0.022$) for 3+ clones compared to 2 clones (left). For the distinction based on signature change, without signature change (HR= 1.14, CI= $[1.04, 1.26]$, $p = 7.11\text{e}-3$), and with signature change (HR= 1.51, CI= $[1.37, 1.67]$, $p = 3.30\text{e}-16$). For heterogeneous tumors with a signature change, compared to without, the HR is 1.33 (CI= $[1.22, 1.44]$, $p = 5.22\text{e}-11$) (right)

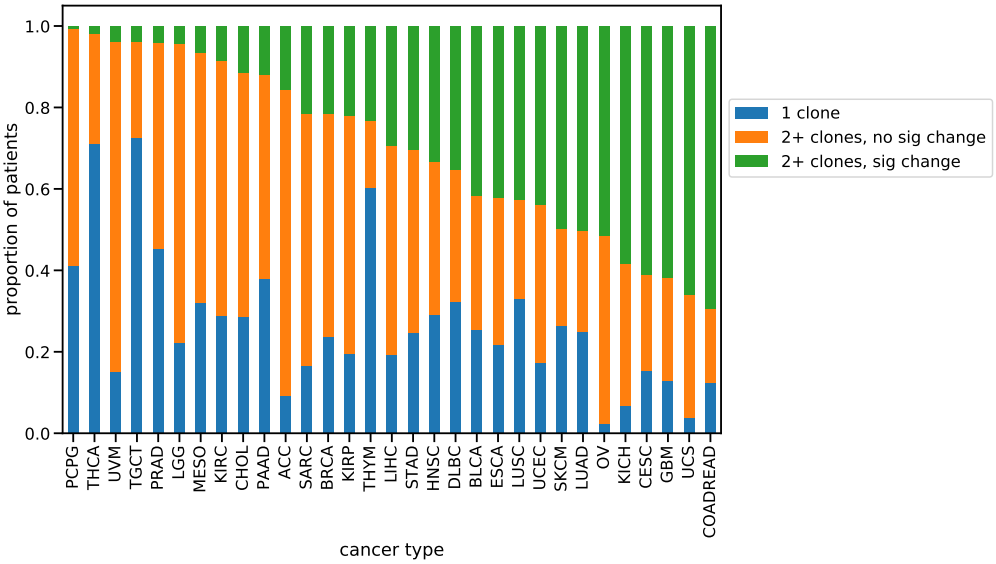


Figure 8: Proportion of patients among the three categories: "1 clone", "2 clones and more without a change in signature activity", and "2 clones and more with a change in signature activity" for the different cancer types considered in the TCGA. Cancer types are sorted from left to right by increasing proportion of heterogeneous samples with change in signature activity.

Supplementary Note S3.

3 Discussion

In recent years, a large number of methods have been developed to unravel ITH in tumors [7, 18, 8, 22, 6], and have been applied to different cohorts, including the TCGA. Recent analyses illustrate limits encountered when applying those methods to bulk WES [23, 10], as the number of observed mutations is small, the variance

in read counts can be high, and a unique sample may miss the heterogeneity of the tumor. As sequencing costs are continuously decreasing, WGS, multi-sample sequencing and single cell sequencing will constitute relevant alternatives and simplify the study of ITH. However, to date a much larger number of tumor samples with sufficient clinical annotation (in particular survival data) is available with WES compared to other more advanced technologies, and can lead to interesting insights. Beyond the number of clones present in a tumor, another relevant aspect of tumor evolution is the presence of changes in mutational signatures activities [5], which could have clinical implications in cancer prevention and treatment, and unravel the evolutionary constraints shaping early tumor development. To the best of our knowledge, TrackSig [15] and Palimpsest [14] are the only methods addressing the problem of systematic detection of signature changes, but they both present serious limitations: Palimpsest first detects ITH, and then performs signature deconvolution, which has the major drawback that if this first step fails, no signature change can be detected. Moreover, Palimpsest simply aims to distinguish subclonal from clonal mutations, thus ignoring more complex patterns. TrackSig is only applicable to WGS data, and though avoiding the caveat of relying on a previous detection of ITH, the final step of associating signature changes to the subclonal reconstruction is manual. Finally, none of these methods leverages the changes in signature activity to inform and improve the ITH detection step. To overcome these limitations, we have developed CloneSig, the first method to offer joint inference of both subclonal reconstruction and signature deconvolution, which can be applied to WGS as well as to WES data.

3.1 Improved ITH and signature detection in WES

CloneSig is a generative probabilistic graphical model that considers somatic mutations as derived from a mixture of clones where different mutational signatures are active. We demonstrated with a thorough simulation study the benefits of the joint inference in detecting ITH, both in WES and WGS samples. We showed that CloneSig is competitive with or outperforms state-of-the art ITH methods, even in the absence of signature activity change between the clones, and is particularly efficient for the detection of samples with one or a few subclones. Interestingly, several other methods we considered including PyClone [7], SciClone [18] and Ccube [8], are fully Bayesian and choose the number of clones by maximizing of the posterior probability of the data. In those methods the prior has a regularizing role, and they exhibit a decrease of accuracy as the number of observed mutations increases. This may be related to the fact that the regularizing prior is less influential as more mutations are taken into account. We instead developed a specific adaptive criterion to estimate the number of clones, as we observed that standard statistical tools for model selection performed poorly in preliminary experiments.

When applied to real data, CloneSig's results on the TCGA exhibit a strong association with survival when comparing homogeneous and heterogeneous samples. This effect on survival is stronger than the one reported in [24], also on the TCGA. This may be due to a better accuracy of CloneSig, as well as to the better statistical power of our analysis with larger sample sizes. Regarding the signature deconvolution problem, results on simulations (Score_sig_1C) suggest that CloneSig exhibits an improved sensitivity. Application to the TCGA also indicates such increased sensitivity: in the TCGA pancreatic ductal adenocarcinoma cohort (PAAD), the original study using deconstructSigs could not detect signature 3 activity in samples with somatic subclonal mutations in genes BRCA1 and BRCA2 [25], while CloneSig reports signature 3 exposure in some PAAD tumors.

3.2 Clinical relevance of signature variations

An original result of this study is the ability to further stratify heterogeneous tumors based on the presence of a significant change in signature activity, which seems associated with a worse prognosis. This could be illustrative of a more advanced stage of tumor development where a new generation of driver events supplant the initial drivers of the tumor. However, we could not reproduce those results observed on the whole TCGA cohort in a cancer-type-specific way. There are several possibilities explaining this observation: smaller cohorts may lack statistical power, or there could be a confounding effect where larger proportions of cancer types of bad prognosis are heterogeneous and have a significant change in signature activity compared to cancer types with better prognosis. Even in this latter hypothesis, this stratification can still be the manifestation of a true biological process, and not just an artifact. Indeed, other factors may explain this phenomena, like systematic later diagnosis.

To further assess the clinical relevance of signature changes, we have systematically analyzed whether we could identify an association between the exact pattern of signature change and clinical variables, but found no significant association. However, more refined or complete analyses may be necessary to uncover the full significance of signature activity changes. Previous studies report important signature activity differences between early and metastatic tumors in endometrial and breast cancers [26, 27], with impact on the survival in the breast cancer study [27]. We could not perform a similar analysis using the TCGA with only untreated primary tumors, but this constitutes new directions and opportunities of research using CloneSig on metastatic cohorts, for instance to refine findings of [27], that compares signatures deconvoluted from the whole metastasis, and could benefit from subclonal analysis to distinguish early and late mutations.

A final potential clinical application could be usage as a marker for personalized treatment. Signature 3 is associated with homologous recombination repair defect (HRD), and a targeted therapy, PARP inhibitors, can successfully target cells with such defect. A first idea is to use detection of signature 3 to identify patients that can benefit from such therapy, and CloneSig exhibits better identification of active signatures, as illustrated in the simulation studies. Indeed, several mutations in genes like BRCA1 and 2, RAD51 are known to cause HRD, but some other mutations are less frequent, or other events may result in HRD and be undetectable using regular genome sequencing, such as epigenetic inactivation [28]. In addition, the intensity of HRD mutational process may be predictive of the treatment response. Pursuing this line of thought, the change in signature activity can also be exploited as an indicator of the current driver status of HRD in tumor development. As the underlying processes of signatures will keep being uncovered, more examples of such applications are likely to arise.

3.3 Importance of input signatures and challenges

As illustrated in simulations, and based on our experience with the TCGA, the choice of the input signatures is key to CloneSig's optimal performances. This is related to the unidentifiability of the signature deconvolution problem. Several solutions have been proposed: use of a pre-defined cancer-specific matrix [12, 15], selection of signatures based on other genomic information, such as patterns of indels or structural variants, or strand biases [12], or with other molecular or clinical covariates [29]. The probabilistic framework of CloneSig is well suited to integrate other mutation types (indels, structural variants), as well as prior knowledge on signature co-occurrence, and a prior based on other molecular and clinical covariates. The difficulty of this approach is the possibility to learn such association patterns. Another direction for further development would be to use CloneSig's model to learn the signatures, or to allow some variations, as suggested in [30].

4 Materials and methods

4.1 CloneSig model

CloneSig is a probabilistic graphical framework, represented in Figure 2, to model the joint distribution of SNV frequency and mutational context using several latent variables to capture the subclonal composition of a tumor and the mutational processes involved in each clone. For a given SNV it assumes that we observe the following variables: D, the total number of reads covering the SNV; $B \leq D$, the number of mutated reads; $T \in \{1, \ldots, 96\}$ the index of the mutation type (i.e., the mutation and its flanking nucleotides, up to symmetry by reverse complement); and $C = (C_{normal}, C_{tumor}^{major}, C_{tumor}^{minor})$ the allele-specific copy number at the SNV locus, as inferred using existing tools such as ASCAT [31]. Here C_{normal} is the total copy number in normal cells, and $(C_{tumor}^{major}, C_{tumor}^{minor})$ are respectively the copy number in the cancer cells of the major and minor allele, respectively. We therefore also observe $C_{tumor} = C_{tumor}^{major} + C_{tumor}^{minor}$, the total copy number in cancer cells. Finally, we assume observed the tumor sample purity p, i.e., the fraction of cancer cells in the sample.

In addition to those observed variables, CloneSig models the following unobserved variables: $U \in \{1, \ldots, J\}$, the index of the clone where the SNV occurs (assuming a total of J clones); $S \in \{1, \ldots L\}$ the index of the mutational signature that generated the SNV (assuming a total of L possible signatures, given *a priori*); and $M \in \{1, \ldots, C_{tumor}^{major}\}$, the number of chromosomes where the SNV is present. Note that here we assume that SNVs can only be present in one of the two alleles, hence the upper bound of M by C_{tumor}^{major}.

Denoting for any integer d by $\Sigma_d = \{u \in \mathbb{R}_+^d, \sum_{i=1}^d u_i = 1\}$ the d-dimensional probability simplex, and for $u \in \Sigma_d$ by $\mathrm{Cat}(u)$ the categorical distribution over $\{1, \dots, d\}$ with probabilities $u_1, \dots, u_d$ (i.e., $X \sim \mathrm{Cat}(u)$ means that $P(X = i) = u_i$ for $i = 1, \dots, d$), let us now describe the probability distribution encoded by CloneSig for a single SNV; its generalization to several SNVs is simply obtained by assuming they are independent and identically distributed (i.i.d.) according to the model for a single SNV. We do not model the law of C and D, which are observed root nodes in Figure 2, and therefore only explicit the conditional distribution of (U, S, T, M, B) given (C, D).

Given parameters $\xi \in \Sigma_J$, $\pi \in (\Sigma_L)^J$ and $\mu \in (\Sigma_{96})^L$, we simply model U, S and T as categorical variables:

$$\begin{aligned} U &\sim \mathrm{Cat}(\xi)\,, \\ S \,|\, U &\sim \mathrm{Cat}(\pi_U)\,, \\ T \,|\, S &\sim \mathrm{Cat}(\mu_S)\,. \end{aligned}$$

Conditionally on C, we assume that the number of mutated chromosomes M is uniformly chosen between 1 and C_{tumor}^{major}, i.e.,

$$M \,|\, C \sim \mathrm{Cat}(1/C_{tumor}^{major})\,,$$

where $1/C_{tumor}^{major} \in \Sigma_{C_{tumor}^{major}}$ represents the vector of constant probability. Finally, to define the law of B, the number of mutated reads, we follow a standard approach in previous studies that represent ITH as a generative probabilistic model [7, 9, 8, 18] where the law of the mutated read counts for a given SNV must take into account the purity of the tumor, the proportion of cells in the tumor sample carrying that mutation (cancer cell fraction, CCF), as well as the various copy numbers of the normal and tumor cells. More precisely, as reviewed by [6], one can show that the expected fraction of mutated reads (variant allele frequency, VAF) satisfies

$$\mathrm{VAF} = \frac{p \times \mathrm{CCF} \times M}{p \times C_{tumor} + (1-p) \times C_{normal}}\,.$$

Note that this only holds under the classical simplifying assumption that all copy number events are clonal and affect all cells in the sample. If we now denote by $\phi \in [0,1]^J$ the vector of CCF for each clone, and introduce a further parameter $\rho \in \mathbb{R}_+^*$ to characterize the possible overdispersion of mutated read counts compared to their expected values, we finally model the number of mutated reads using a beta binomial distribution as follows:

$$\begin{aligned} B \,|\, D, U, C, M &\sim \mathrm{BetaBinomial}\left(D, \rho\phi_U\eta(M,C), \rho(1-\phi_U\eta(M,C))\right) \\ \text{with } \eta(M,C) &= \frac{p \times M}{p \times C_{tumor} + (1-p) \times C_{normal}}\,. \end{aligned}$$

4.2 Parameter estimation

Besides the tumor purity p, we assume that the matrix of mutational processes $\mu \in (\Sigma_{96})^L$ is known, as provided by databases like COSMIC and discussed below in Section 4.10. We note that we could consider μ unknown and use CloneSig to infer a new set mutational signatures from large cohorts of sequenced tumors, but prefer to build on existing work on mutational processes in order to be able to compare the results of CloneSig to the existing literature. Besides p and μ, the free parameters or CloneSig are J, the number of clones, and $\boldsymbol{\theta} = (\xi, \phi, \pi, \rho)$ which define the distributions of all random variables. On each tumor, we optimize $\boldsymbol{\theta}$ separately for $J = 1$ to $J_{max} = 8$ clones to maximize the likelihood of the observed SNV data in the tumor. The optimization is achieved approximately by an expectation-maximization (EM) algorithm [32] detailed in Supplementary Section S1.1. The number of clones $J^* \in [1, J_{max}]$ is then estimated by maximizing an adaptive model selection criterion, detailed in Supplementary Section S1.2.

4.3 Test of mutational signature changes

We use a likelihood ratio test to determine the significance of a signature change, by comparing a regular CloneSig fit to a fit with a single mixture of signatures common to all clones. To adapt the test, the parameter of the chi-squared distribution needs a calibration, that we perform on simulated data under the

null hypothesis (without change of signatures between clones). We obtain the optimal parameter using a ridge regression model with the number of clones and the degree of freedom of the input signature matrix as covariates. The coefficient values are averaged over 10-fold cross-validation to ensure robustness. We provide more details about this test in Supplementary Section S1.3.

4.4 Simulations

We use several simulation strategies to evaluate the performance of CloneSig and other methods in various situations. We also use simulations to adjust several aspects of CloneSig, in particular the setting of a custom stopping criterion and the calibration of the statistical test to detect a significant signature change along tumor evolution.

4.4.1 Default simulations

We implemented a class `SimLoader` to perform data simulation in CloneSig package. The user sets the number of clones J, the number of observed mutations N, and the matrix of L possible signatures μ. She can also specify the desired values for the CCF of each clone $\phi \in [0,1]^J$, the proportion of each clone $\xi \in \Sigma_J$, the exposure of each signature in each clone $\pi \in (\Sigma_L)^J$, and the overdispersion parameter $\rho \in \mathbb{R}^{+*}$ for the beta-binomial distribution, as well as the proportion of the genome that is diploid. If the user does not provide values for one or several parameters, we generate them randomly as follows:

π the number of active signatures follows a $Poisson(7) + 1$ distribution, and the signatures are chosen uniformly among the L available signatures. Then for each subclone, the exposures of active signatures follow a Dirichlet distribution of parameter 1 for each active signature;

ϕ the cancer cell fraction of each clone is set such that the largest clone has a CCF of 1, and each subsequent CCF is uniformly drawn in decreasing order to be greater than 0.1, and at a distance at least 0.05 from the previous clone;

ξ the proportions of clones are drawn from a Dirichlet distribution of parameter 1 for each clone. The proportions are repeatedly drawn until the minimal proportion of a clone is greater than 0.05;

ρ follows a normal distribution of mean 60 and of variance 5.

The same strategy is used for random initialization of the parameters for the EM algorithm.

The total copy number status is drawn for a user-set diploid proportion of the genome with a bell-like distribution centered in 2, and skewed towards the right (see Supplementary Figure S57 for examples), or from a rounded log-normal distribution of parameters 1 and 0.3. The minor copy number is then drawn as the rounded product between a beta distribution of parameters 5 and 3 and the total copy number. The multiplicity of each mutation n is uniformly drawn between 1 and $C_{n,tumor_{major}}$. The purity is drawn as the minimum between a normal variable of mean 0.7 and of variance 0.1, and 0.99. The other observed variables (T, B, D) are drawn according to CloneSig probabilistic model.

4.4.2 Simulations for comparison with other ITH and signature methods

To calibrate the custom stopping criterion and for further evaluation of CloneSig, we simulated $6,300$ datasets using the previously described setting, with a few adjustments: we set the minimal proportion of each clone to 0.1, the minimal difference between 2 successive clone CCFs to 0.1, and we chose the active signatures among the active signatures for each of the 35 cancer types described in the file `signatures_in_samples_and_cancer_types.mat`, extracted from the SigProfiler MATLAB package (version 2.5.1.7, downloaded from Mathworks on May 16th 2019). We draw the number of active signatures as the minimum of a $Pois(7) + 1$ distribution and the number of active signatures for this cancer type. We required a cosine distance of at least 0.05 between the mutational profiles of two successive clones.

In total, for each of the 35 cancer types, we generated a simulated sample for each combination of a number of mutations from the set $\{100, 300, 600, 1000, 5000\}$ covering the range observed in WES and WGS data, a percentage of the genome that is diploid from the set $\{0\%, 20\%, 40\%, 60\%, 80\%, 100\%\}$ to assess the impact of copy number variations, and finally, between 1 and 6 clones.

4.4.3 Simulations without signature change between clones

We generated a set of simulations similar in all points to the one for comparison with other ITH and signature methods, except that there is a unique signature mixture common to all clones. We used this dataset in two contexts: (i) to evaluate CloneSig in comparison to other methods in the absence of signature change, and (ii) to design a statistical test to assess the significance of a change in mutational signatures. For the latter, the dataset was limited to the first ten cancer types to avoid unnecessary computations.

4.4.4 Simulations to assess the separating power of CloneSig

To assess the separating power of CloneSig, we generated a dataset of 5,400 simulated tumor samples with two clones, where each clone represents 50% of the observed SNVs. Our objective was to explore the set of the distance between two clones, in terms of CCF distance, and of cosine distance between the two mutational profiles. For that purpose we first drew ten possible CCF distances evenly on a log scale between 0 and 1, and set to 1 the largest clone CCF. We also generated 30 matrices π with cosine distances covering regularly the possible cosine distances; to obtain them, we first generated 10,000 such π matrices to estimate an empirical distance distribution, and we implemented a rejection sampling strategy to obtain 30 samples from a uniform distribution. For each pair of CCF distance and π matrix, several samples were generated with the number of mutations varying among $\{100, 300, 1000\}$, the diploid proportion of the genome among $\{0.1, 0.5, 0.9\}$, and the sequencing depth among $\{100, 500\}$.

4.4.5 Simulations to assess the sensitivity of the statistical test

To measure the sensitivity of the statistical test to detect a significant signature change along tumor evolution, we generated a dataset of 2,700 simulated tumor samples with 2 to 6 clones. We used again a rejection sampling strategy to explore the space of the maximal distance between the profiles between any 2 clones, but the target distribution is here a beta distribution of parameters 1.5 and 8 as a target distribution, as the objective was to sample more thoroughly the small cosine distances. We repeated the sampling of 30 π matrices for 2 to 6 clones, and in each case, and generated several samples with the number of mutations varying among $\{100, 300, 1000\}$, the diploid proportion of the genome among $\{0.1, 0.5, 0.9\}$, and the sequencing depth among $\{100, 500\}$.

4.5 Evaluation metrics

We use several evaluation metrics to assess the quality of CloneSig and other comparable methods. Some assess specifically the accuracy of the subclonal decomposition, while others assess the performance of signature deconvolution.

4.5.1 Metrics evaluating the subclonal decomposition

The metrics described in this section evaluate the accuracy of the subclonal deconvolution. They are adapted from [17].

Score1B measures the adequacy between the true number of clones J_{true} and the estimated number of clones J_{pred}. It is computed as $\frac{J_{true}+1-\min(J_{true}+1,|J_{pred}-J_{true}|)}{J_{true}+1}$.

Score1C is the Wasserstein similarity, defined as 1 minus the Wasserstein distance between the true and the predicted clustering, defined by the CCFs of the different clones and their associated weights (proportion of mutations), implemented as the function `stats.wasserstein_distance` in the Python package `scipy`.

Score2A measures the correlation between the true and predicted binary co-clustering matrices in a vector form, M_{true} and M_{pred}. It is the average of 3 correlation coefficients:

Pearson correlation coefficient $PCC = \frac{\mathrm{Cov}(M_{true},M_{pred})}{\sigma_{M_{true}},\sigma_{M_{pred}}}$, implemented as the function `pearsonr` in the Python package `scipy`,

Matthews correlation coefficient MCC $= \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, implemented as the function `metrics.matthews_corrcoef` in the Python package `scikit-learn`,

V-measure is the harmonic mean of a homogeneity score that quantifies the fact that each cluster contains only members of a single class, and a completeness score measuring if all members of a given class are assigned to the same cluster [33]; here the classes are the true clustering. We used the function `v_measure_score` in the Python package `scikit-learn`.

Before averaging, all those scores were rescaled between 0 and 1 using the score of the minimal score between two "bad scenarios": all mutations are in the same cluster, or all mutations are in their own cluster ($M_{pred} = \mathbf{1}_{N \times N}$ or $M_{pred} = \mathbb{I}_{N \times N}$).

Score2C quantifies the accuracy of each method prediction of clonal and subclonal mutations. We report the accuracy, and the area under the ROC curve (implemented in function `metrics.roc_auc_score` in the Python package `scikit-learn`), sensitivity and specificity in Supplementary Note S2

4.5.2 Metrics evaluating the identification of mutational signatures

The metrics described in this section evaluate the accuracy of the mutational signature deconvolution.

Score_sig_1A computes the Euclidean distance between normalized mutation type counts (empirical), and the reconstituted profile. This is the objective function of most signature reconstruction approaches (including deconstructSigs[13] and Palimpsest [14]).

Score_sig_1B is the Euclidean distance between simulated and estimated signature profiles (weighted sum over all clones). This is closer to the objective of CloneSig and TrackSig [15].

Score_sig_1C measures the ability of each method to correctly identify present signatures. For CloneSig, no signature has a null contribution to the mixture, so for each clone, the signatures are considered in the decreasing order of their contribution to the mixture, and selected until the cumulative sum reaches 0.95. This rule is applied to all methods. For that metric, the area under the ROC curve (implemented in function `metrics.roc_auc_score` in the Python package `scikit-learn`) is reported, as well as the accuracy, sensitivity, and specificity in Supplementary Note S2

Score_sig_1D is the percent of mutations with the right signature. For each mutation, the most likely signature is found by taking into account the distribution of each mutation type in each signature, and the contribution of the signature to the mixture.

Score_sig_1E measures for each mutation the cosine distance between the clonal mutation type distribution that generated the mutation and the reconstituted one. We consider a unique global distribution for deconstructSigs. This allows us to measure the relevance of the reconstruction even if the wrong signatures are selected, as several signatures have very similar profiles. The result is a distribution of distances over all mutations, and we report the median of this distribution. We also report in Supplementary Note S2 more results with the minimum, the maximum, and the standard deviation of this distribution (`max_diff_distrib_mut`, `median_diff_distrib_mut`), as well as the proportions of mutations with a distance below 0.05 or 0.1 (`perc_dist_5` and `perc_dist_10`).

4.6 Implementation

CloneSig is implemented in Python, and is available as a Python package at `https://github.com/judithabk6/clonesig`. A wrapper function implements the successive optimization of CloneSig with increasing number of clones. For two clones and more, the model is initialized using results from the precedent run with one fewer clone, by splitting the subclone with the largest contribution to the mixture entropy as described in [34]. This process is stopped when the maximum number of subclones is reached, or when the selection criterion decreases for two successive runs. A class for simulating data according to the CloneSig model is also implemented, as detailed above.

4.7 Data

We downloaded data from the GDC data portal `https://portal.gdc.cancer.gov/`. We gathered annotated somatic mutations, both raw variant calling output, whose access is restricted and public mutations, from the new unified TCGA pipeline `https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/`, with alignment to the GRCh38 assembly, and variant calling using 4 variant callers: MuSe, Mutect2, VarScan2 and SomaticSniper. Instructions for download can be found in the companion Github repository (`https://github.com/judithabk6/CloneSig_analysis`).

4.8 Copy number calling and purity estimation

We obtained copy number alterations (CNA) data from the ASCAT complete results on TCGA data partly reported on the COSMIC database [31, 35]. We then converted ASCAT results on hg19 to GRCh38 coordinates using the `segment_liftover` Python package [36]. ASCAT results also provide an estimate of purity, which we used as input to ITH methods when possible. Other purity measures are available [37]; however we selected the ASCAT estimate to ensure consistency with CNV data.

4.9 Variant calling filtering

Variant calling is known to be a challenging problem. It is common practice to filter variant callers output, as ITH methods are deemed to be highly sensitive to false positive SNVs. We filtered out indels from the public dataset, and considered the union of the 4 variant callers output SNVs. For the protected data, we also removed indels, and then filtered SNVs on the `FILTER` columns output by the variant caller ("PASS" only VarScan2, SomaticSniper, "PASS" or "panel_of_normals" for Mutect2, and "Tier1" to "Tier5" for MuSe). In addition, for all variant callers, we removed SNVs with a frequency in 1000 genomes or Exac greater than 0.01, except if the SNV was reported in COSMIC. A coverage filter was added, and we kept SNVs with at least 6 reads at the position in the normal sample, of which 1 maximum reports the alternative nucleotide (or with a variant allele frequency (VAF) <0.01), and for the tumor sample, at least 8 reads covering the position, of which at least 3 reporting the variant, or a VAF>0.2. The relative amount of excluded SNVs from protected to public SNV sets varied significantly between the 3 cancer types (see Table S3). All annotations are the ones downloaded from the TCGA, using VEP v84, and GENCODE v.22, sift v.5.2.2, ESP v.20141103, polyphen v.2.2.2, dbSNP v.146, Ensembl genebuild v.2014-07, Ensembl regbuild v.13.0, HGMD public v.20154, ClinVar v.201601. We further denote the filtered raw mutation set as "Protected SNVs" and the other one, which is publicly available, as "Public SNVs"

4.10 Construction of a curated list of signatures associated with each cancer type

A very important input for CloneSig is the signature matrix. For application to the TCGA data, we restrict ourselves to signatures known to be active in each subtype. To that end, we downloaded the signatures found in the TCGA using SigProfiler [12] from synapse table syn11801497. The resulting list was not satisfactory as it lacked important known patterns; for instance signature 3, associated with homologous recombination repair deficiency was not found to be active in any tumor of the prostate cohort, while signature 3 in prostate cancer is well described in the literature [2, 19, 38]. We therefore completed the signatures present in each cancer type based on the literature [2, 39, 20, 40, 41, 42, 21, 43, 19, 44, 26, 45, 46], and used the resulting matrix in all CloneSig runs on the TCGA. Our curated list of signatures present in each cancer type is provided in Table S4.

4.11 Survival analysis

We used the Python package `lifelines` to compute the Kaplan-Meier curves and multivariate Cox models.

List of abbreviations

WES whole exome sequencing

WGS whole genome sequencing

SNV single nucleotide variant

ITH intra-tumor heterogeneity

VAF variant allele frequency

CCF cancer cell fraction

EM expectation maximization

TCGA the cancer genome atlas

ICGC international cancer genome consortium

ACC Adrenocortical carcinoma

BLCA Bladder Urothelial Carcinoma

LGG Brain Lower Grade Glioma

BRCA Breast invasive carcinoma

CESC Cervical squamous cell carcinoma and endocervical adenocarcinoma

CHOL Cholangiocarcinoma

COAD Colon adenocarcinoma

ESCA Esophageal carcinoma

GBM Glioblastoma multiforme

HNSC Head and Neck squamous cell carcinoma

KICH Kidney Chromophobe

KIRC Kidney renal clear cell carcinoma

KIRP Kidney renal papillary cell carcinoma

LIHC Liver hepatocellular carcinoma

LUAD Lung adenocarcinoma

LUSC Lung squamous cell carcinoma

DLBC Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

MESO Mesothelioma

OV Ovarian serous cystadenocarcinoma

PAAD Pancreatic adenocarcinoma

PCPG Pheochromocytoma and Paraganglioma

PRAD Prostate adenocarcinoma

READ Rectum adenocarcinoma

SARC Sarcoma

SKCM Skin Cutaneous Melanoma

STAD Stomach adenocarcinoma

TGCT Testicular Germ Cell Tumors

THYM Thymoma

THCA Thyroid carcinoma

UCS Uterine Carcinosarcoma

UCEC Uterine Corpus Endometrial Carcinoma

UVM Uveal Melanoma

References

[1] Nowell P. The clonal evolution of tumor cell populations. *Science*, 194(4260):23–28, oct 1976. ISSN 0036-8075. doi:10.1126/science.959840.

[2] Dentro S C, Leshchiner I, Haase K, Tarabichi M, Wintersinger J, Deshwar A G, Yu K, Rubanova Y, Macintyre G, Vazquez-Garcia I, Kleinheinz K, Livitz D G, et al. Portraits of genetic intra-tumour heterogeneity and subclonal selection across cancer types. Technical Report 312041, bioRxiv, 2018. doi:10.1101/312041.

[3] Sottoriva A, Kang H, Ma Z, Graham T A, Salomon M P, Zhao J, Marjoram P, Siegmund K, Press M F, Shibata D, and Curtis C. A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3):209–216, mar 2015. ISSN 1061-4036. doi:10.1038/ng.3214.

[4] Turajlic S, Xu H, Litchfield K, Rowan A, Horswell S, Chambers T, O'Brien T, Lopez J I, Watkins T B, Nicol D, Stares M, Challacombe B, et al. Deterministic evolutionary trajectories influence primary tumor growth: TRACERx Renal. *Cell*, 173(3):595–610.e11, apr 2018. ISSN 00928674. doi:10.1016/j.cell.2018.03.043.

[5] Fittall M W and Van Loo P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Medicine*, 11(1):20, dec 2019. ISSN 1756-994X. doi:10.1186/s13073-019-0632-z.

[6] Dentro S C, Wedge D C, and Van Loo P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor Perspectives in Medicine*, 7(8):a026625, aug 2017. ISSN 2157-1422. doi:10.1101/cshperspect.a026625.

[7] Roth A, Khattra J, Yap D, Wan A, Laks E, Biele J, Ha G, Aparicio S, Bouchard-Côté A, and Shah S P. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, apr 2014. ISSN 1548-7091. doi:10.1038/nmeth.2883.

[8] Yuan K, Macintyre G, Liu W, Group P E, Working H, and Markowetz F. Ccube: A fast and robust method for estimating cancer cell fractions. Technical Report 484402, bioRxiv, 2018. doi:10.1101/484402.

[9] Deshwar A G, Vembu S, Yung C K, Jang G, Stein L, and Morris Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, 16(1):35, 2015. ISSN 1465-6906. doi:10.1186/s13059-015-0602-8.

[10] Shi W, Ng C K, Lim R S, Jiang T, Kumar S, Li X, Wali V B, Piscuoglio S, Gerstein M B, Chagpar A, Weigelt B, Pusztai L, Reis-Filho J S, and Hatzis C. Reliability of whole-exome sequencing for assessing intratumor genetic heterogeneity. *SSRN Electronic Journal*, 25(6):1446–1457, 2018. ISSN 1556-5068. doi:10.2139/ssrn.3155634.

[11] Alexandrov L B, Nik-Zainal S, Wedge D C, Aparicio S a J R, Behjati S, Biankin A V, Bignell G R, Bolli N, Borg A, Børresen-Dale A L, Boyault S, Burkhardt B, et al. Signatures of mutational processes in human cancer. *Nature*, 500(7463):415–421, aug 2013. ISSN 0028-0836. doi:10.1038/nature12477.

[12] Alexandrov L B, Kim J, Haradhvala N J, Huang M N, Ng A W, Boot A, Covington K R, Gordenin D A, Bergstrom E, Lopez-Bigas N, Klimczak L J, McPherson J R, et al. The repertoire of mutational signatures in human cancer. Technical Report 322859, bioRxiv, 2018. doi:10.1101/322859.

[13] Rosenthal R, McGranahan N, Herrero J, Taylor B S, and Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, 17(1):31, dec 2016. ISSN 1474-760X. doi:10.1186/s13059-016-0893-4.

[14] Shinde J, Bayard Q, Imbeaud S, Hirsch T Z, Liu F, Renault V, Zucman-Rossi J, and Letouzé E. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*, 34(19):3380–3381, may 2018. ISSN 1367-4803. doi:10.1093/bioinformatics/bty388.

[15] Rubanova Y, Shi R, Li R, Wintersinger J, Sahin N, Deshwar A, Morris Q, PCAWG Evolution and Heterogeneity Working Group, and PCAWG network. TrackSig: reconstructing evolutionary trajectories of mutations in cancer. Technical Report 260471, BioRxiv, 2018. doi:10.1101/260471.

[16] Koller D and Friedman N. *Probabilistic Graphical Models*. MIT Press, 2009.

[17] Salcedo A, Tarabichi M, Espiritu S M G, Deshwar A G, David M, Wilson N M, Dentro S, Wintersinger J A, Liu L Y, Ko M, Sivanandan S, Zhang H, et al. Creating standards for evaluating tumour subclonal reconstruction. Technical Report 310425, bioRxiv, 2018.

[18] Miller C A, White B S, Dees N D, Griffith M, Welch J S, Griffith O L, Vij R, Tomasson M H, Graubert T A, Walter M J, Ellis M J, Schierding W, et al. SciClone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Computational Biology*, 10(8):e1003665, aug 2014. ISSN 1553-7358. doi:10.1371/journal.pcbi.1003665.

[19] Espiritu S M G, Liu L Y, Rubanova Y, Bhandari V, Holgersen E M, Szyca L M, Fox N S, Chua M L, Yamaguchi T N, Heisler L E, Livingstone J, Wintersinger J, et al. The evolutionary landscape of localized prostate cancers drives clinical aggression. *Cell*, 173(4):1003–1013.e15, may 2018. ISSN 00928674. doi:10.1016/j.cell.2018.03.029.

[20] Roerink S F, Sasaki N, Lee-Six H, Young M D, Alexandrov L B, Behjati S, Mitchell T J, Grossmann S, Lightfoot H, Egan D A, Pronk A, Smakman N, et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature*, 556(7702):457–462, apr 2018. ISSN 0028-0836. doi:10.1038/s41586-018-0024-3.

[21] Warsow G, Hübschmann D, Kleinheinz K, Nientiedt C, Heller M, Van Coile L, Tolstov Y, Trennheuser L, Wieczorek K, Pecqueux C, Gasch C, Kuru T, et al. Genomic features of renal cell carcinoma with venous tumor thrombus. *Scientific Reports*, 8(1):7477, dec 2018. ISSN 2045-2322. doi:10.1038/s41598-018-25544-z.

[22] Turajlic S, McGranahan N, and Swanton C. Inferring mutational timing and reconstructing tumour evolutionary histories. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1855(2):264–275, apr 2015. ISSN 0304419X. doi:10.1016/j.bbcan.2015.03.005.

[23] Abécassis J, Hamy A S, Laurent C, Sadacca B, Bonsang-Kitzis H, Reyal F, and Vert J P. Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data. *PLoS One*, 2019.

[24] Andor N, Graham T A, Jansen M, Xia L C, Aktipis C A, Petritsch C, Ji H P, and Maley C C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, 22(1):105–113, jan 2016. ISSN 1078-8956. doi:10.1038/nm.3984.

[25] Raphael B J, Hruban R H, Aguirre A J, Moffitt R A, Yeh J J, Stewart C, Robertson A G, Cherniack A D, Gupta M, Getz G, Gabriel S B, Meyerson M, et al. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell*, 32(2):185–203.e13, 2017. ISSN 18783686. doi:10.1016/j.ccell.2017.07.007.

[26] Ashley C W, Da Cruz Paula A, Kumar R, Mandelker D, Pei X, Riaz N, Reis-Filho J S, and Weigelt B. Analysis of mutational signatures in primary and metastatic endometrial cancer reveals distinct patterns of DNA repair defects and shifts during tumor progression. *Gynecologic Oncology*, 152(1):11–19, jan 2019. ISSN 00908258. doi:10.1016/j.ygyno.2018.10.032.

[27] Bertucci F, Ng C K Y, Patsouris A, Droin N, Piscuoglio S, Carbuccia N, Soria J C, Dien A T, Adnani Y, Kamal M, Garnier S, Meurice G, et al. Genomic characterization of metastatic breast cancers. *Nature*, 569(7757):560–564, may 2019. ISSN 0028-0836. doi:10.1038/s41586-019-1056-z.

[28] Knijnenburg T A, Wang L, Zimmermann M T, Chambwe N, Gao G F, Cherniack A D, Fan H, Shen H, Way G P, Greene C S, Liu Y, Akbani R, et al. Genomic and molecular landscape of DNA damage repair deficiency across The Cancer Genome Atlas. *Cell Reports*, 23(1):239–254.e6, apr 2018. ISSN 22111247. doi:10.1016/j.celrep.2018.03.076.

[29] Robinson W, Sharan R, and Leiserson M D M. Modeling clinical and molecular covariates of mutational process activity in cancer. *Bioinformatics*, 35(14):i492–i500, jul 2019. ISSN 1367-4803. doi:10.1093/bioinformatics/btz340.

[30] Volkova N V, Meier B, González-Huici V, Bertolini S, Gonzalez S, Abascal F, Martincorena I, Campbell P J, Gartner A, and Gerstung M. Mutational signatures are jointly shaped by DNA damage and repair. Technical Report 686295, bioRxiv, 2019. doi:10.1101/686295.

[31] Martincorena I, Raine K M, Gerstung M, Dawson K J, Haase K, Van Loo P, Davies H, Stratton M R, and Campbell P J. Universal patterns of selection in cancer and somatic tissues. *Cell*, 171(5):1029–1041.e21, nov 2017. ISSN 00928674. doi:10.1016/j.cell.2017.09.042.

[32] Dempster A P, Laird N M, and Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, sep 1977. ISSN 00359246. doi:10.1111/j.2517-6161.1977.tb01600.x.

[33] Rosenberg A and Hirschberg J. V-Measure: A conditional entropy-based external cluster evaluation measure. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 1(June):410–420, 2007.

[34] Baudry J P and Celeux G. EM for mixtures: Initialization requires special care. *Statistics and Computing*, 25(4):713–726, 2015. ISSN 15731375. doi:10.1007/s11222-015-9561-x.

[35] Forbes S A, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, Cole C G, Ward S, Dawson E, Ponting L, Stefancsik R, Harsha B, et al. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Research*, 45(D1):D777–D783, jan 2017. ISSN 0305-1048. doi:10.1093/nar/gkw1121.

[36] Gao B, Huang Q, and Baudis M. segment_liftover : a Python tool to convert segments between genome assemblies. *F1000Research*, 7:319, mar 2018. ISSN 2046-1402. doi:10.12688/f1000research.14148.1.

[37] Aran D, Sirota M, and Butte A J. Systematic pan-cancer analysis of tumour purity. *Nature Communications*, 6(1):8971, dec 2015. ISSN 2041-1723. doi:10.1038/ncomms9971.

[38] Riaz N, Blecua P, Lim R S, Shen R, Higginson D S, Weinhold N, Norton L, Weigelt B, Powell S N, and Reis-Filho J S. Pan-cancer analysis of bi-allelic alterations in homologous recombination DNA repair genes. *Nature Communications*, 8(1):857, dec 2017. ISSN 2041-1723. doi:10.1038/s41467-017-00921-w.

[39] Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X, Martincorena I, Alexandrov L B, Martin S, Wedge D C, Van Loo P, Ju Y S, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534(7605):47–54, jun 2016. ISSN 0028-0836. doi:10.1038/nature17676.

[40] Letouzé E, Shinde J, Renault V, Couchy G, Blanc J F, Tubacher E, Bayard Q, Bacq D, Meyer V, Semhoun J, Bioulac-Sage P, Prévôt S, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications*, 8(1):1315, dec 2017. ISSN 2041-1723. doi:10.1038/s41467-017-01358-x.

[41] Shibata T, Arai Y, and Totoki Y. Molecular genomic landscapes of hepatobiliary cancer. *Cancer Science*, 109(5):1282–1291, may 2018. ISSN 13479032. doi:10.1111/cas.13582.

[42] Ren W, Ye X, Su H, Li W, Liu D, Pirmoradian M, Wang X, Zhang B, Zhang Q, Chen L, Nie M, Liu Y, et al. Genetic landscape of hepatitis B virus–associated diffuse large B-cell lymphoma. *Blood*, 131(24):2670–2681, jun 2018. ISSN 0006-4971. doi:10.1182/blood-2017-11-817601.

[43] Royer-Bertrand B, Torsello M, Rimoldi D, El Zaoui I, Cisarova K, Pescini-Gobert R, Raynaud F, Zografos L, Schalenbourg A, Speiser D, Nicolas M, Vallat L, et al. Comprehensive genetic landscape of uveal melanoma by whole-genome sequencing. *The American Journal of Human Genetics*, 99(5):1190–1198, nov 2016. ISSN 00029297. doi:10.1016/j.ajhg.2016.09.008.

[44] Macintyre G, Goranova T E, De Silva D, Ennis D, Piskorz A M, Eldridge M, Sie D, Lewsley L A, Hanif A, Wilson C, Dowson S, Glasspool R M, et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nature Genetics*, 50(9):1262–1270, sep 2018. ISSN 1061-4036. doi:10.1038/s41588-018-0179-8.

[45] Liu Y, Sethi N S, Hinoue T, Schneider B G, Cherniack A D, Sanchez-Vega F, Seoane J A, Farshidfar F, Bowlby R, Islam M, Kim J, Chatila W, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell*, 33(4):721–735.e8, apr 2018. ISSN 15356108. doi:10.1016/j.ccell.2018.03.010.

[46] Verhagen C V, Vossen D M, Borgmann K, Hageman F, Grénman R, Verwijs-Janssen M, Mout L, Kluin R J, Nieuwland M, Severson T M, Velds A, Kerkhoven R, et al. Fanconi anemia and homologous recombination gene variants are associated with functional DNA repair defects in vitro and poor outcome in patients with advanced head and neck squamous cell carcinoma. *Oncotarget*, 9(26):18198–18213, apr 2018. ISSN 19492553. doi:10.18632/oncotarget.24797.

[47] Bertsekas D P. Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246, mar 1982. ISSN 0363-0129. doi:10.1137/0320018.

[48] Martinez E Z, Achcar J A, and Aragon D C. Estimação dos parâmetros da distribuição beta-binomial: Uma aplicação usando o software SAS. *Ciência e Natura*, 37(3):12–19, sep 2015. ISSN 2179-460X. doi:10.5902/2179460X17512.

[49] Schwartz G. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978. doi:10.1214/aos/1176344136.

[50] Akaike H. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, pages 199–213. Springer, 1998. doi:10.1007/978-1-4612-1694-0_15.

[51] Biernacki C, Celeux G, and Govaert G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, jul 2000. ISSN 01628828. doi:10.1109/34.865189.

[52] Maugis C and Michel B. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: Probability and Statistics*, 15:41–68, jan 2011. ISSN 1292-8100. doi:10.1051/ps/2009004.

[53] Arlot S. Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3):1–160, 2019.

[54] Neyman J and Pearson E S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 231(694-706):289–337, jan 1933. ISSN 1364-503X. doi:10.1098/rsta.1933.0009.

[55] Wilks S S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, mar 1938. ISSN 0003-4851. doi:10.1214/aoms/1177732360.

Note S1 Supplementary methods

S1.1 EM algorithm for parameter estimation

In this section we detail the EM algorithm used to estimate the parameters $\boldsymbol{\theta} = (\xi, \phi, \pi, \rho)$ of CloneSig, for a given number of clones J. To lighten notations, we use in this section the notation $M_{max_n} = (C_{tumor}^{major})_n$ for the maximum value that M_n can take. We do not model the distributions of the observed variables C_n (copy number information) and D_n (total read count), and therefore only consider the following complete conditional log-likelihood:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \log \left[\prod_{n=1}^{N} \mathbb{P}(B_n, T_n, U_n, S_n, M_n | D_n, C_n; \boldsymbol{\theta}, p, \mu) \right] \\
&= \log \left[\prod_{n=1}^{N} \prod_{u=1}^{J} \prod_{s=1}^{L} \prod_{m=1}^{M_{max_n}} (\mathbb{P}(U_n = u; \boldsymbol{\theta}) \mathbb{P}(S_n = s | U_n = u; \boldsymbol{\theta}) \mathbb{P}(T_n = t | S_n = s; \mu) \right. \\
&\quad \left. \mathbb{P}(M_n = m | C_n) \mathbb{P}(B_n | D_n, C_n, M_n = m, U_n = u; \boldsymbol{\theta}, p))^{\mathbb{I}(S_n = s, U_n = u, M_n = m)} \right] \\
&= \sum_{n=1}^{N} \sum_{u=1}^{J} \sum_{s=1}^{L} \sum_{m=1}^{M_{max_n}} \mathbb{I}(S_n = s, U_n = u, M_n = m) \log \left[\xi_u \pi_{us} \mu_{st} M_{max_n}^{-1} \, \mathrm{BB}(B_n \, ; \, D_n, \rho \phi_u \eta_{nm}, \rho(1 - \phi_u \eta_{nm})) \right] ,
\end{aligned}$$

where BB is the beta-binomial density:

$$\mathrm{BB}(k \, ; \, n, \alpha, \beta) = \binom{n}{k} \frac{\Gamma(k + \alpha) \Gamma(n - k + \beta) \Gamma(\alpha + \beta)}{\Gamma(n + \alpha + \beta) \Gamma(\alpha) \Gamma(\beta)} ,$$

and

$$\eta_{nm} = \frac{pm}{p \times (C_{tumor})_n + (1 - p) \times (C_{normal})_n} .$$

To maximize $\mathcal{L}(\boldsymbol{\theta})$, we introduce the auxiliary function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta'})$ as the expected value of the loglikelihood function of $\boldsymbol{\theta}$ when the latent variables follow the law with parameters $\boldsymbol{\theta'}$, that will be alternatively computed and maximized in the two steps of the EM algorithm. For that purpose, let us denote by $\mathbf{X}_n = (C_n, T_n, B_n, D_n)$ the set observed variables for the n-th SNV, and $\mathcal{D} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$ the totality of observed variables. Then we define:

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta'}) &= \mathbb{E}(\mathcal{L}(\boldsymbol{\theta}) | \mathcal{D}; \boldsymbol{\theta'}, p, \mu) \\
&= \sum_{n=1}^{N} \sum_{u=1}^{J} \sum_{s=1}^{L} \sum_{m=1}^{M_{max_n}} q_{nu} r_{nus} v_{mnu} \log \left[\xi_u \pi_{us} \mu_{st} M_{max_n}^{-1} \, \mathrm{BB}(B_n \, ; \, D_n, \rho \phi_u \eta_{nm}, \rho(1 - \phi_u \eta_{nm})) \right] ,
\end{aligned} \tag{1}$$

with

$$q_{nu} = \mathbb{P}(U_n = u | \mathbf{X}_n; \boldsymbol{\theta'}) , \tag{2}$$

$$r_{nus} = \mathbb{P}(S_n = s | U_n = u, \mathbf{X}_n; \boldsymbol{\theta'}) , \tag{3}$$

$$v_{mnu} = \mathbb{P}(M_n = m | U_n = u, \mathbf{X}_n; \boldsymbol{\theta'}) . \tag{4}$$

The EM algorithm iteratively builds a sequence of estimate $\boldsymbol{\theta^1}, \boldsymbol{\theta^2}, \ldots$ by solving recursively

$$\boldsymbol{\theta^i} = \mathrm{argmax}_{\boldsymbol{\theta}} \, \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}}) .$$

For that purpose, at each iteration i, the expectation (E) step first consists in computing the function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})$ with the current parameters $\boldsymbol{\theta^{i-1}}$. In other words, we must estimate the variables (2)-(4). Given the conditional independence relationships encoded in the graphical model (Figure 2), one easily gets:

$$q_{nu} = \frac{\sum_{s=1}^{L} \sum_{m=1}^{M_{max_n}} \xi_u^{i-1} \, \mathrm{BB}(B_n | D_n, \rho^{i-1} \phi_u^{i-1} \eta_{nm}^{i-1}, \rho^{i-1}(1 - \phi_u^{i-1} \eta_{nm}^{i-1})) \mu_{sT_n} \pi_{us}^{i-1}}{\sum_{u'=1}^{J} \sum_{s=1}^{L} \sum_{m=1}^{M_{max_n}} \xi_{u'}^{i-1} \, \mathrm{BB}(B_n | D_n, \rho^{i-1} \phi_{u'}^{i-1} \eta_{nm}^{i-1}, \rho^{i-1}(1 - \phi_{u'}^{i-1} \eta_{nm}^{i-1})) \mu_{sT_n} \pi_{u's}^{i-1}} , \tag{5}$$

$$r_{nus} = \frac{\mu_{sT_n} \pi_{us}^{i-1}}{\sum_{s'=1}^{M} \mu_{s'T_n} \pi_{us'}^{i-1}} , \tag{6}$$

$$v_{mnu} = \frac{\sum_{s=1}^{L} \xi_u^{i-1} \, \mathrm{BB}(B_n | D_n, \rho^{i-1} \phi_u^{i-1} \eta_{nm}^{i-1}, \rho^{i-1}(1 - \phi_u^{i-1} \eta_{nm}^{i-1})) \mu_{sT_n} \pi_{us}^{i-1}}{\sum_{s=1}^{L} \sum_{m'=1}^{M_{max_n}} \xi_u^{i-1} \, \mathrm{BB}(B_n | D_n, \rho^{i-1} \phi_u^{i-1} \eta_{nm'}^{i-1}, \rho^{i-1}(1 - \phi_u^{i-1} \eta_{nm'}^{i-1})) \mu_{sT_n} \pi_{us}^{i-1}} . \tag{7}$$

In the maximization (M) step, we compute $\boldsymbol{\theta^i}$ by plugging the estimates of the E-step (5)-(7) onto (1) and maximizing $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})$ separately for each component of $\boldsymbol{\theta}$. The maximization in ξ and π are easily obtained as:

$$\forall u \in (1 \dots J), \xi_u^i = \sum_{n=1}^{N} \frac{q_{nu}}{N},$$

$$\forall u \in (1 \dots J), \forall s \in (1 \dots L), \pi_{us}^i = \frac{\sum_{n=1}^{N} r_{nus} q_{nu}}{\sum_{n'=1}^{N} q_{n'u}}.$$

The optimization of ϕ and ρ inside the beta-binomial density term are not computable using a close formula. We therefore resort to numerical optimization and use a projected Newton method, with line search to set the Newton step at each iteration [47], in order to compute approximations of ϕ^i and ρ^i that respect constraints on their domain. Indeed, ρ must be non-negative and ϕ is a proportion so in the unit interval. For that purpose, we now compute the first and second derivatives of $\mathcal{Q}$ with respect to ϕ and $\tau = 1/\rho$:

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}}) = \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \Bigg[& \log(\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1}) + \log(\binom{d_n}{b_n}) \\ & + \log(\Gamma(b_n + \frac{\phi_u \eta_{nm}}{\tau})) + \log(\Gamma(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n)) + \log(\Gamma(\frac{1}{\tau})) \\ & - \log(\Gamma(\frac{1}{\tau} + d_n)) - \log(\Gamma(\frac{\phi_u \eta_{nm}}{\tau})) - \log(\Gamma(\frac{1-\phi_u\eta_{nm}}{\tau}))\Bigg]\end{aligned}$$

Let's now compute derivatives. ψ_0 and ψ_1 denote the digamma and trigamma functions respectively.

$$\begin{aligned}\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau} = & \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} \frac{q_{nu} v_{mnu}}{\tau^2} \Bigg[-\eta_{nm}\phi_u \psi_0(b_n + \frac{\phi_u \eta_{nm}}{\tau}) - (1-\eta_{nm}\phi_u)\psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ & -\psi_0(\frac{1}{\tau}) + \psi_0(\frac{1}{\tau} + d_n) + \eta_{nm}\phi_u \ \psi_0(\frac{\eta_{nm}\phi_u}{\tau}) + (1-\eta_{nm}\phi_u) \ \psi_0(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau^2} = & \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \Bigg[\frac{2}{\tau^3} \Bigg(\eta_{nm}\phi_u \psi_0(b_n + \frac{\phi_u\eta_{nm}}{\tau}) + (1-\eta_{nm}\phi_u)\psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ & +\psi_0(\frac{1}{\tau}) - \psi_0(\frac{1}{\tau} + d_n) - \eta_{nm}\phi_u \ \psi_0(\frac{\eta_{nm}\phi_u}{\tau}) - (1-\eta_{nm}\phi_u) \ \psi_0(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg) \\ & + \frac{1}{\tau^4} \Bigg(\eta_{nm} 2\phi_u 2\psi_1(b_n + \frac{\phi_u\eta_{nm}}{\tau}) + (1-\eta_{nm}\phi_u)2\psi_1(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) + \psi_1(\frac{1}{\tau}) \\ & -\psi_1(\frac{1}{\tau} + d_n) - \eta_{nm}2\phi_u 2 \ \psi_1(\frac{\eta_{nm}\phi_u}{\tau}) - (1-\eta_{nm}\phi_u)2 \ \psi_1(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg)\Bigg] \\ \frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u} = & \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \frac{\eta_{nm}}{\tau} \Bigg[\psi_0(b_n + \frac{\phi_u\eta_{nm}}{\tau}) - \psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ & - \ \psi_0(\frac{\eta_{nm}\phi_u}{\tau}) + \ \psi_0(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u^2} = & \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \frac{\eta_{nm}^2}{\tau^2} \Bigg[\psi_1(b_n + \frac{\phi_u\eta_{nm}}{\tau}) + \psi_1(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ & - \ \psi_1(\frac{\eta_{nm}\phi_u}{\tau}) - \ \psi_1(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u \partial \tau} = & \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \frac{\eta_{nm}}{\tau^2} \Bigg[-\psi_0(b_n + \frac{\phi_u\eta_{nm}}{\tau}) - \frac{\eta_{nm} \ \phi_u}{\tau}\psi_1(b_n + \frac{\phi_u\eta_{nm}}{\tau}) + \psi_0(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) \\ & + \frac{(1-\eta_{nm}\phi_u)}{\tau}\psi_1(\frac{1-\phi_u\eta_{nm}}{\tau} + d_n - b_n) + \psi_0(\frac{\eta_{nm}\phi_u}{\tau}) + \frac{\phi_u\eta_{nm}}{\tau}\psi_1(\frac{\eta_{nm}\phi_u}{\tau}) \\ & -\psi_0(\frac{1-\eta_{nm}\phi_u}{\tau}) - \frac{1-\phi_u\eta_{nm}}{\tau}\psi_1(\frac{1-\eta_{nm}\phi_u}{\tau})\Bigg] \\ \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u \partial \phi_{u'}} = & 0\end{aligned}$$

For sake of completeness, we provide below a second, equivalent computation using another formulation following [48].

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}}) &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \\
&\log\left[\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1} \binom{d_n}{b_n} \frac{\Gamma(b_n + \rho\phi_u\eta_{nm})\Gamma(\rho(1-\phi_u\eta_{nm}) + d_n - b_n)}{\Gamma(\rho + d_n)} \frac{\Gamma(\rho)}{\Gamma(\rho\phi_u\eta_{nm})\Gamma(\rho(1-\phi_u\eta_{nm})}\right] \\
&= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \\
&\log\left[\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1} \binom{d_n}{b_n} \frac{\prod_{i=0}^{b_n-1}(\phi_u\eta_{nm} + \frac{i}{\rho}) \prod_{i=0}^{d_n-b_n-1}(1-\phi_u\eta_{nm} + \frac{i}{\rho})}{\prod_{i=0}^{d_n-1}(1+\frac{i}{\rho})}\right] \\
&= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{s=1}^{L}\sum_{m=1}^{M_{max_n}} r_{nus} q_{nu} v_{mnu} \left[\log(\xi_u \mu_{st} \pi_{us} M_{max_n}^{-1}) + \log(\binom{d_n}{b_n}) + \sum_{i=0}^{b_n-1}\left[\log(\phi_u\eta_{nm} + \frac{i}{\rho})\right]\right. \\
&\left. + \sum_{i=0}^{d_n-b_n-1}\left[\log(1-\phi_u\eta_{nm} + \frac{i}{\rho})\right] - \sum_{i=0}^{d_n-1}\left[\log(1+\frac{i}{\rho})\right]\right]
\end{aligned}
$$

Let's set $\tau = \frac{1}{\rho}$. We are trying to compute maximum likelihood estimates for ϕ_u and τ.

$$
\begin{aligned}
\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau} &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[\sum_{i=0}^{b_n-1}\left[\frac{i}{\phi_u\eta_{nm} + i\tau}\right] + \sum_{i=0}^{d_n-b_n-1}\left[\frac{i}{1-\phi_u\eta_{nm} + i\tau}\right] - \sum_{i=0}^{d_n-1}\left[\frac{i}{1+i\tau}\right]\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \tau^2} &= \sum_{n=1}^{N}\sum_{u=1}^{J}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[-\sum_{i=0}^{b_n-1}\left[\frac{i^2}{(\phi_u\eta_{nm} + i\tau)}\right] - \sum_{i=0}^{d_n-b_n-1}\left[\frac{i^2}{(1-\phi_u\eta_{nm} + \tau)^2}\right] + \sum_{i=0}^{d_n-1}\left[\frac{i^2}{(1+\tau)^2}\right]\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u \partial \tau} &= \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[-\sum_{i=0}^{b_n-1}\left[\frac{i\eta_{nm}}{(\phi_u\eta_{nm} + i\tau)^2}\right] + \sum_{i=0}^{d_n-b_n-1}\left[\frac{i\eta_{nm}}{(1-\phi_u\eta_{nm} + i\tau)^2}\right]\right] \\
\frac{\partial \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u} &= \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[\sum_{i=0}^{b_n-1}\left[\frac{\eta_{nm}}{\phi_u\eta_{nm} + i\tau}\right] + \sum_{i=0}^{d_n-b_n-1}\left[\frac{-\eta_{nm}}{1-\phi_u\eta_{nm} + i\tau}\right]\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_u^2} &= \sum_{n=1}^{N}\sum_{m=1}^{M_{max_n}} q_{nu} v_{mnu} \left[-\sum_{i=0}^{b_n-1}\left[\frac{\eta_{nm}}{\phi_u\eta_{nm} + i\tau}\right]^2 - \sum_{i=0}^{d_n-b_n-1}\left[\frac{\eta_{nm}}{1-\phi_u\eta_{nm} + i\tau}\right]^2\right] \\
\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta^{i-1}})}{\partial \phi_{u'} \partial \phi_u} &= 0
\end{aligned}
$$

We can then plug these formulas in the projected Newton algorithm to estimate ϕ^i and ρ^i. We repeat the E and M steps until $\|\boldsymbol{\theta^i} - \boldsymbol{\theta^{i-1}}\| < 10^{-5} \times J \times L$.

S1.2 Selecting the number of clones

As explained in Supplementary Section S1.1, the EM algorithm allows us to optimize all parameters of the CloneSig model for a given number of clones J. Here we explain how to estimate J.A first idea to automatize that choice is to rely on a model selection heuristics, such as the widely used Bayesian Information Criterion (BIC) [49], an asymptotic Bayesian criterion aiming at selecting the model best supported by the data. BIC is defined as

$$BIC(J) = \ell(\mathcal{D}; \boldsymbol{\theta}_J) - \frac{D_J}{2}\log N,$$

where $\ell(\mathcal{D}; \boldsymbol{\theta}_J)$ is the maximum log-likelihood as estimated by the EM procedure with J clones, and D_J is the degree of freedom of the model; by default, we take it equal to the number of free parameters, namely, $D_J = J * (L - 1 + 2)$ for J clones, where L is the number of signatures. Indeed, for each clone, we have $L - 1$ parameters for the signature proportions (π), the frequency of the clone (ϕ_u), and the proportion of the clone ξ_u. We have to remove 1 because $\sum_{u=1}^{J} \xi_u = 1$, and add 1 for the overdispersion parameter τ.

On simulations, however, we found that while BIC correctly identifies the number of clones when the number of SNVs is large, it tends to performs poorly when the number of mutations is low (a few hundreds) in which case it

quasi systematically selects a single clone. On the other hand, when we observe the variation of the log-likelihood with the number of components J as for example in Supplementary Figure S1, we clearly see an "elbow" for some $J > 1$, suggesting that the information about J is properly captured by CloneSig's likelihood but not by BIC.

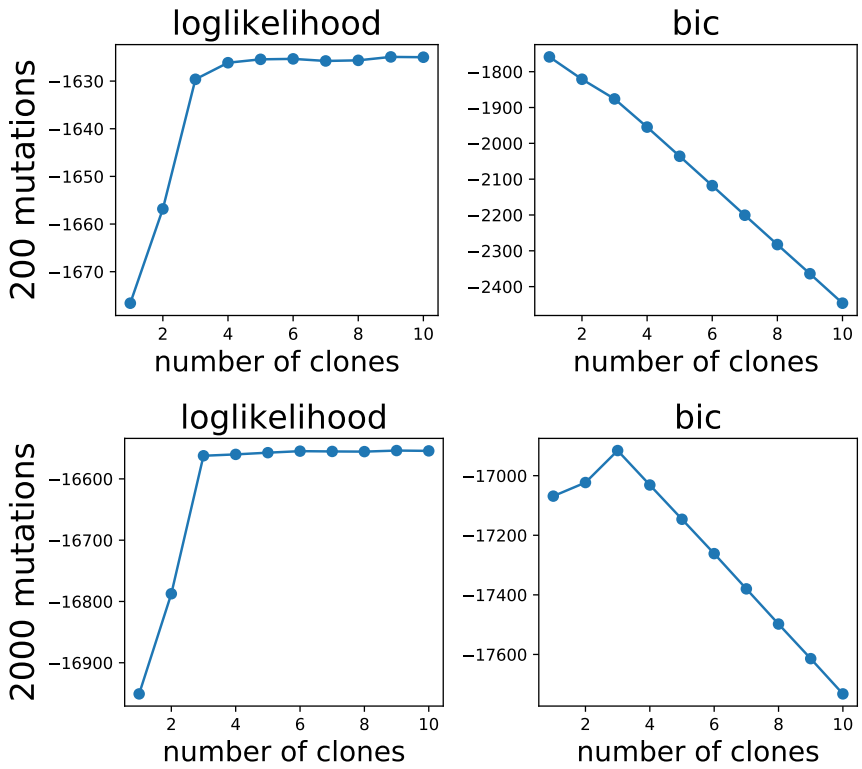


Figure S1: Evolution of the loglikelihood and BIC criterion for 2 simulated samples, with the same parameters and 200 mutations (up panels), and 2000 mutations (bottom panels). In both cases, the loglikelihood has an "elbow" at 3 clones indicating that the likelihood of the data increases much less at the addition of an additional mixture component beyond 3 components. The BIC criterion is maximal at 3 clones when 2000 mutations are observed, but at 1 clone in the case with 200 mutations.

We observed similar behaviors with other classical criteria such as the Akaike Information Criterion (AIC) [50], the Integrated Classification Likelihood (ICL) [51]), or the slope heuristics as described in [52]. This difficulty can be related to results from statistical theory of model selection and penalization suggesting that asymptotic results are known up to a factor when applied to smaller datasets [53], and therefore propose now as an alternative an empirical criterion that can be fit on data with known model, such as simulations. More precisely, we consider the following criterion:

$$BIC_\alpha(J) = \ell(\mathcal{D}; \boldsymbol{\theta}_J) - \alpha D_J \log N \,. \tag{8}$$

with $\alpha > 0$ is a free parameter to be user-defined or estimated, and D_J is a measure complexity of the model.

While we leave α as a user-defined parameter in the CloneSig software, we now propose a systematic approach to estimate it when we can simulate samples. For each simulated sample, we fit CloneSig for 1 to 10 clones. The objective is to estimate a parameter α such that $BIC_{\alpha,J}$ is maximal for the true number of clones J_{true} on all or most simulations. To achieve that, we formulate it as a standard supervised classification problem where for each simulation and each $J \neq J_{true}$, we want $BIC_\alpha(J_{true}) > BIC_\alpha(J)$; since $BIC_\alpha(J)$ is itself a linear function of α, we estimate α by minimizing a convex proxy to the number of errors, namely,

$$\min_\alpha \sum_{\mathcal{D}} \sum_{J \neq J_{true}} \phi\left(BIC_\alpha(J_{true}) - BIC_\alpha(J)\right) \,, \tag{9}$$

where $\phi(u) = \max(0, 1-u)$ is the hinge loss that pushes its argument to be larger than one when minimized; solving (9) is a simple support vector machine (SVM) problem that we solve with a standard SVM solver.

The second important aspect of (8) is D_J, that measures the complexity of the model with J clones. The original BIC penalizes the "dimension of the model" [49], that can be interpreted as the degree of freedom of the model, and

we now discuss different possible definitions for it. The parameters ϕ, ξ and ρ determining the CCFs and proportions of the different clones in the mixture must clearly be counted as in BIC. Regarding the signatures however, one can notice that the signatures are neither orthogonal (some signatures are very similar), nor independent (some signatures are associated with the same underlying biological process). Instead of just counting the number of signatures, we therefore propose to estimate the degree of freedom dof_L of the matrix with L signatures by the number of eigenvalues of the cosine similarity matrix greater than 0.5 in absolute value. As shown in Figure S2, dof_L is roughly proportional to L, at least for L up to 20. Another source of degree of freedom is the copy number. Indeed, for each observed mutation, several values of the number of mutated copies are considered, so if the maximal average multiplicity for mutations in the sample is $M_{max_{avg}}$, a unique clone CCF corresponds in average to $M_{max_{avg}}$ possible VAFs, adding some freedom to the model. We therefore consider four possible definitions for D_J, indexed with letters A to D.

$$D_J^A = J \times (L+1) \times M_{max_{avg}}\,, \tag{10}$$

$$D_J^B = J \times (L+1)\,, \tag{11}$$

$$D_J^C = J \times (\text{dof}_L + 1) \times M_{max_{avg}}\,, \tag{12}$$

$$D_J^D = J \times (\text{dof}_L + 1)\,. \tag{13}$$

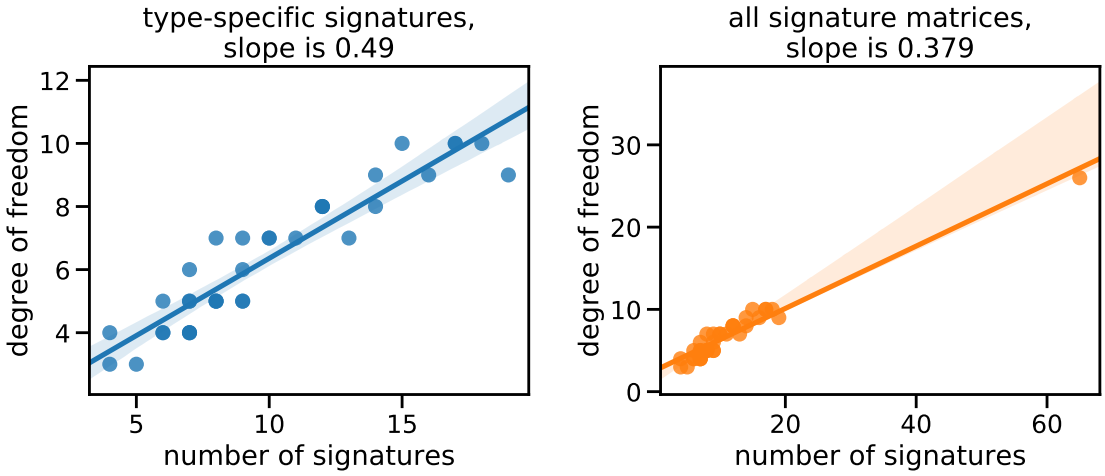


Figure S2: Variation of the degree of freedom of a subset of cancer type-specific signatures (35 distinct types) or for all available signatures depending on the number of signatures. The left panel shows the dependence for subsets of the 65 signatures only, and the right panel for the 65 signatures additionally. We see that the dependence with the number of signature (slope) is different in the two cases.

Moreover, if we consider the variations of the degree of freedom associated with L signatures, dof_L, as a function of L for the 35 available cancer types, and for the all 65 signatures, we note that there is a gap, as the maximal number of signatures in one cancer type is 19, and that the slope seems different for a subset or for all the signatures (see Supplementary Figure S2). The dependency being quite different, this raises the question of whether we should estimate a single α for all situations (i.e., a unique BIC model), or whether we should fit two BIC models: one for the cases where CloneSig is run with only cancer type-specific signatures, and one for the case where CloneSig is run with all the 65 signatures.

For each possible definition of D_J (10)-(13), and for each setting (estimating a unique or two separate BIC models), we ran simulations to estimate the value of α such that $BIC_{\alpha,j}, j \in \{1,\ldots,10\}$ is maximal for the true value of J, by solving (9). To evaluate the results, we split the dataset into a train (80% of data) and a test set (20%), and assess the accuracy of J estimation on the test set. To evaluate the stability of the learnt parameter α, we compute the 95% confidence interval over 10 independent train-test splits. The values for learnt coefficients, averaged over 10 independent train/test splits for each case are presented in Table S1. The test accuracies for different criteria and different learning settings are presented in Figure S3. We first see that, as mentioned earlier, standard model selection criteria (BIC, AIC, ICL) perform overall poorly. Second, we notice that the "separate" strategy is usually slightly better than the "full" strategy, i.e., learning a single α for CloneSig with all 65 signatures or only a subset is not as good as learning two different α's. As for the definition of D_J, we see in both cases that using the degree of freedom of the signature matrix is better than counting the number of columns, and that taking into account the variations

	D_J^A	D_J^B	D_J^C	D_J^D
separate model (subset)	-0.037 ± 0.000215	-0.061 ± 0.000268	-0.056 ± 0.000336	-0.092 ± 0.000404
unique model	-0.014 ± 0.000072	-0.023 ± 0.000101	-0.034 ± 0.000173	-0.055 ± 0.000233
separate model (65 signatures)	-0.012 ± 0.000060	-0.020 ± 0.000087	-0.030 ± 0.000146	$-0.0490.000214\pm$

Table S1: Values for the coefficients α for different penalty shapes and training subset. We see that the coefficients for the whole dataset and for the 65 signatures examples are close. Overall, the confidence interval for the coefficients are small.

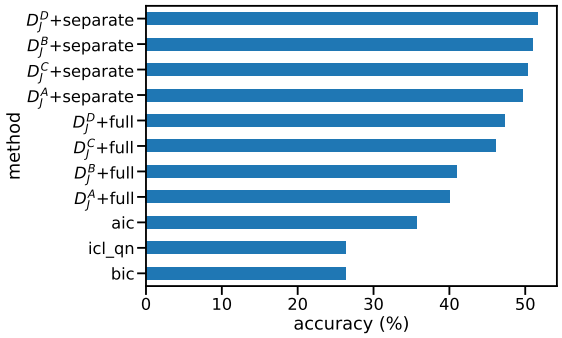


Figure S3: Test accuracy of various model selection criteria. BIC, AIC and ICL are standard model selection. The others are attempts to learn a valid criterion on simulated data.

in copy numbers through $M_{max_{avg}}$ does not bring any benefit. A complete overview of the number of clones found over the test set for each penalization strategy is given in Figure S4. In conclusion, we use in all our experiments an adaptive BIC criterion based on D_J^D as a measure of degree of freedom, and α estimated separately when CloneSig is fitted with 65 signatures or with a cancer-specific subset.

S1.3 Statistical test for signature change

To assess whether a signature change between clones is statistically significant, we design and calibrate a statistical test. To that end, we compare the likelihood of a CloneSig model with J clones as determined by the model selection criterion, and the likelihood of a model with the same clones but a single mixture of signatures common to all the clones (and found by fitting all observed mutations together). The objective of the test is to determine whether the difference between the two likelihoods is significant. To that end, we implement a likelihood-ratio test based on the statistics:

$$\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}} \, .$$

Following Neyman-Pearson lemma [54] one can set a threshold c to reject the null hypothesis that there is no signature change if λ is lower or equal to c with a certain level of significance α determined by the distributions of the likelihood of the model. As this distribution is unknown, we apply the Wilks theorem stating that asymptotically, $-2\log(\lambda)$ follows a chi-squared distribution of parameter the difference in dimensionality between the two alternative models [55].

As previously illustrated for the model selection criterion, the number of parameters is different from the degree of freedom in the case of CloneSig, so we resort to simulations to fit the degree of freedom of the test. We simulate a dataset with a similar mixture of signatures for all clones of each sample, and focused on samples with at least 2 clones, as described in Material and Methods. For the purpose of calibration, we use the true number of clones to fit the two alternative models. The objective of this approach is to fit a chi-squared distribution on the empirical distribution of $-2\log(\lambda)$ obtained in simulations. This is achieved again in two settings: fitting with all 65 signatures or with a cancer type-specific subset of signatures. In both cases, the distribution for each number of clones J evokes indeed a chi-squared distribution (Figure S5)

To fit the degree of freedom to use in the implementation of the test, as the degree of freedom of a chi-squared-distributed variable is its mean, we train a linear ridge regression model to fit $-2\log(\lambda)$ to relevant covariates. Four

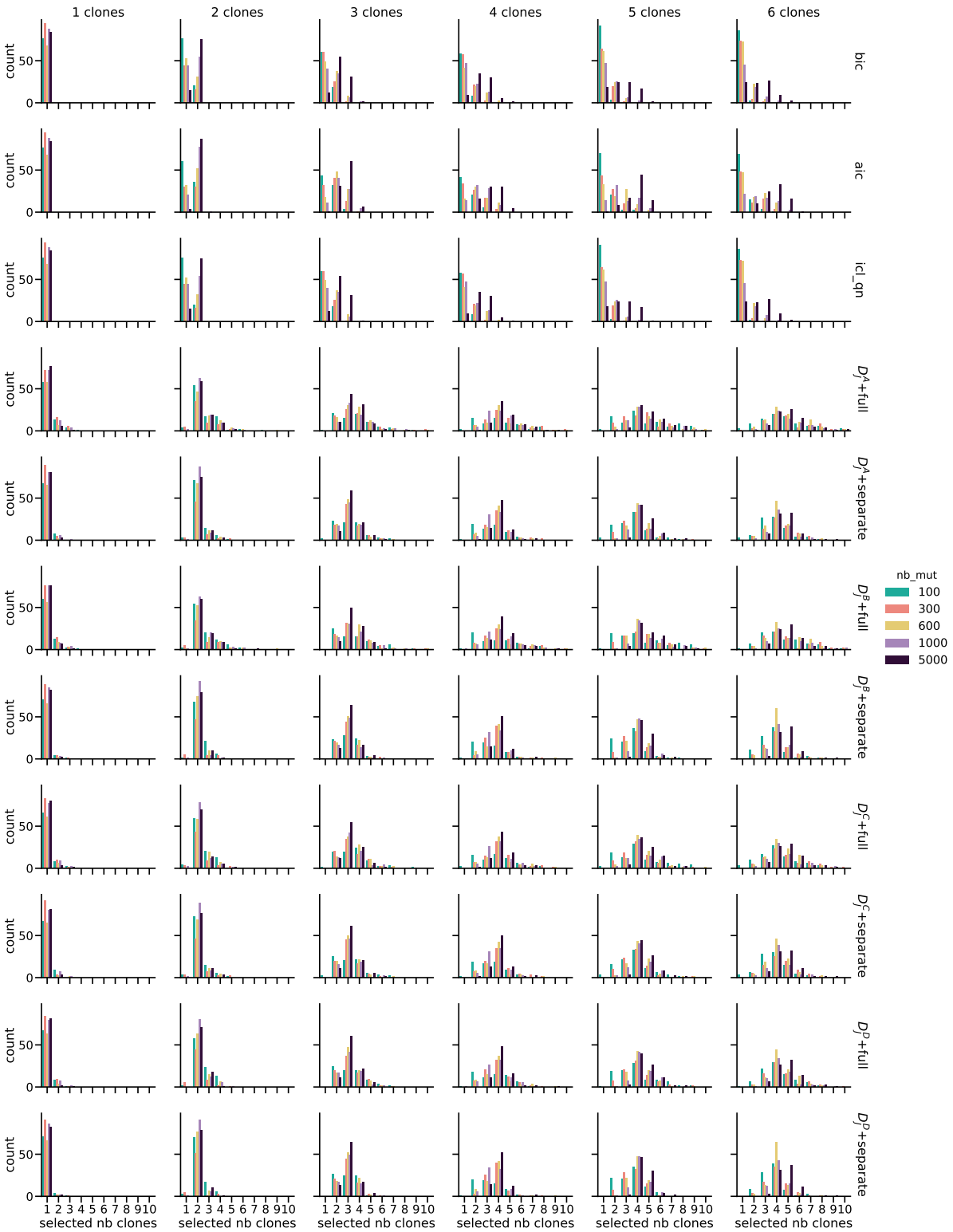


Figure S4: Number of clones found with different model selection criteria on the test set (not used to fit the model selection criteria). This illustrates the improved accuracy of the adapted BIC criterion compared to classical criteria

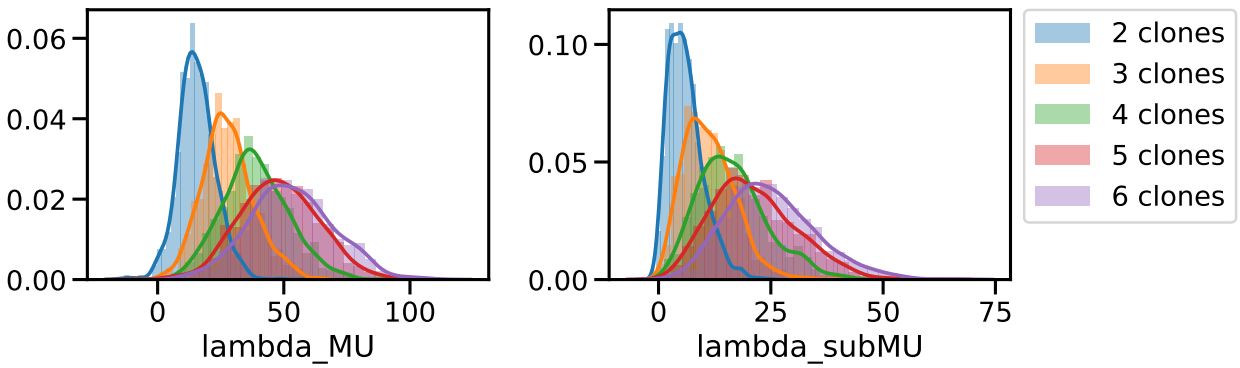


Figure S5: Empirical distribution of $-2\log(\lambda)$, with $\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}}$ obtained by fitting CloneSig with the true number of clones on simulated data, either with all 65 signatures (left), or with a subset of cancer type-specific signatures. The distribution is estimated separately for each number of clones.

covariates were initially considered: the number of clones, the degree of freedom of the input signature matrix, the number of mutations, and the diploid proportion of the genome. We found that the last two variables have no visible correlation with the target variable (see Supplementary Figure S6). Additionally, when added to the model, with

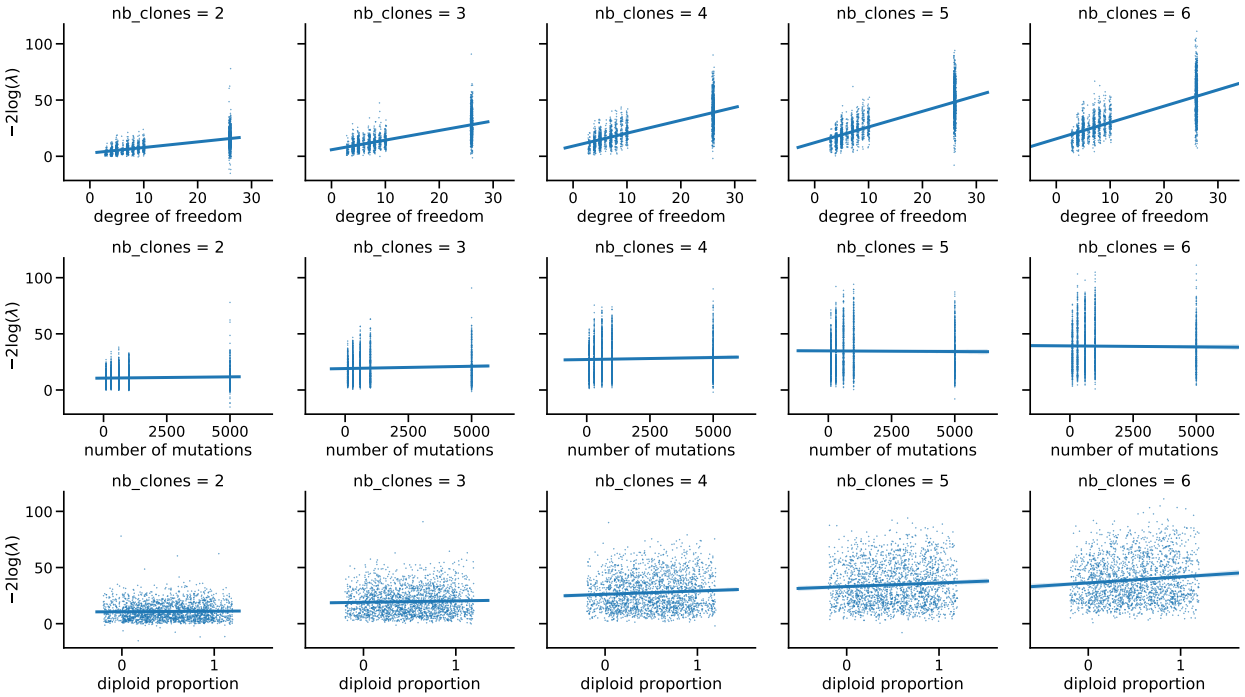


Figure S6: Correlation of $-2\log(\lambda)$, with $\lambda = \frac{\ell_{sigCst}}{\ell_{sigChange}}$ with potentially relevant covariates.

standard scaling of input variables, they have coefficients more than ten times smaller than the ones of the number of clones, and the signature degree of freedom. We therefore compute the final model on the two retained (unscaled) variables, and we average the values of the coefficients over 10-fold cross-validation. The resulting coefficients are reported in Table S2.

To finally ascertain the validity of the test, we now check the uniform distribution of the p-values for negative samples in Figure S7. There is a slight deviation from the uniform distribution, probably due to the fact that CloneSig does not necessarily converge to the true model likelihood (and instead to a local maxima), and thus does not respect the conditions of application of Wilks theorem.

	Intercept	Number of Clones coefficient	Degree of freedom coefficient
separate model (subset)	-13.677 ± 0.0778	4.777 ± 0.0117	1.662 ± 0.00991
unique model	-19.420 ± 0.0589	7.124 ± 0.0169	1.069 ± 0.00210
separate model (65 signatures)	-1.156 ± 0.107	9.470 ± 0.0279	0 ± 0

Table S2: values for the coefficient α for different penalty shapes and training subset. We see that the coefficients for the whole dataset and for the 65 signatures examples are close. Overall, the confidence interval for the coefficients are small.

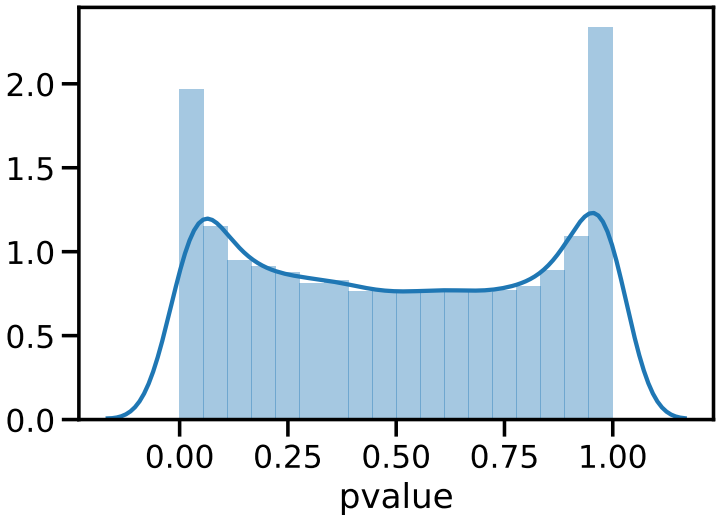


Figure S7: Empirical distribution of the p-values of the calibrated test of significance of signature change for negative simulated samples.

We finally explore the sensitivity of the test on the maximum cosine distance between signatures. The dataset used for that purpose consists of 2,700 samples with the number of clones varying between 2 and 6. For each number of clones, we drew 30 distinct π matrices with distinct maximal cosine distances between the mutation type profiles. For each number of clones and π matrix, we generated a sample with varying number of observed mutations, diploid percent of the genome, and sequencing depth. Figure S8 illustrates the proportion of samples where the test p-value is below 0.05 depending on the maximal distance between two subclones. We observe that detection is more efficient as the distance between clones becomes larger. Dependence on other variables is explored in Supplementary Figure S9.

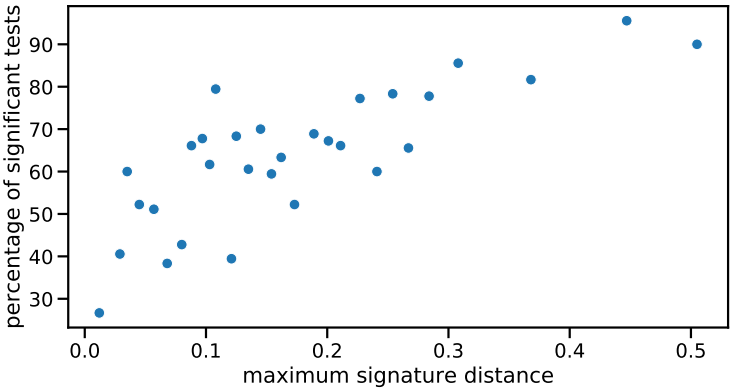


Figure S8: Percentage of significant tests depending on the max distance between 2 clones, quantized in 30 bins.

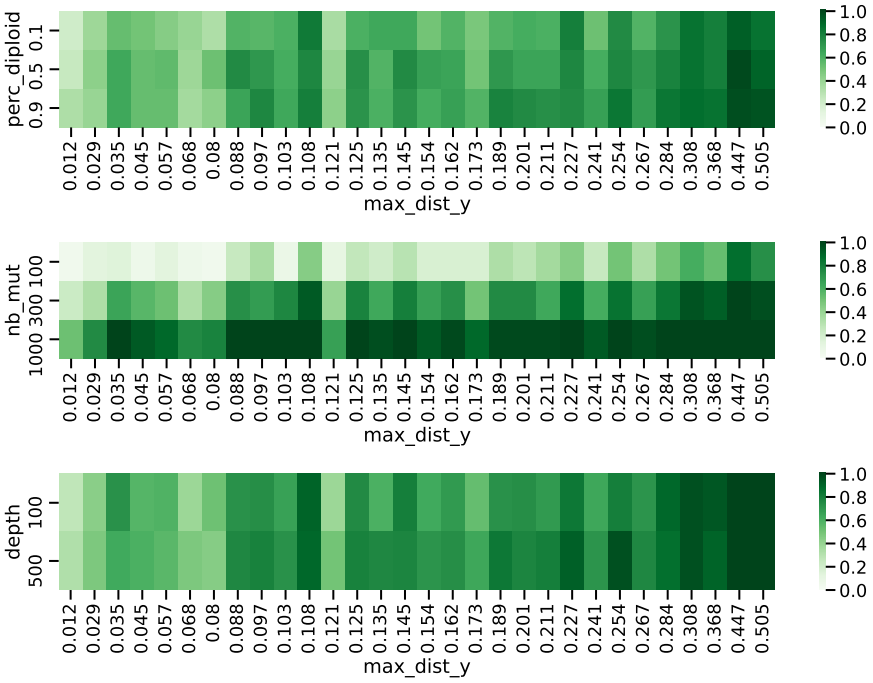


Figure S9: Percentage of significant tests depending several variables: number of mutations, and percentage of diploid genome, and sequencing depth

S1.4 Several "modes" to run CloneSig

A crucial difficulty in performing mutational signature deconvolution is the identifiability of the problem. Indeed, several mixtures of signatures may provide satisfying results. The most common approach to address this issue is to reduce the number of candidate signatures, in particular by using only signatures known to be active in the cancer type of the considered tumor sample [12] (approach `cancer_type`). An alternative approach is to perform two successive fits, the first one on all mutations in the sample in order to select potentially active signatures by keeping those with a contribution greater than a threshold, and the second one to refit those selected signatures with varying number of clones. This avoids the situation where a lot of signatures have very small contributions to the final mixture [15] (approach `prefit`). Those two alternatives are implemented in CloneSig (see Figure S10) and also tested for all methods tested (see supplementary Figures S11-S24). For the subclonal reconstruction problem, we see that the two approaches that limit the number of signatures have similar performance and improve the accuracy of CloneSig, especially in cases with few mutations. However, for the signature deconvolution problem, even though

the **prefit** approach exhibits improved performance compared to taking all signatures, the **cancer_type** approach shows significantly better results. The results were similar for the other signature deconvolution methods, so for the rest of the analysis, we retain the **cancer_type** approach, and report only one result per method, to simplify the interpretation.

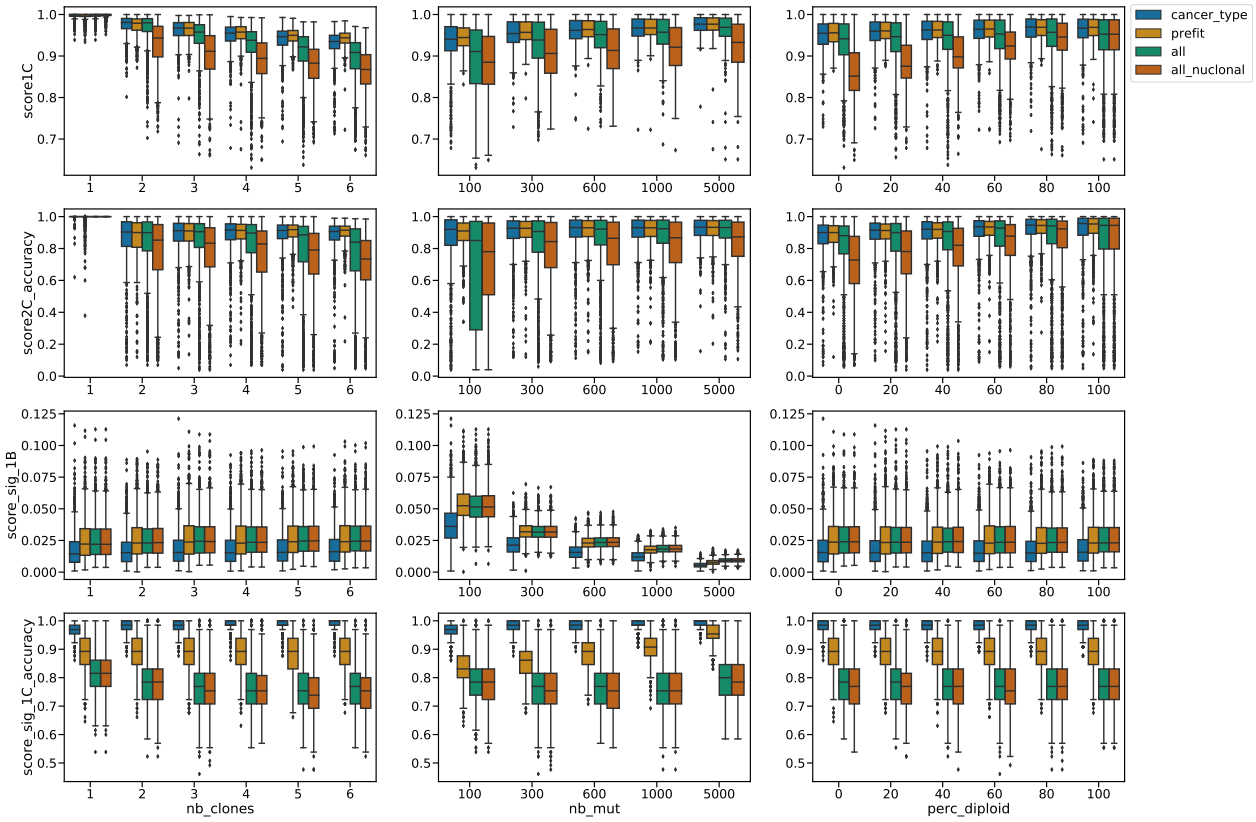


Figure S10: CloneSig's performance for 3 different input signature strategies: use all available signature (all), a subset of cancer type-specific signatures (cancer_type), or proceed in two steps by first fitting all mutations together to select potential signatures, and then actally run CloneSig with the selected subset (prefit). Additionally, the contribution of CloneSig'a approach for accounting for copy number was evaluated, by implementing the simpler approach from Palimpsest [40] (all_nuclonal).

Note S2 Full benchmarking results

To fully assess CloneSig's performance in simulations, in comparison with other state-of-the-art approaches for subclonal reconstruction and signature deconvolution, we report here the full results with all tested "modes" (all signatures, a subset of cancer-type-specific signatures, or a pre-fit step where only the most prominent signatures found on the whole set of mutations are then retained for the true signature deconvolution for CloneSig, TrackSig and Palimpsest). In this extensive version of the results, we report all metrics used to create score2C (AUC, specificity, sensitivity), score_sig_1C and score_sig_1E (`max_diff_distrib_mut`, `median_diff_distrib_mut`, `perc_dist_5` and `perc_dist_10`).

Regarding the subclonal reconstruction problem, for all metrics, there is little difference between the different modes of each signature-aware method, except for `score2C_sensitivity` for CloneSig, where the use of the cancer-type-specific subset exhibits better results. For signature deconvolution, there is a higher variability of results with respect to the run mode. CloneSig is the best performing method, except for one metric: `max_diff_distrib_mut`. For `Score_sig_1C`, the mode cancer-type-specific subset for CloneSig achieves a very good specificity, but the other modes have a high proportion of false positive signatures.

Additionally, we conduct a similar benchmark in the case where there is no signature change between subclones, and present results in Supplementary Figures S11 to S25, panels b, c, f. The improvement of CloneSig over other methods in subclonal reconstruction is partially lost in this setting, but CloneSig remains competitive, and the best performing method for score 1B up to 3 clones. A similar trend is visible for all scores for the subclonal reconstruction problem, with slightly worse scores, and higher inter-quartile space when there is no signature variation between clones. For the signature deconvolution problem, most metrics are unaffected, except for score_sig_1E, where all methods perform better and close the gap with CloneSig. Overall, CloneSig performs better than other methods when there are differences of signature activities between subclones, and remains competitive with other approaches in the absence of signature change.

The runtimes of all methods for those simulations are presented in Figure S25. The main determinant of runtime is the number of input mutations for all methods. CloneSig is slower than methods involving variational inference for the subclonal reconstruction problem, but is significantly faster than PyClone, especially for high numbers of mutations, thus illustrating its scalability to both WES and WGS data.

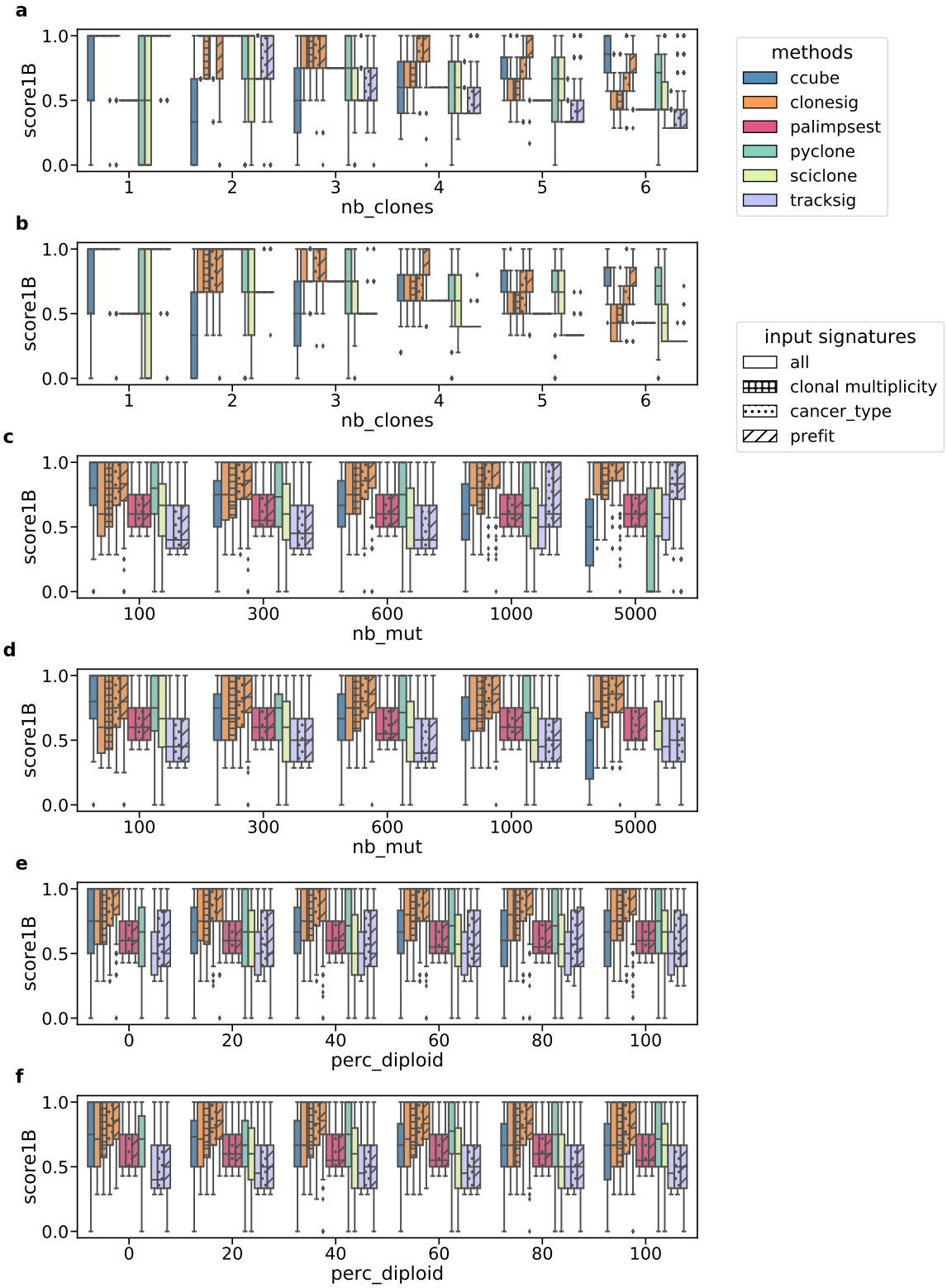


Figure S11: Score_1B for ITH methods on simulated data, with varying number of clones (a,b), number of observed mutations (c,d) and diploid percent of the genome (e,f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

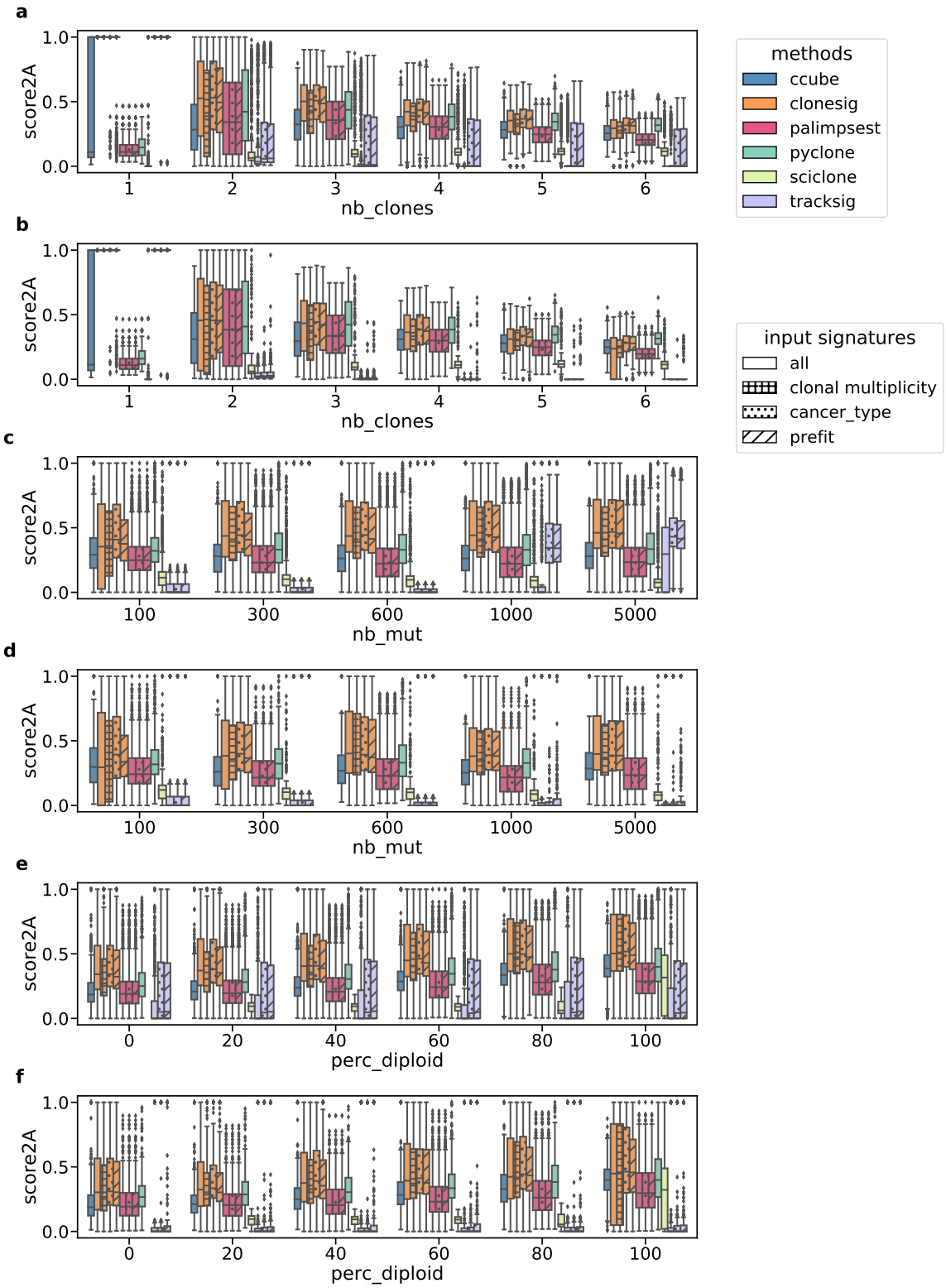


Figure S12: Score_2A for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

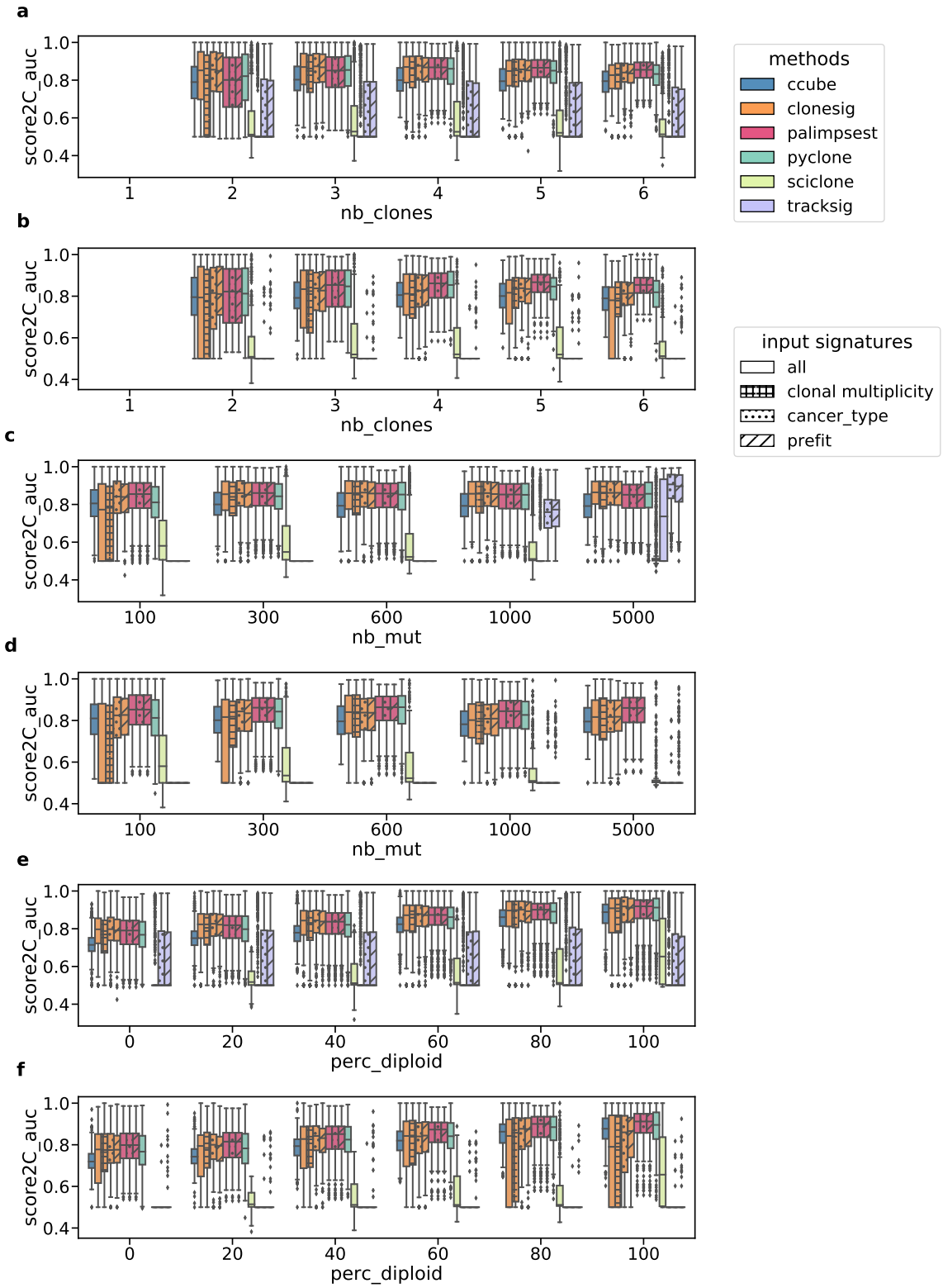


Figure S13: Score_2C (area under the curve) for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

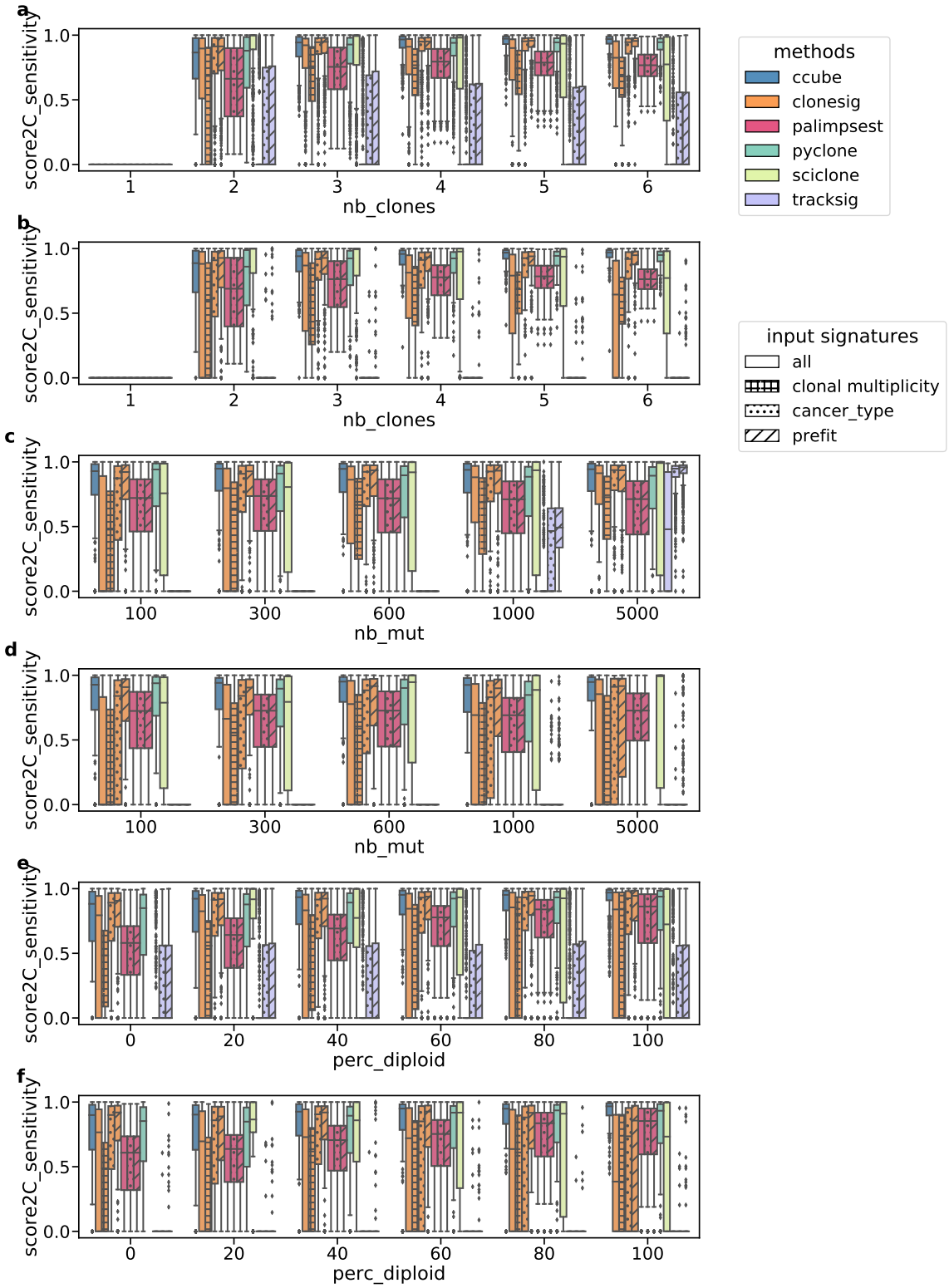


Figure S14: Score_2C (sensitivity) for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

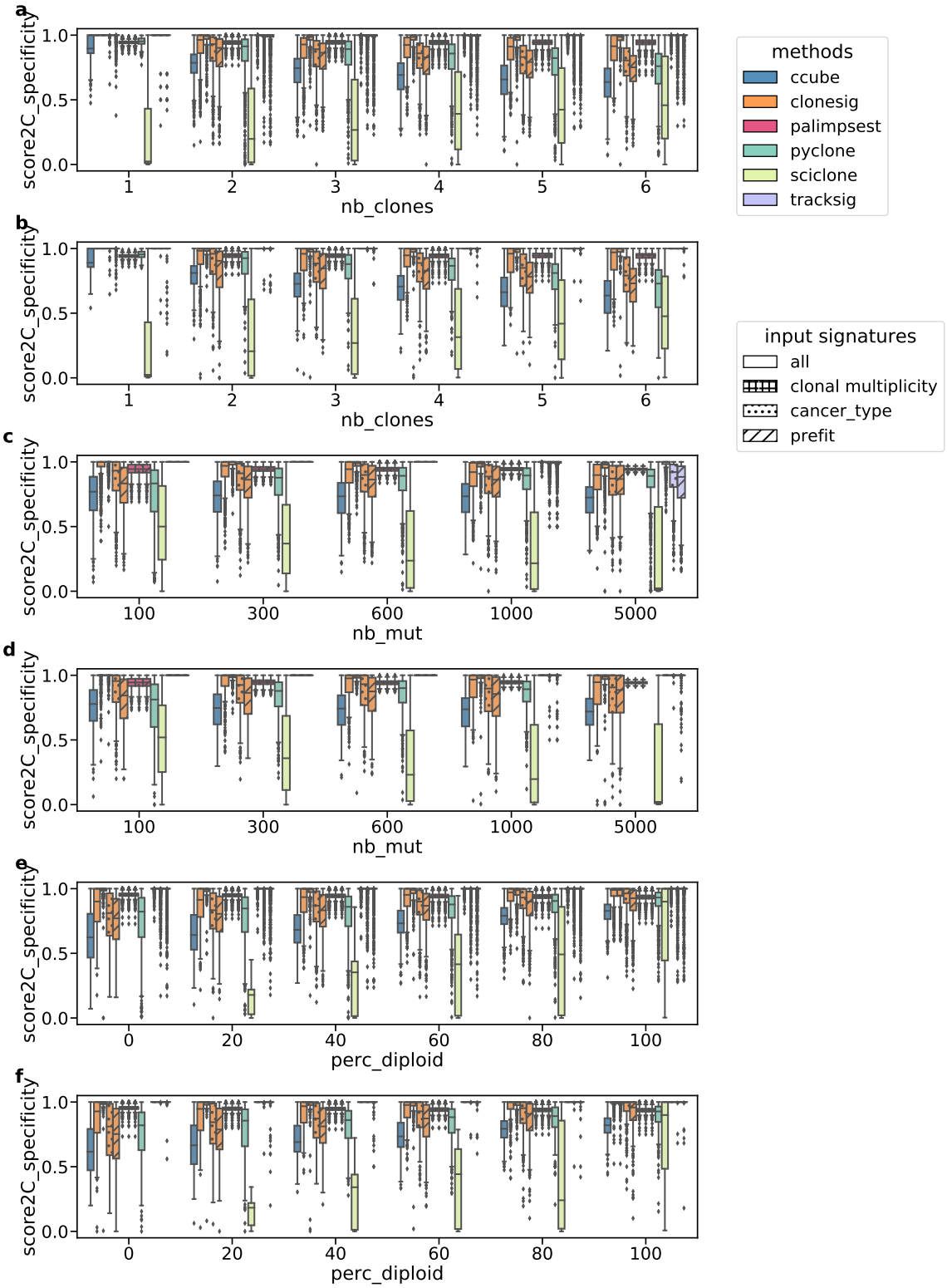


Figure S15: Score_2C (specificity) for ITH methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

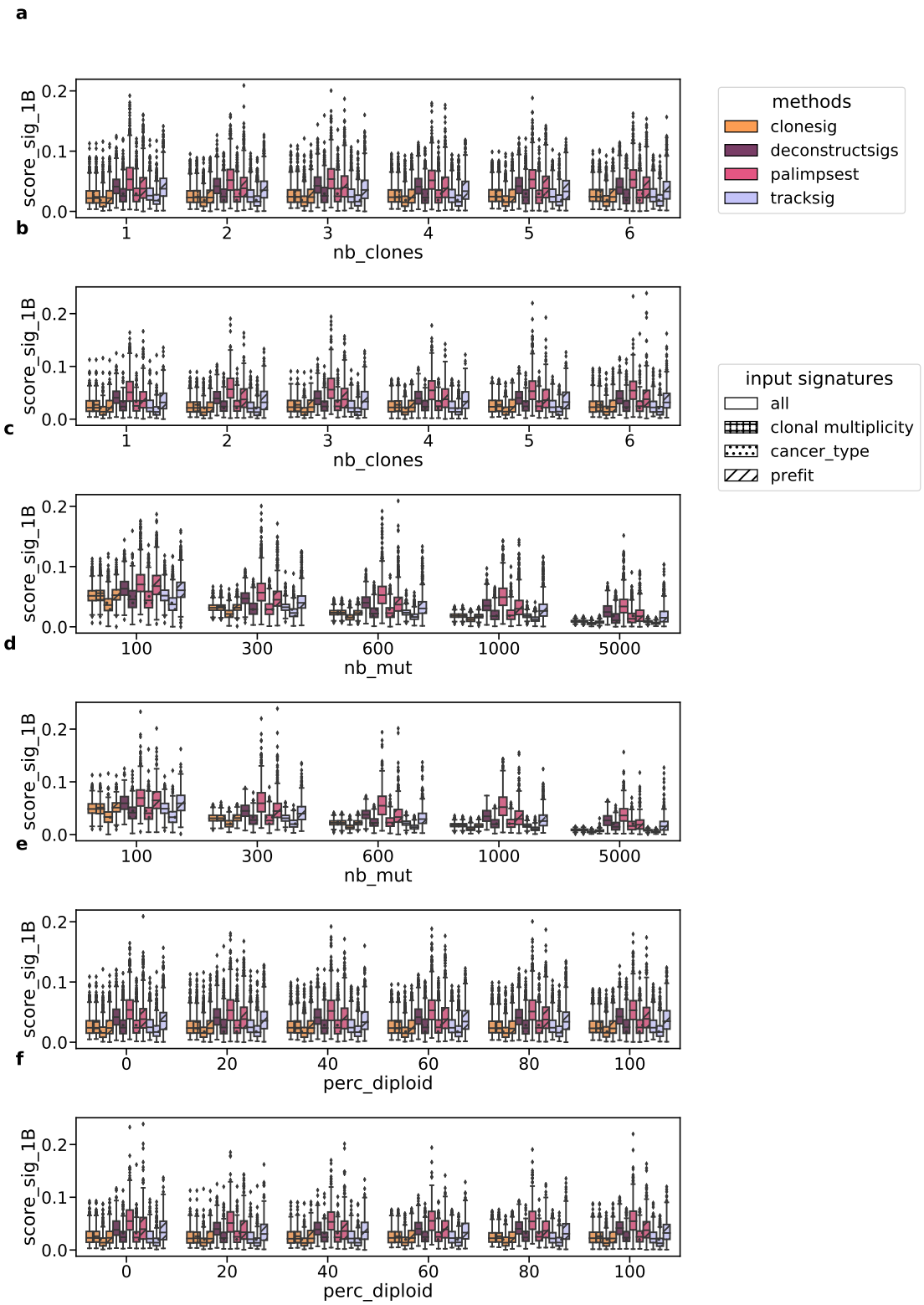


Figure S16: Score_sig_1B for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

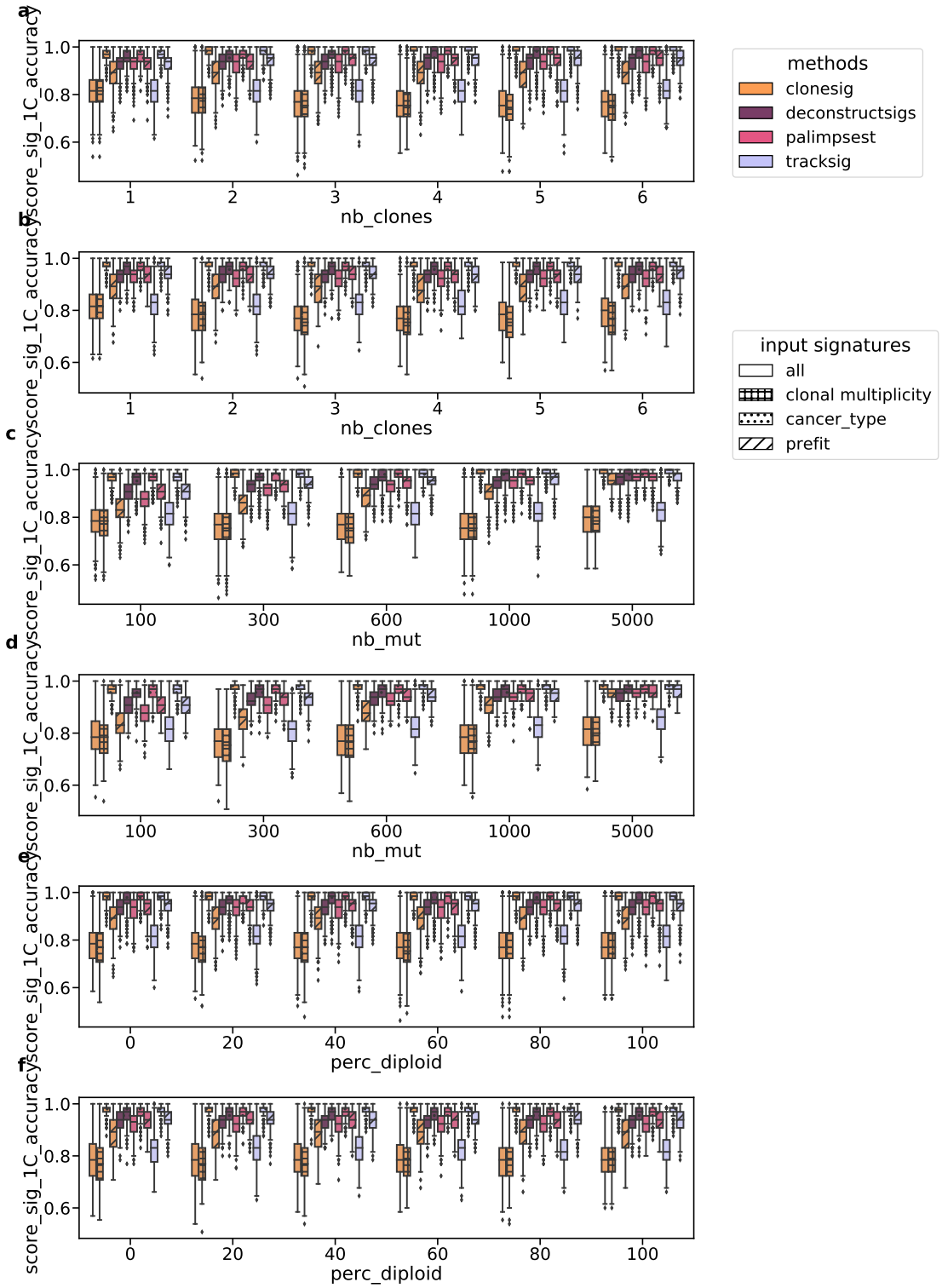


Figure S17: Score_sig_1C (accuracy) for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

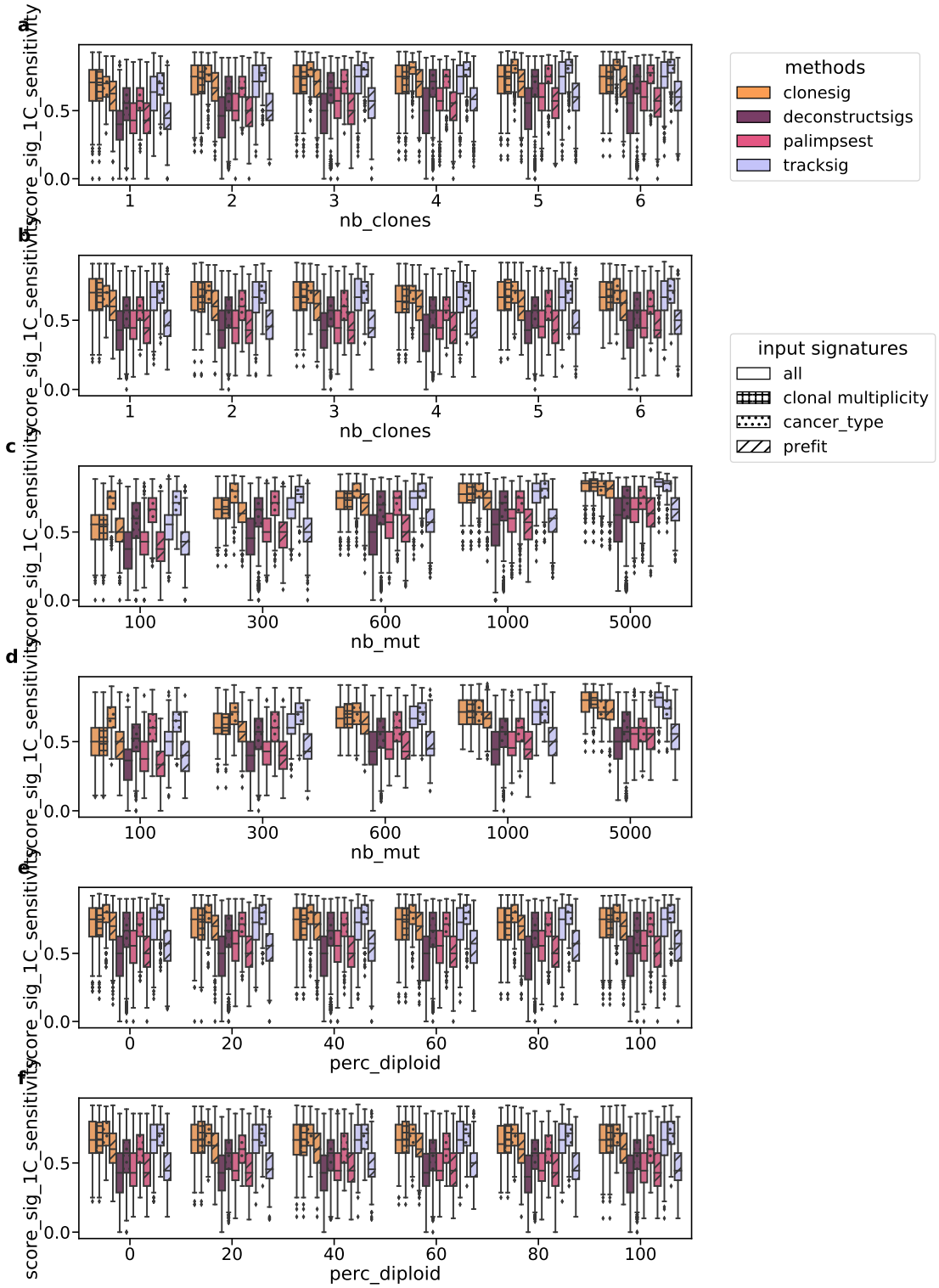


Figure S18: Score_sig_1C (sensitivity) for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

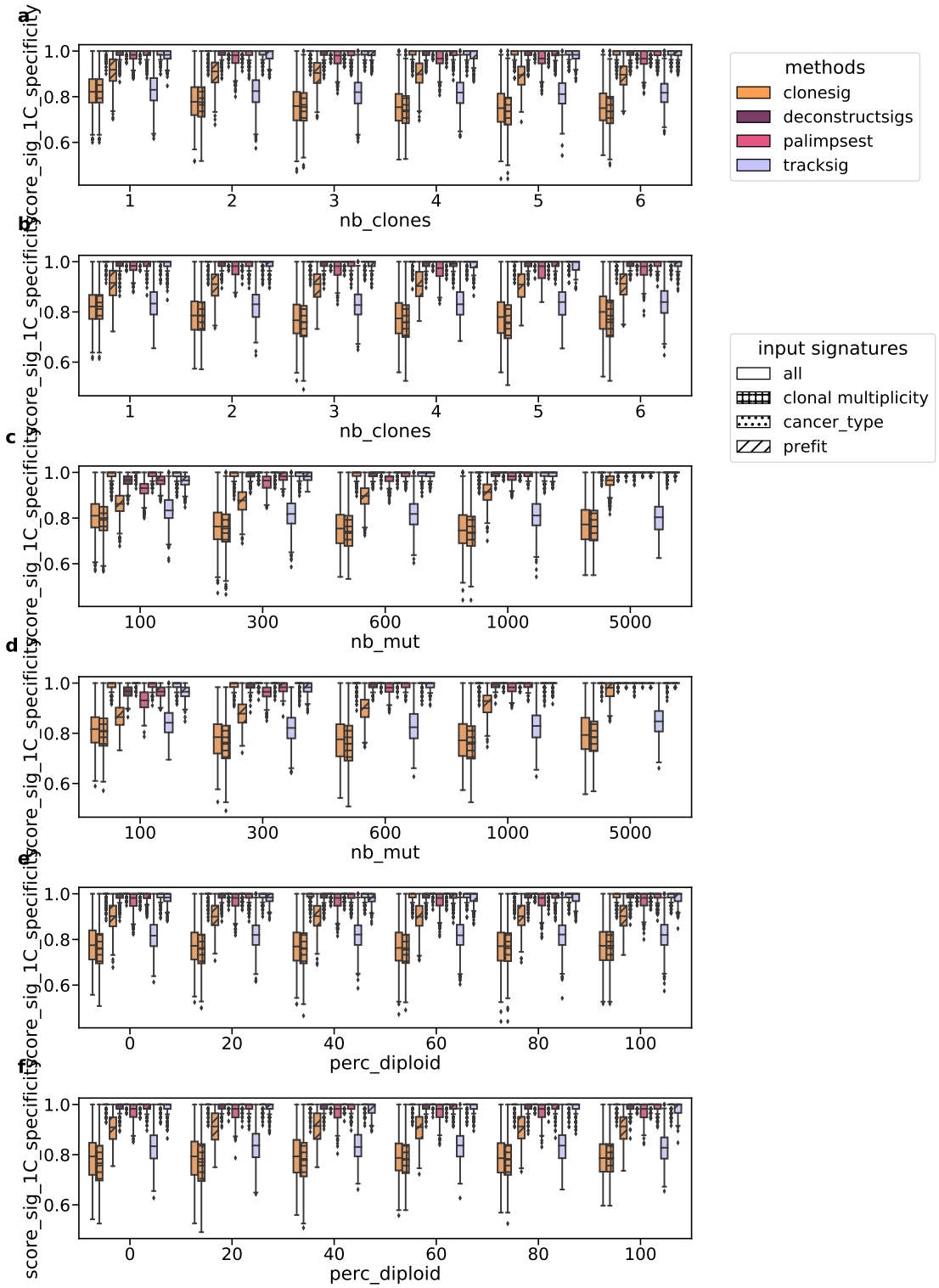


Figure S19: Score_sig_1C (specificity) for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

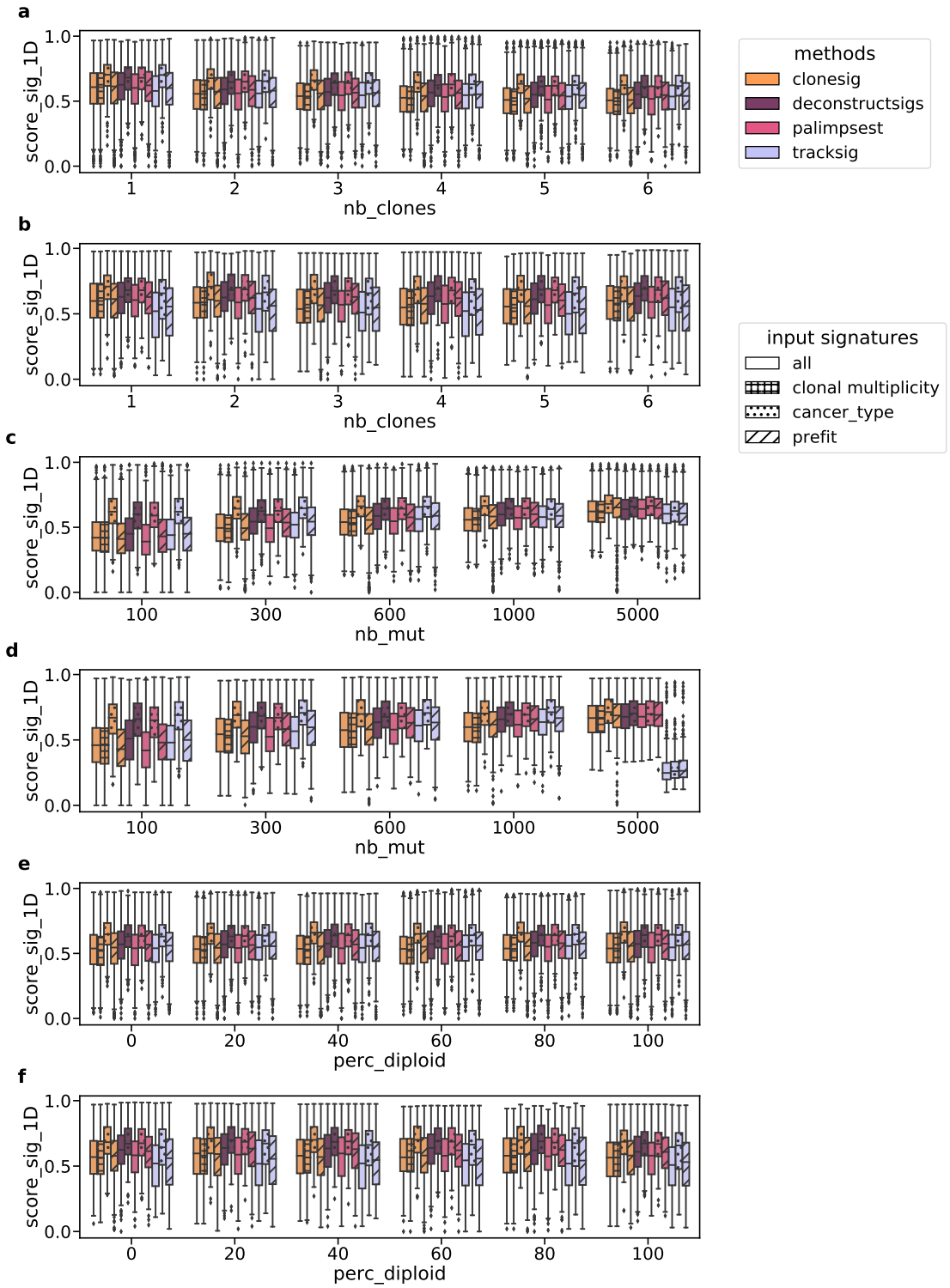


Figure S20: Score_sig_1D for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

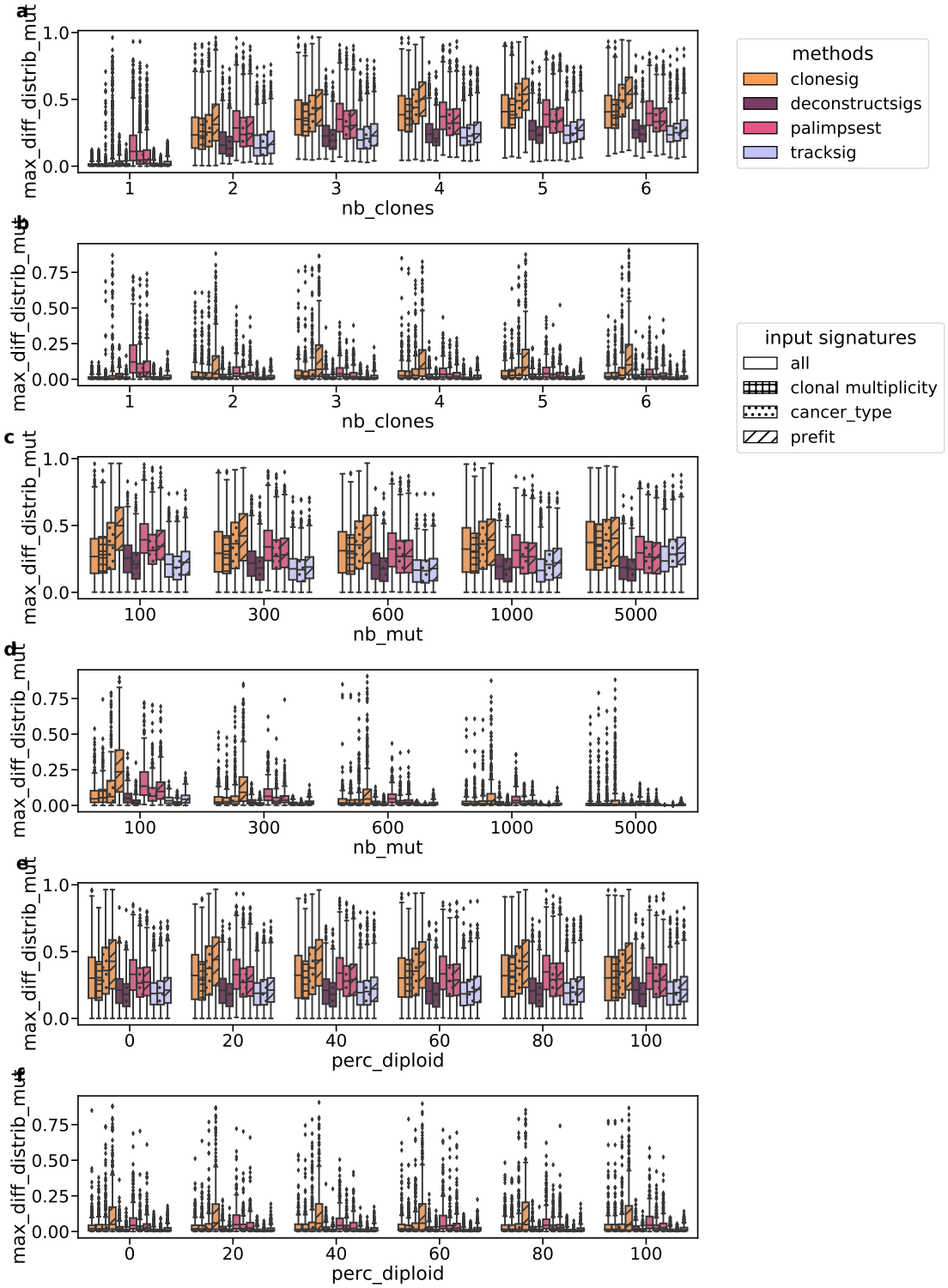


Figure S21: Maximal cosine distance between the true and estimated mutation type profile for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

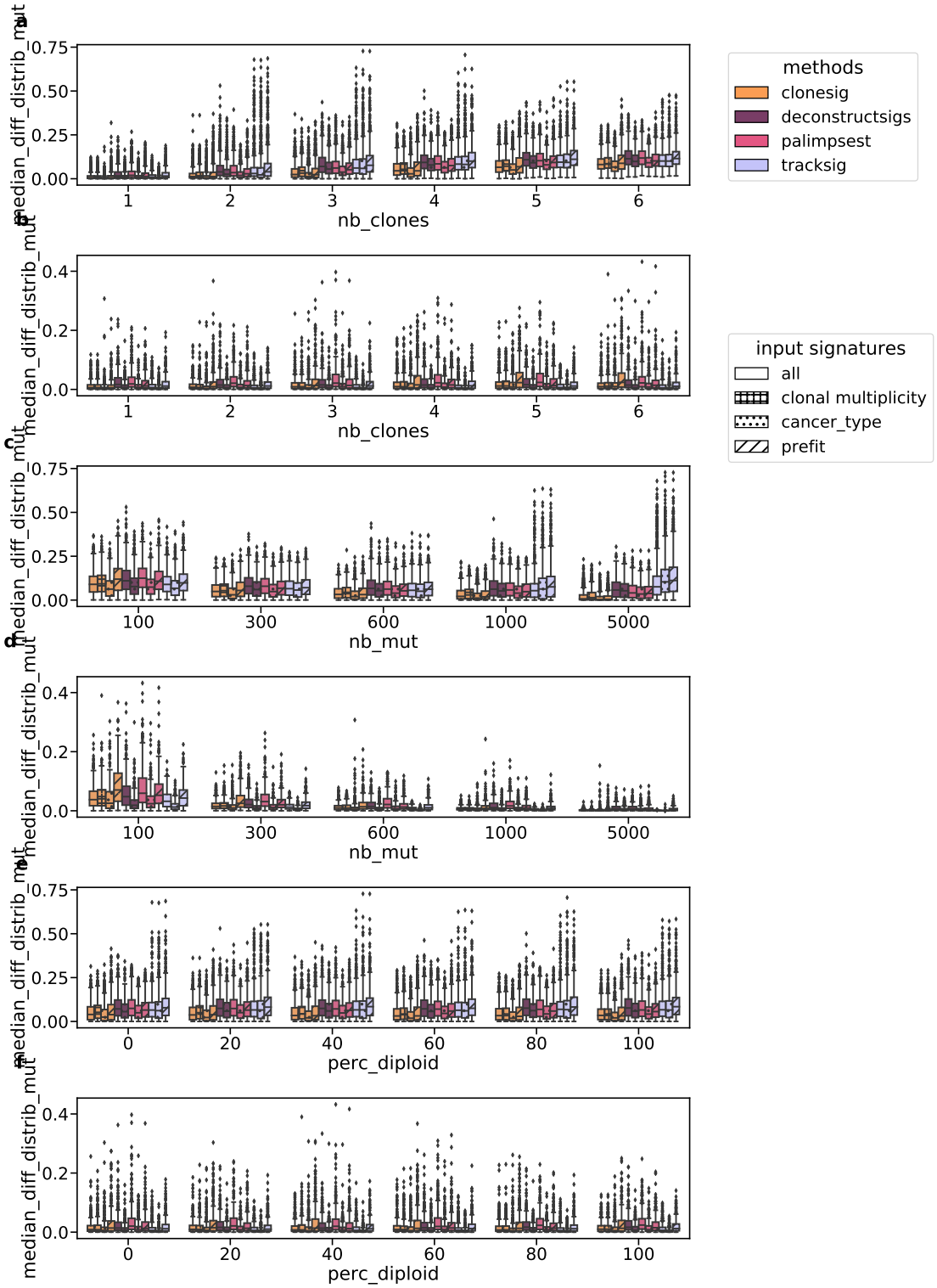


Figure S22: Median cosine distance between the true and estimated mutation type profile for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

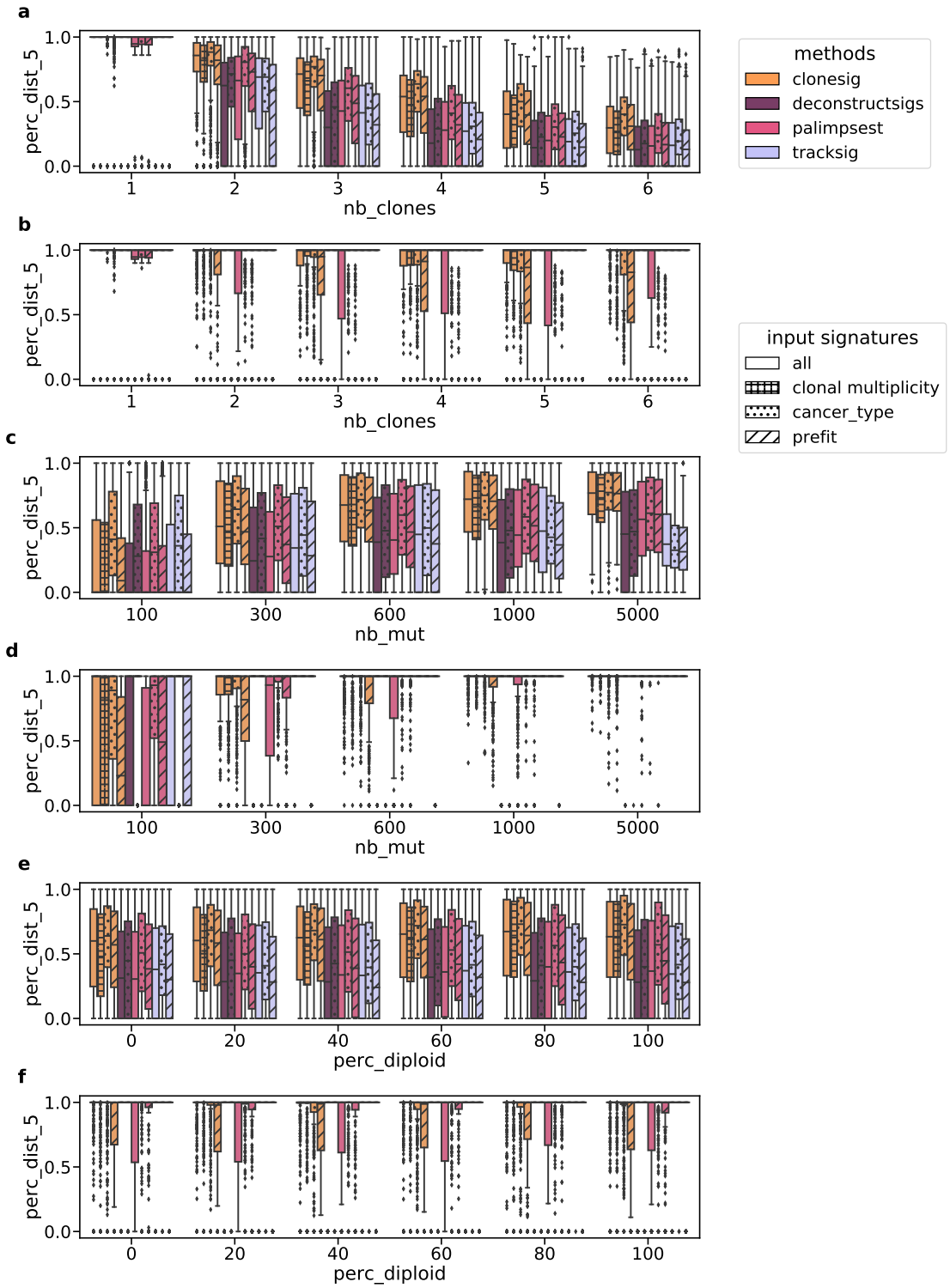


Figure S23: Proportion of SNVs with cosine distance between the true and estimated mutation type profile under 0.05 for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

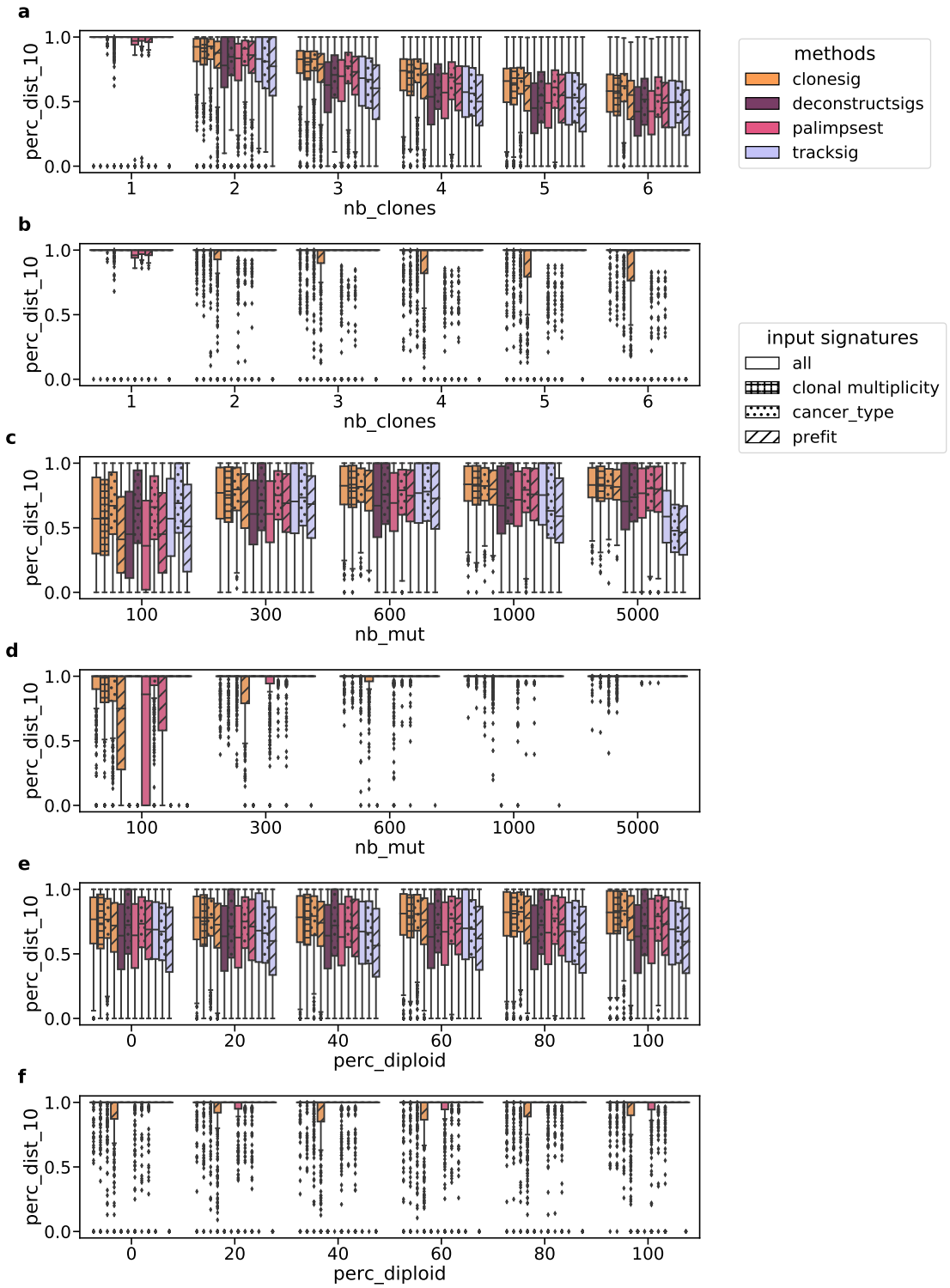


Figure S24: Proportion of SNVs with cosine distance between the true and estimated mutation type profile under 0.10 for signature deconvolution methods on simulated data, with varying number of clones (a, b), number of observed mutations (c, d) and diploid percent of the genome (e, f). Panels a, c and e correspond to simulations with varying signature between clones, and b, d, f to simulations with constant signatures.

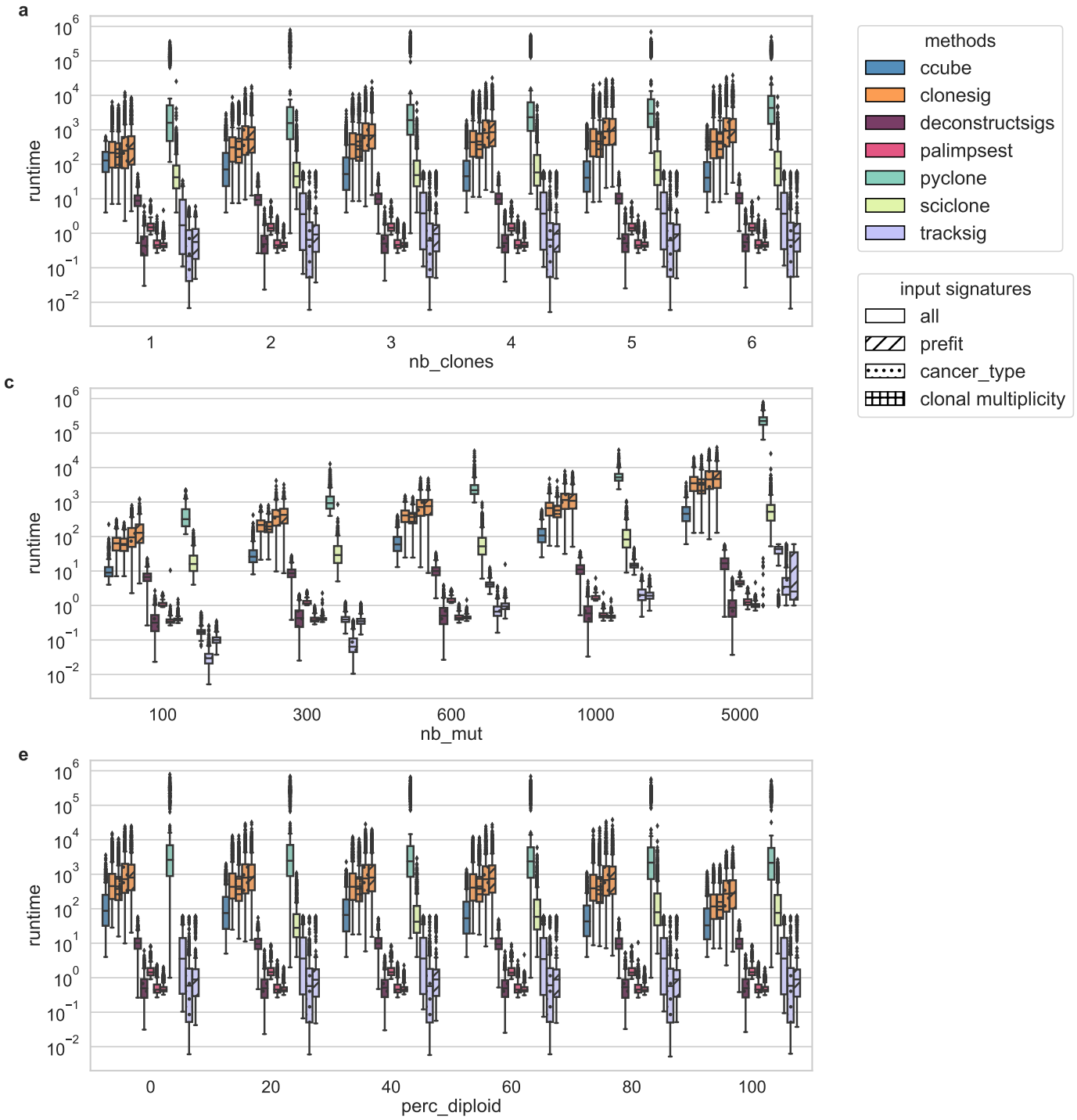


Figure S25: Runtime for ITH reconstruction and signature deconvolution methods on simulated data, with varying number of clones (a), number of observed mutations (b) and diploid percent of the genome (c). Results with varying signature between clones only are shown but similar results were obtained on simulations with constant signatures.

Note S3 Complete overview of TCGA results

To complete the analysis of the TCGA, we present here heatmaps to delineate an overview of each cancer type in Figures S26 to S56. For each type, the first panel represents the difference between subclonal and clonal signature activities (in case of a significant change in activity), and the bottom panel represents the absolute values of each signature activity for clonal SNVs (belonging to the clone of largest CCF estimated by CloneSig), and in the main subclone (in terms of number of SNVs). This allows researchers to fully explore CloneSig's results on the TCGA, and further compare their results in future studies. For each panel, we have added several clinical variables, in particular, the patient's age at diagnosis, the stage of the tumor, the size class of the primary tumor, and the patient's sex. Overall, we found no trend of association between signature activities or change in activities and those clinical characteristics, as previously observed in the particular case of prostate cancer [19].

In most types, like CESC (Figure S29), HNSC (Figure S35) and others, we observe groups of patients with different patterns of signature activity. The clinical significance of such groups remains to be further explored.

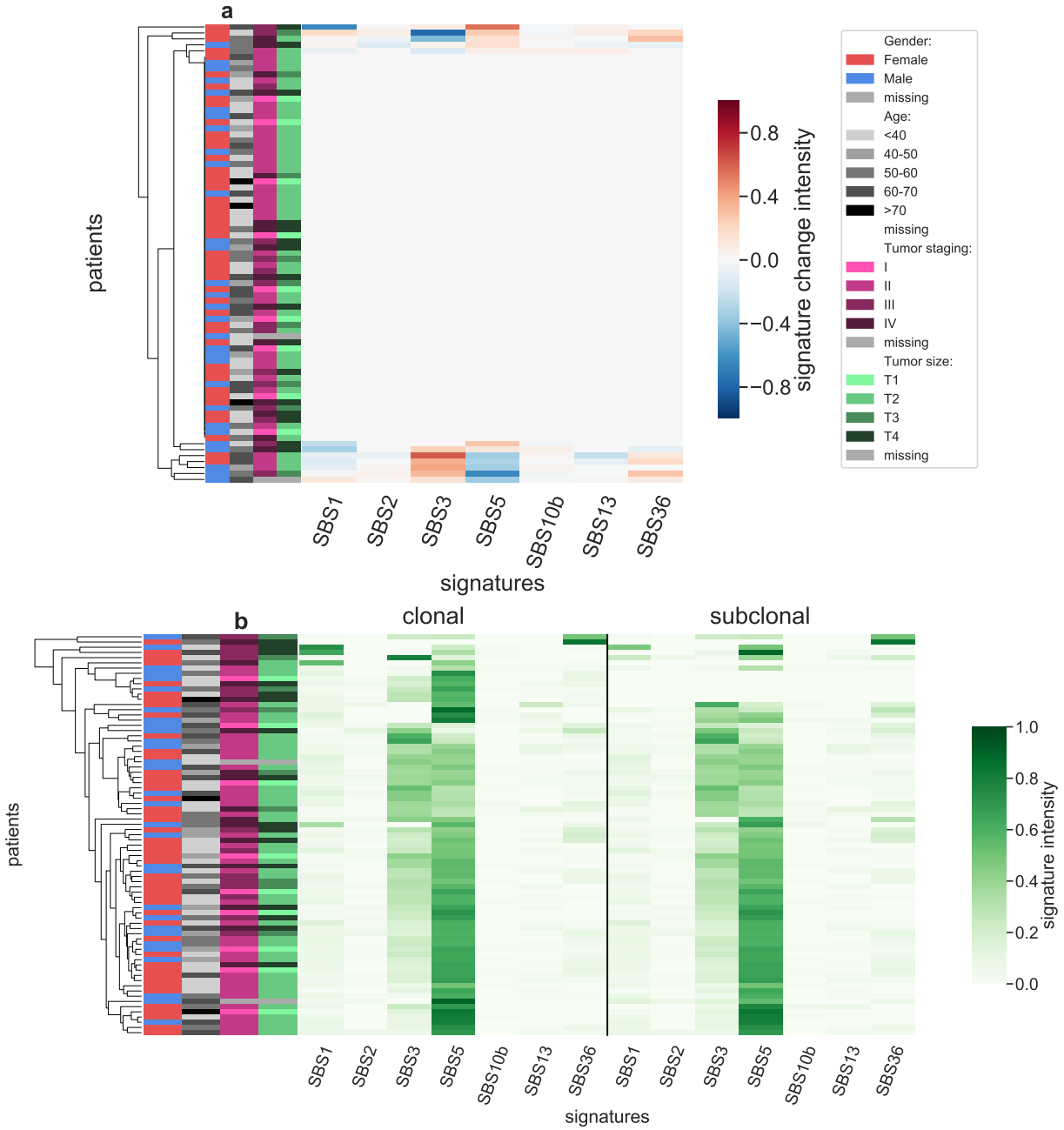


Figure S26: Panel a: Stratification of patients depending on their pattern of signature change for ACC patients (77 patients, including 12 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

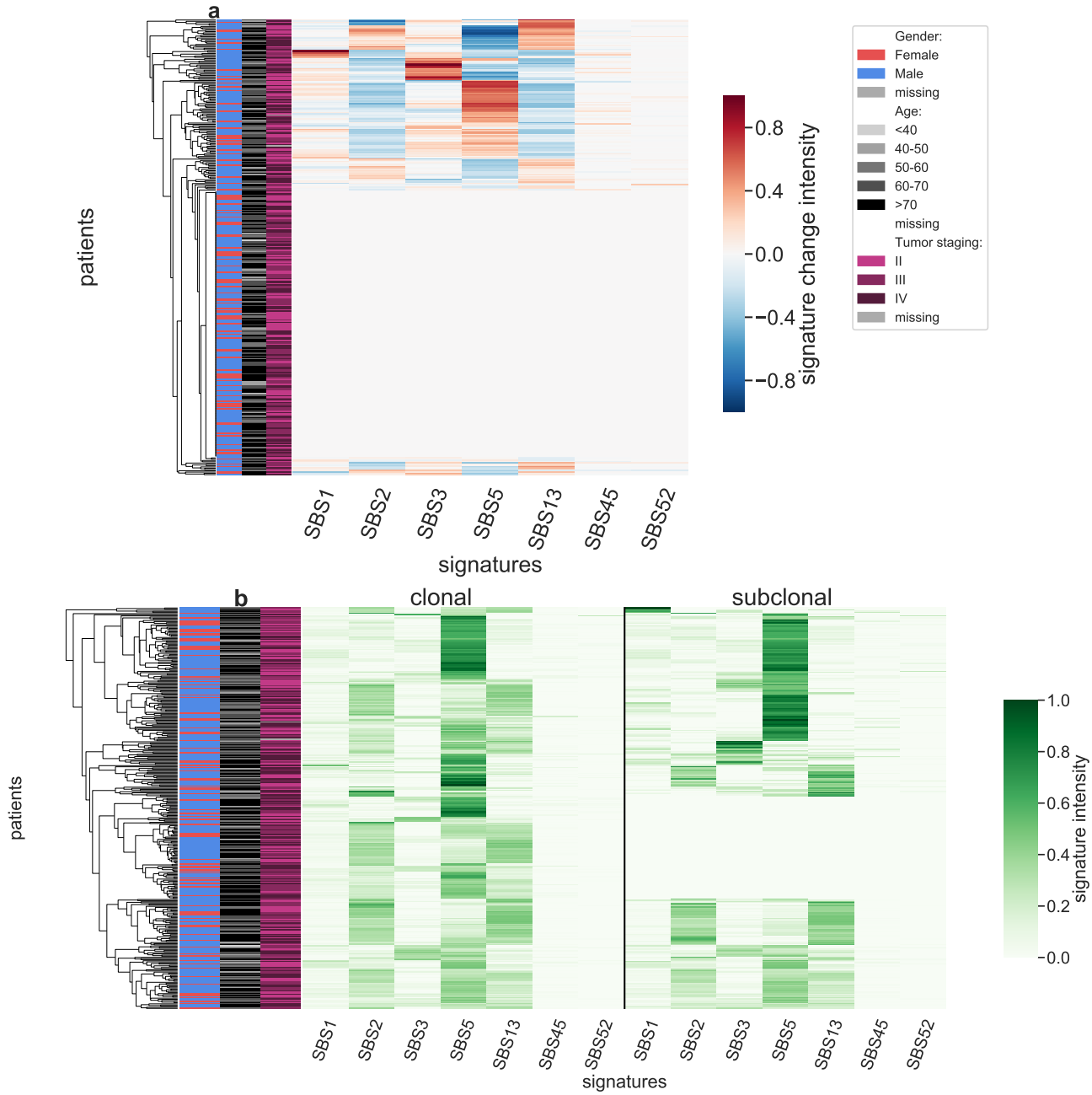


Figure S27: Panel a: Stratification of patients depending on their pattern of signature change for BLCA patients (354 patients, including 147 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

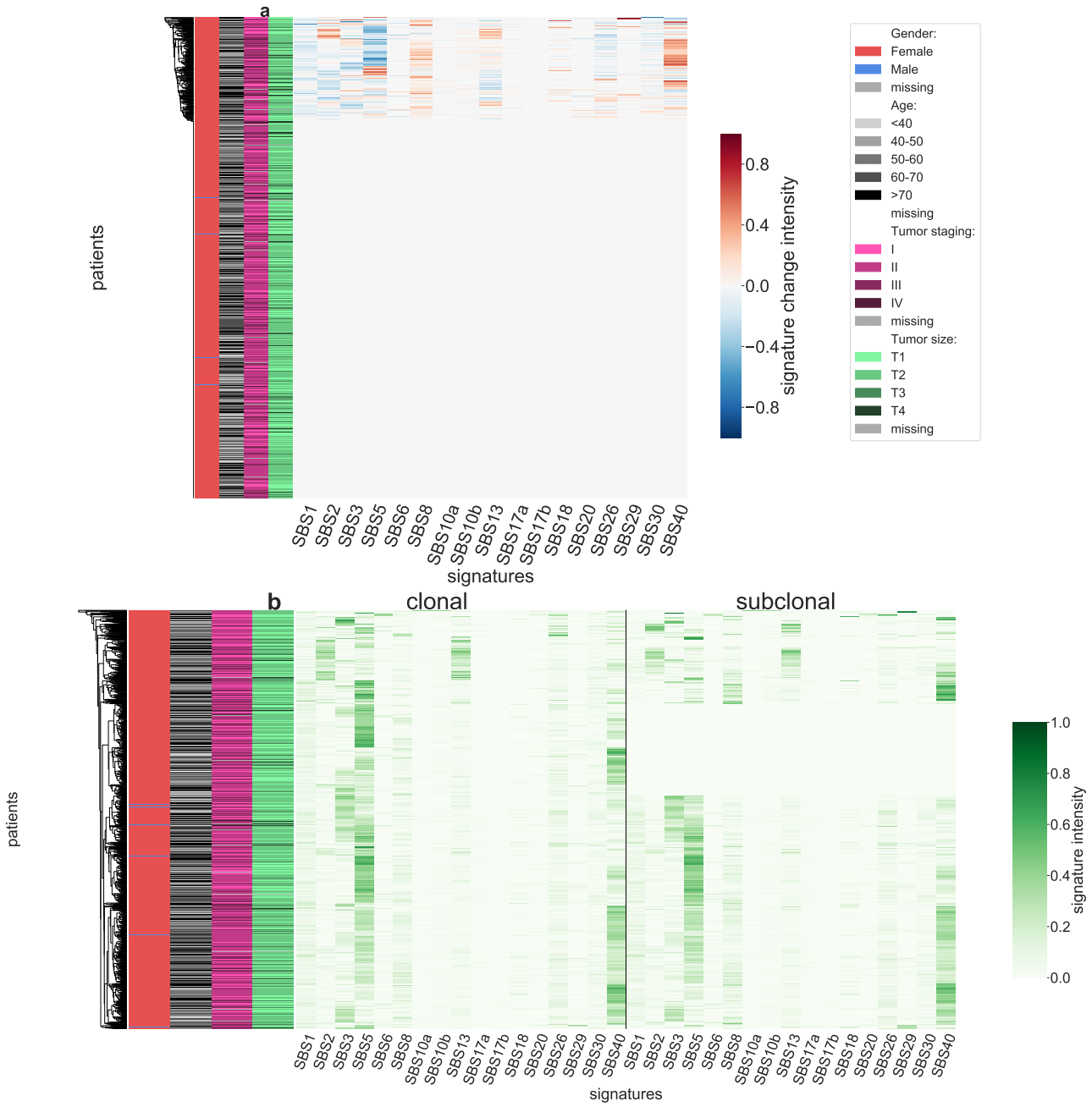


Figure S28: Panel a: Stratification of patients depending on their pattern of signature change for BRCA patients (931 patients, including 200 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

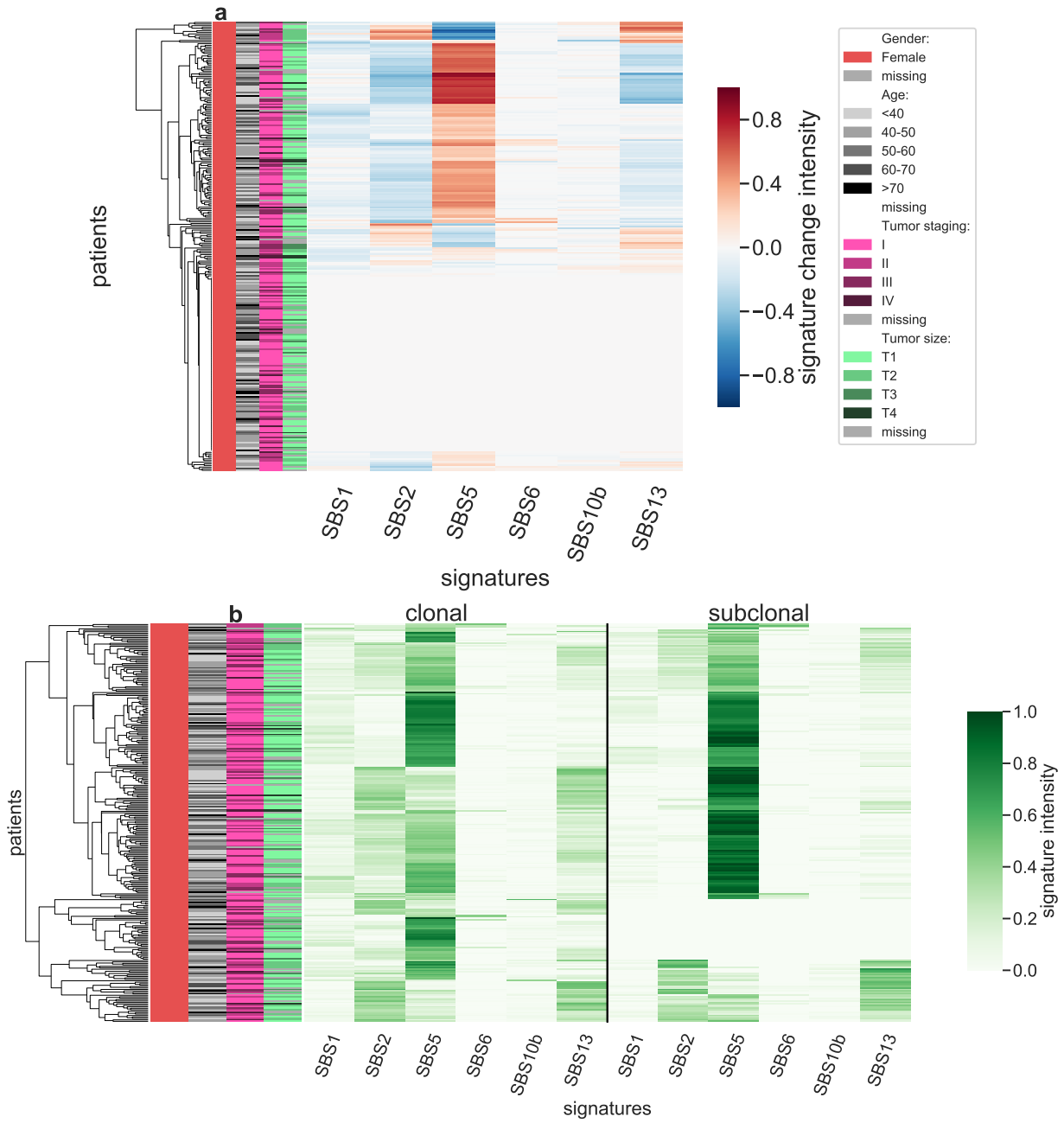


Figure S29: Panel a: Stratification of patients depending on their pattern of signature change for CESC patients (275 patients, including 168 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

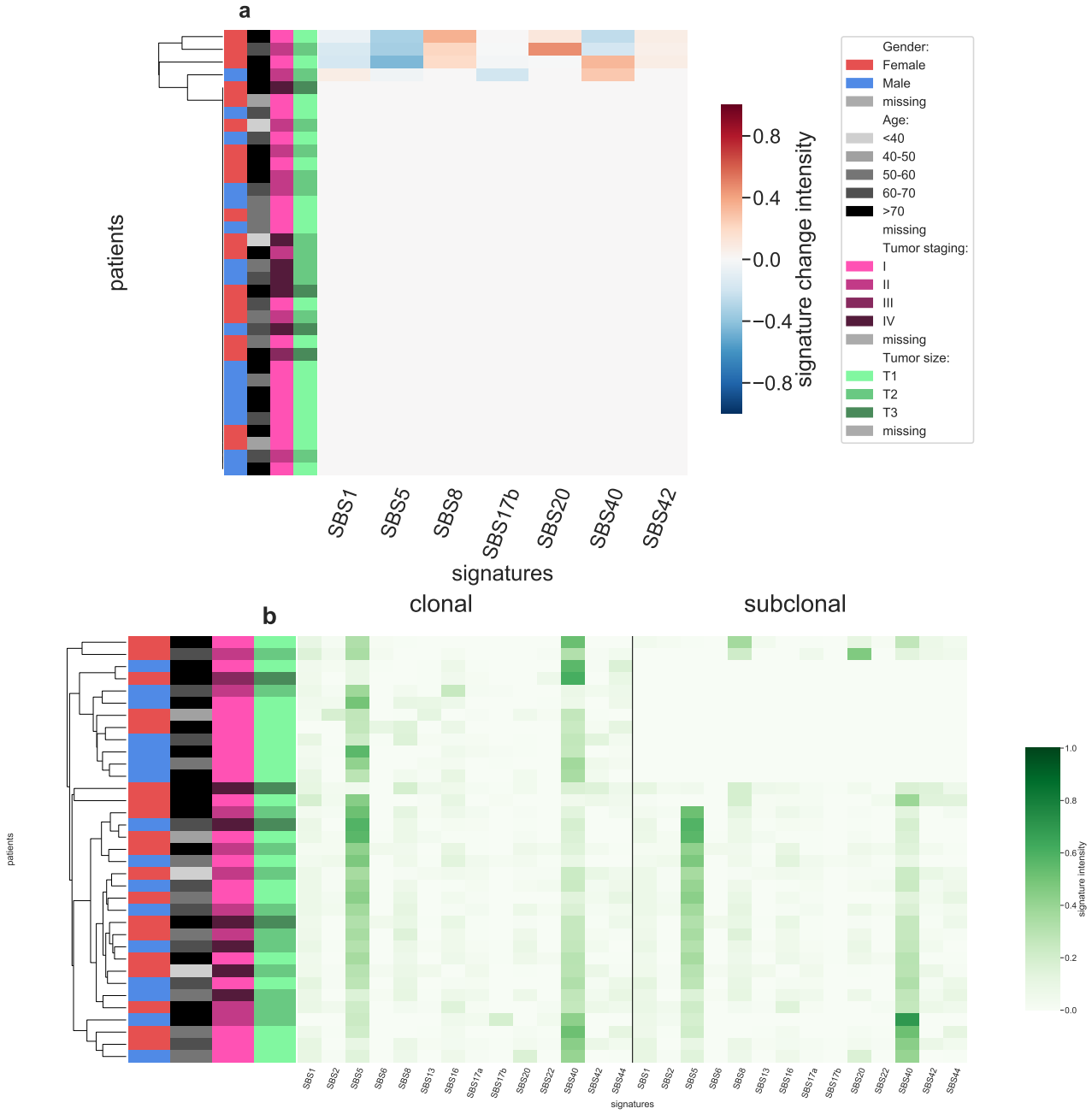


Figure S30: Panel a: Stratification of patients depending on their pattern of signature change for CHOL patients (35 patients, including 4 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

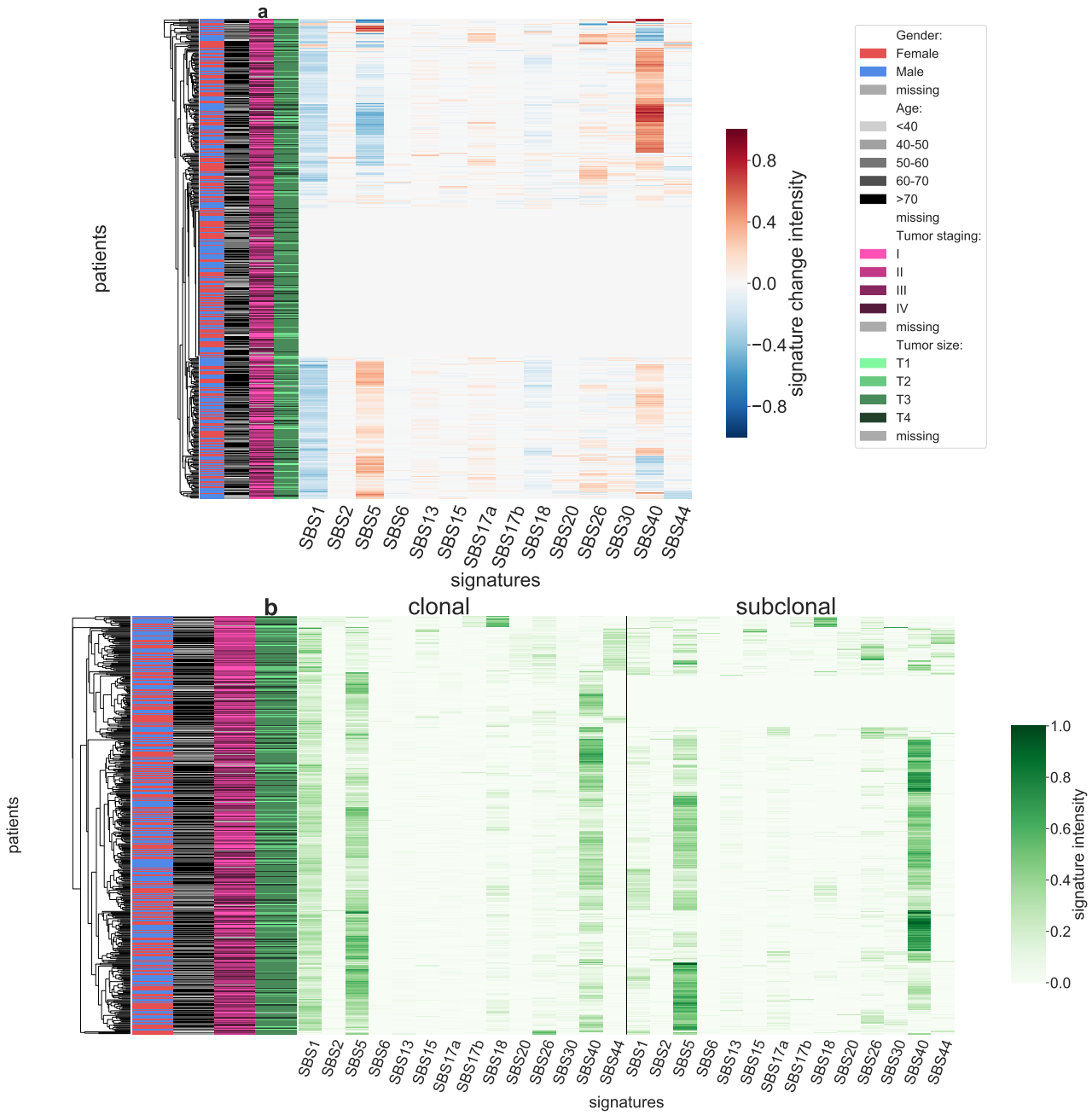


Figure S31: Panel a: Stratification of patients depending on their pattern of signature change for COAD-READ patients (458 patients, including 318 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

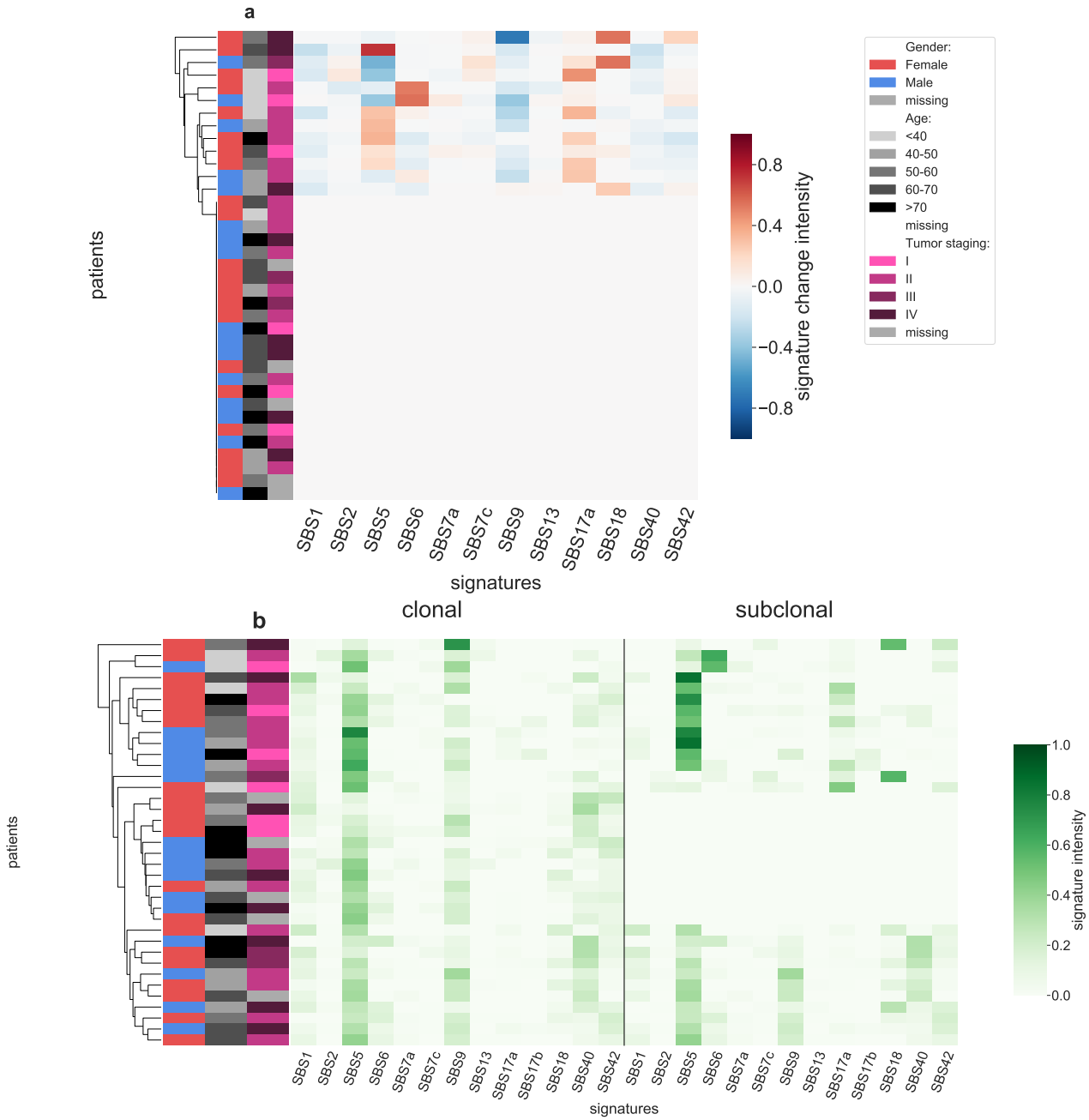


Figure S32: Panel a: Stratification of patients depending on their pattern of signature change for DLBC patients (37 patients, including 13 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

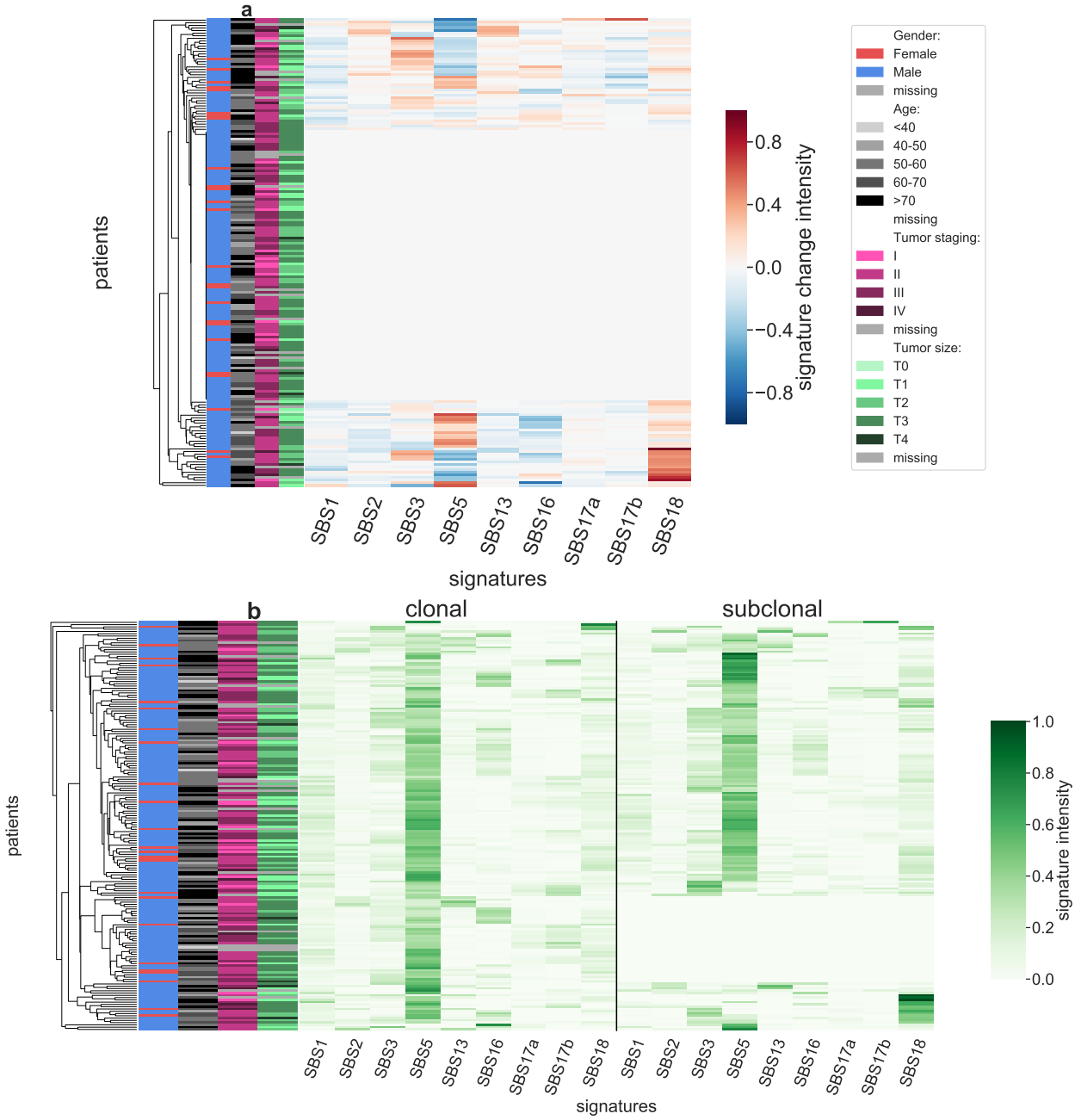


Figure S33: Panel a: Stratification of patients depending on their pattern of signature change for ESCA patients (180 patients, including 76 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

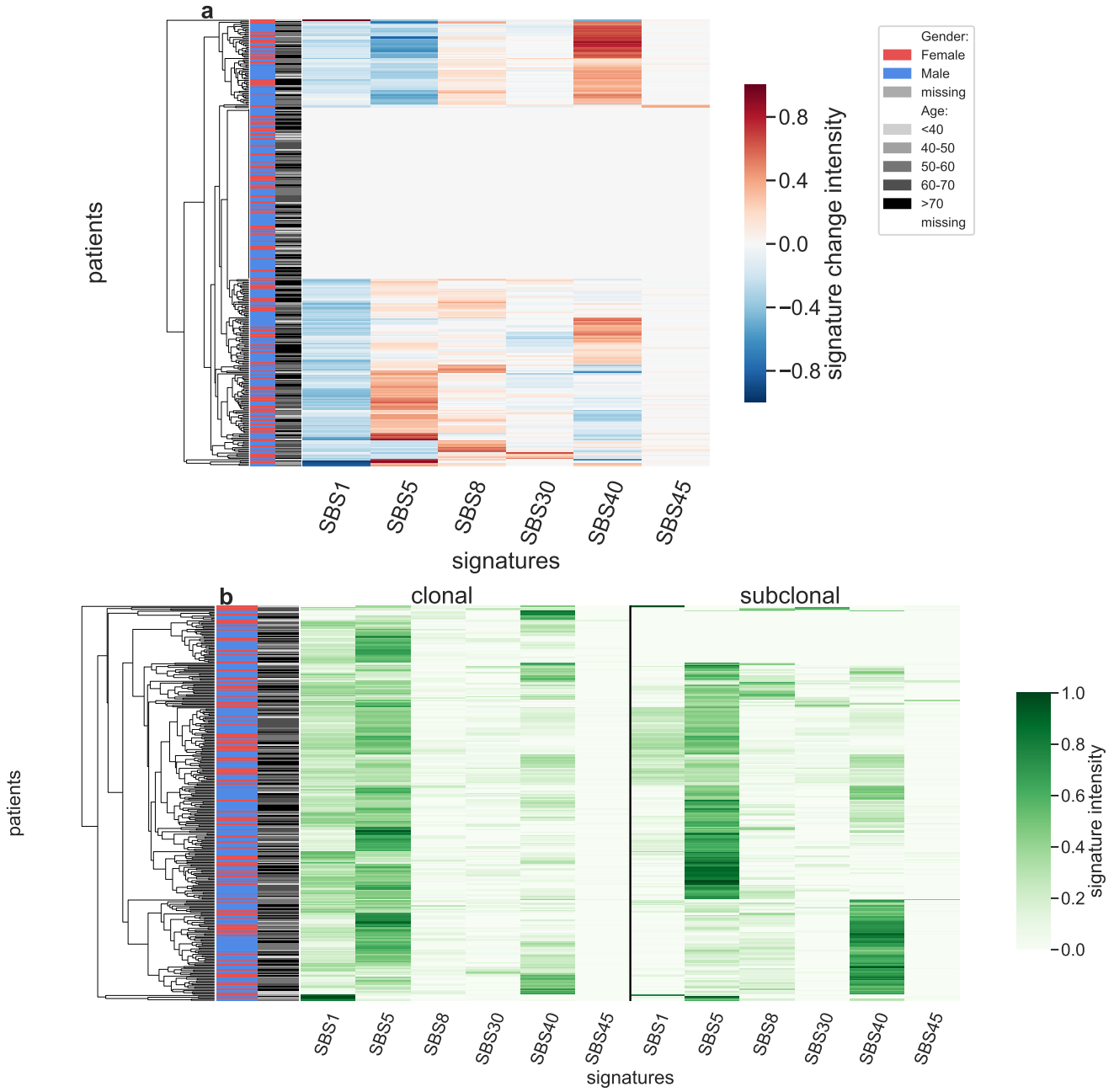


Figure S34: Panel a: Stratification of patients depending on their pattern of signature change for GBM patients (327 patients, including 202 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

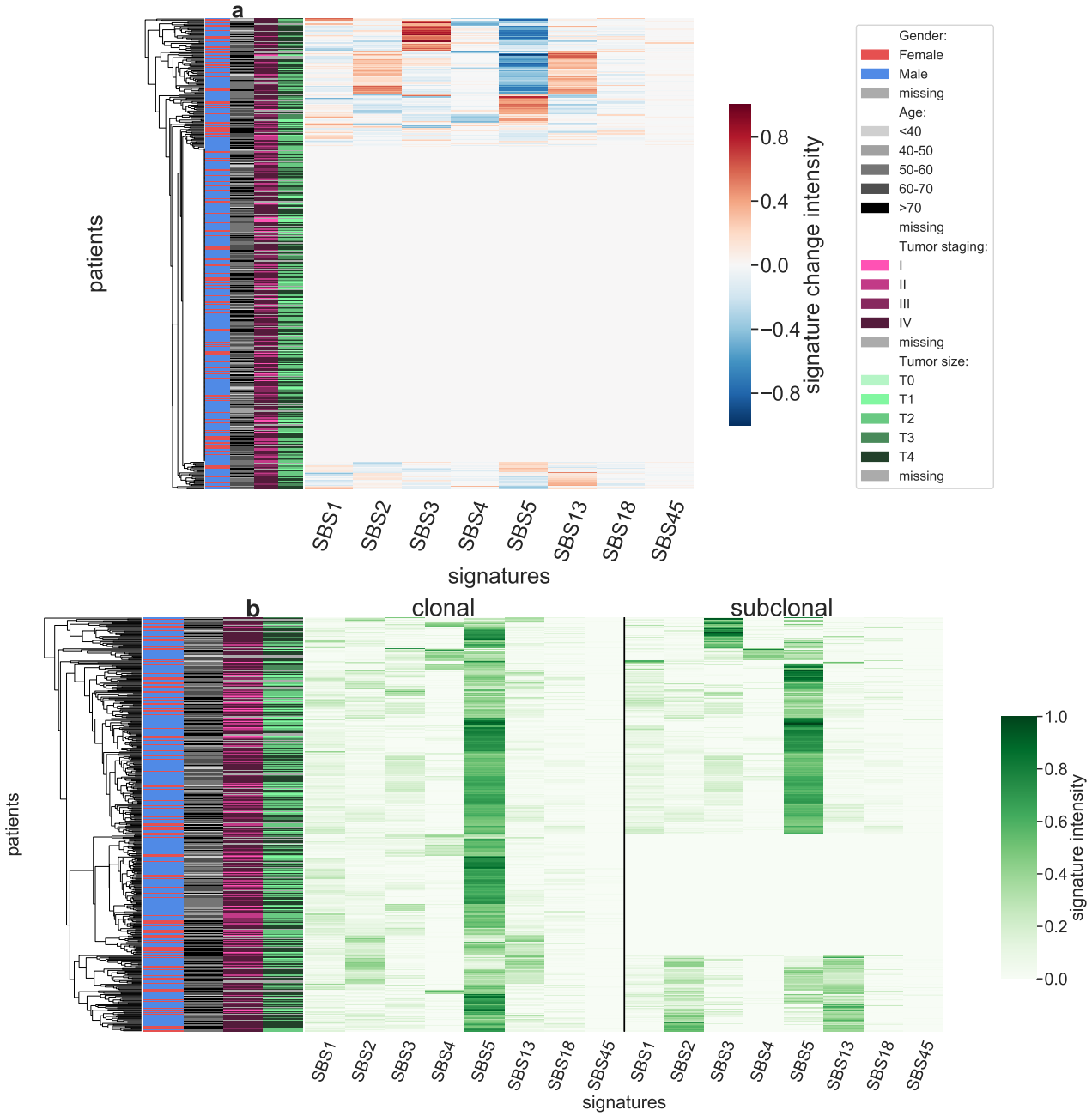


Figure S35: Panel a: Stratification of patients depending on their pattern of signature change for HNSC patients (445 patients, including 148 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

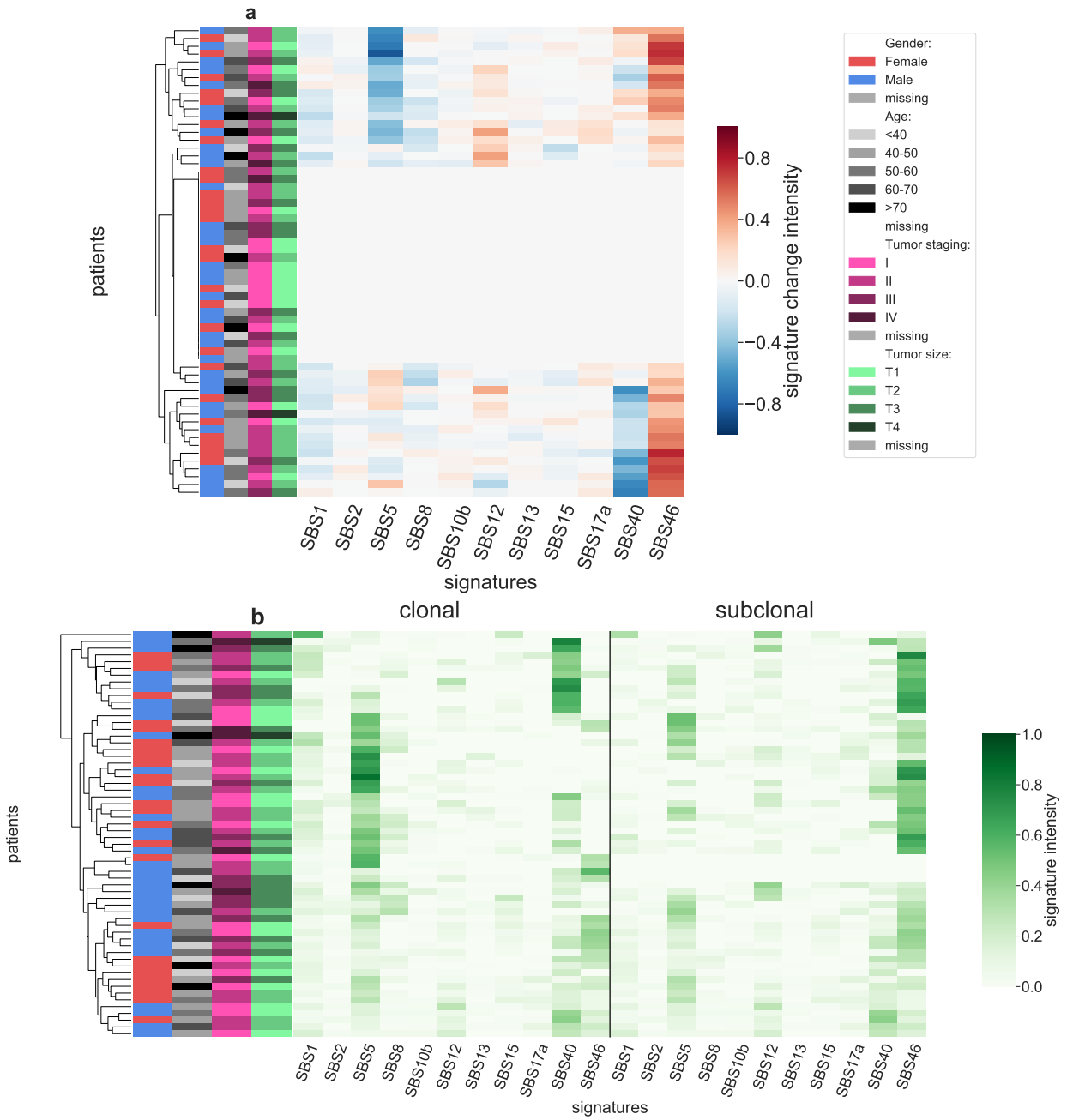


Figure S36: Panel a: Stratification of patients depending on their pattern of signature change for KICH patients (60 patients, including 35 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

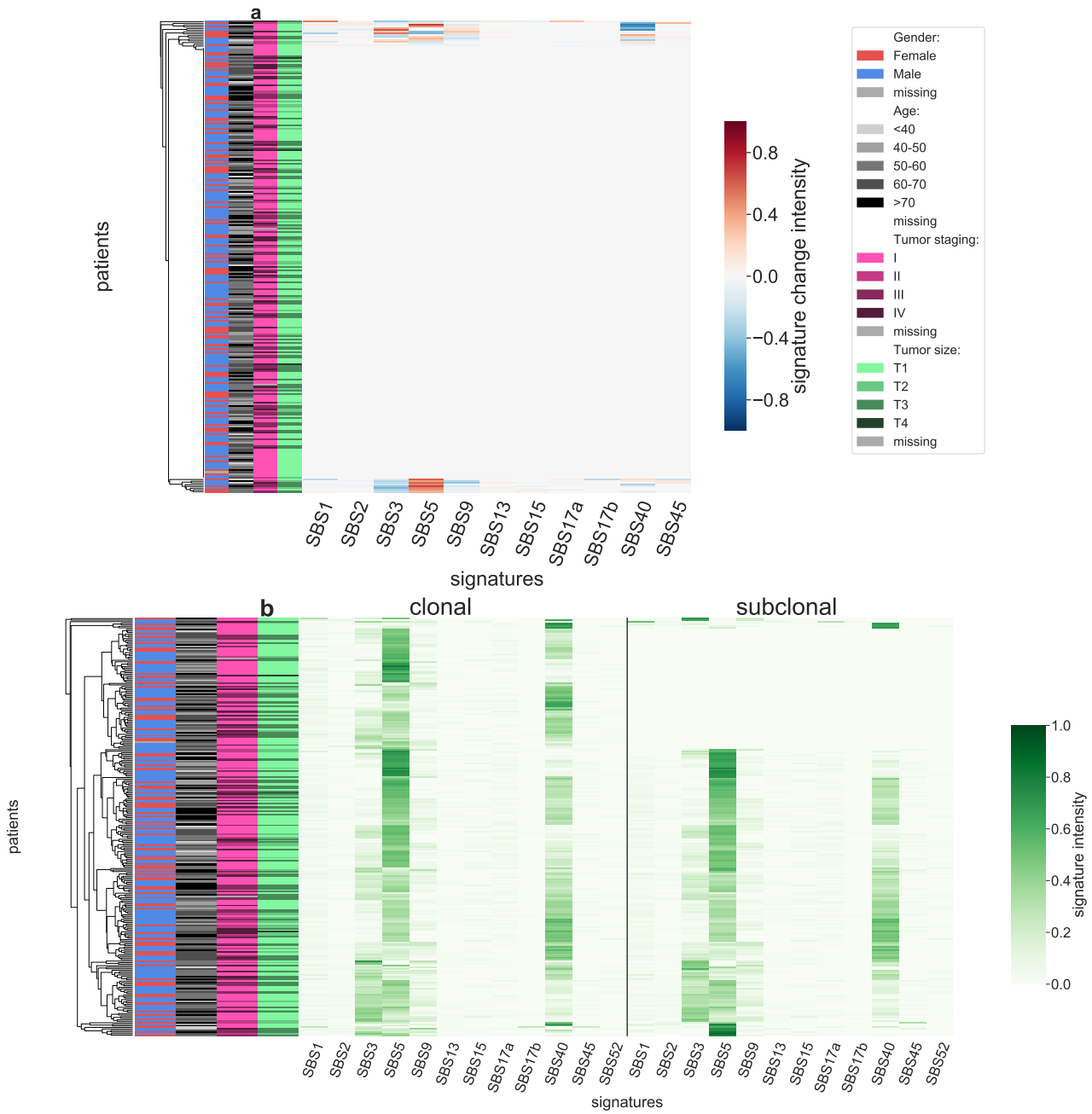


Figure S37: Panel a: Stratification of patients depending on their pattern of signature change for KIRC patients (271 patients, including 23 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

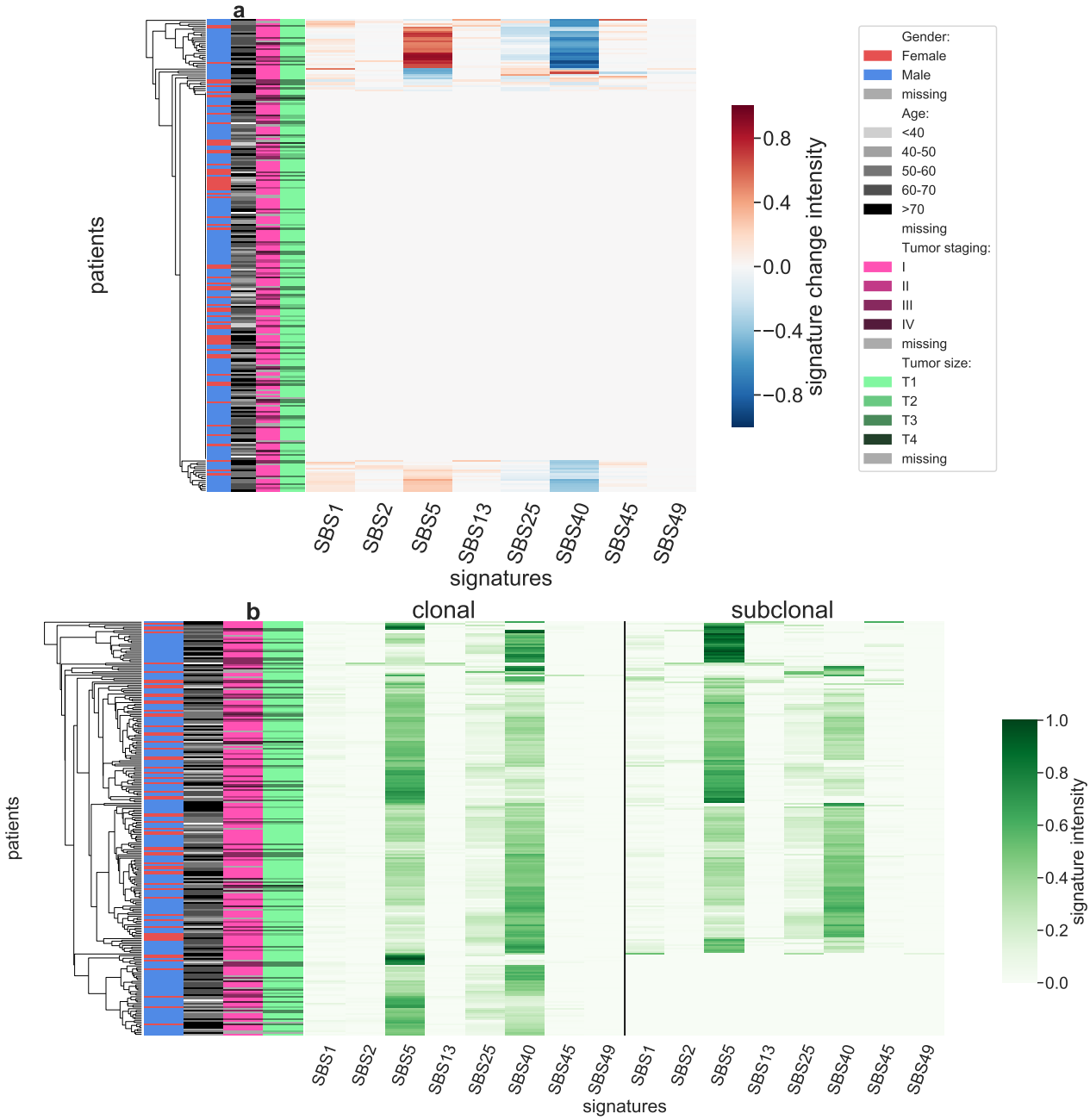


Figure S38: Panel a: Stratification of patients depending on their pattern of signature change for KIRP patients (242 patients, including 53 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

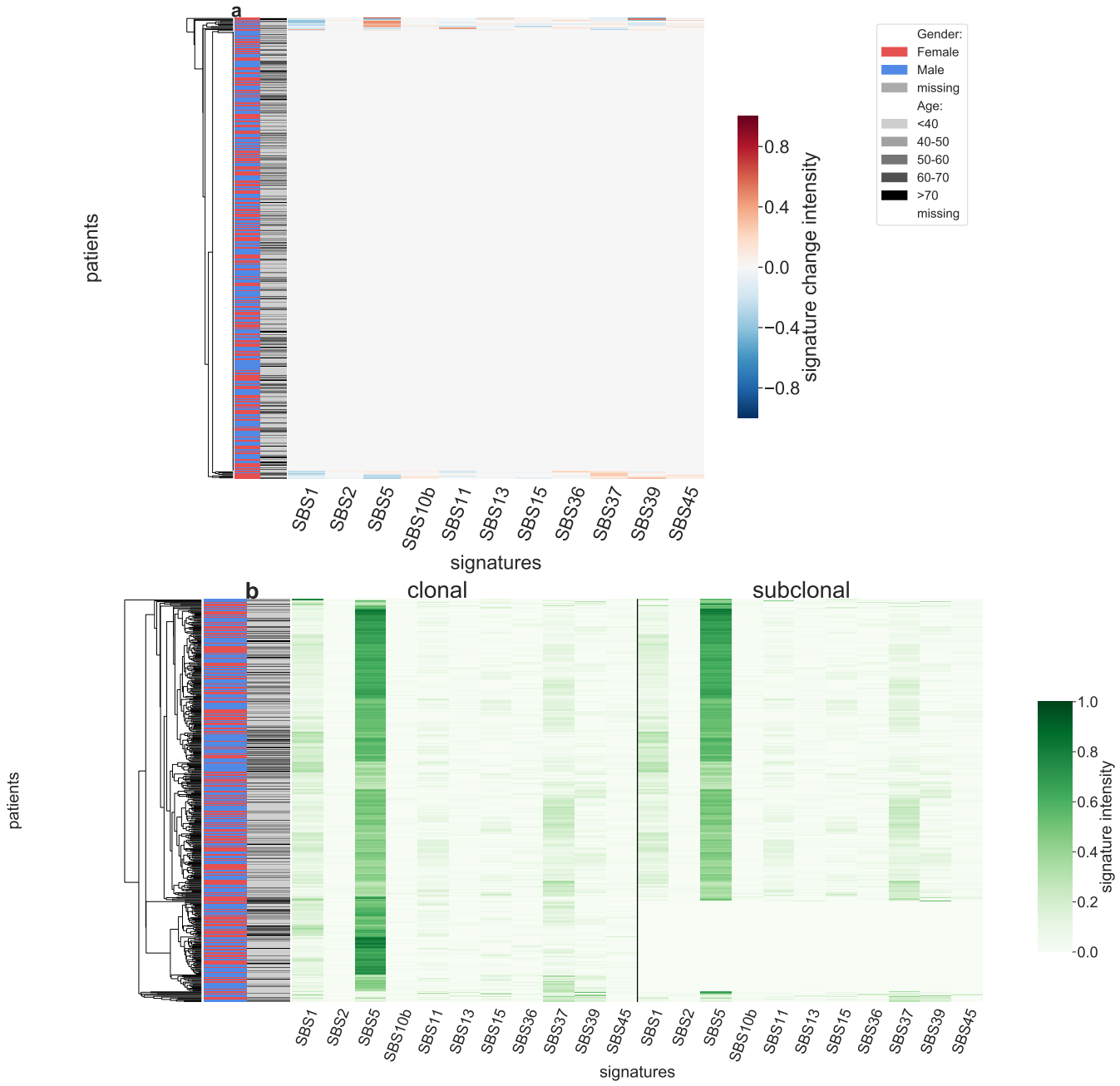


Figure S39: Panel a: Stratification of patients depending on their pattern of signature change for LGG patients (455 patients, including 20 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

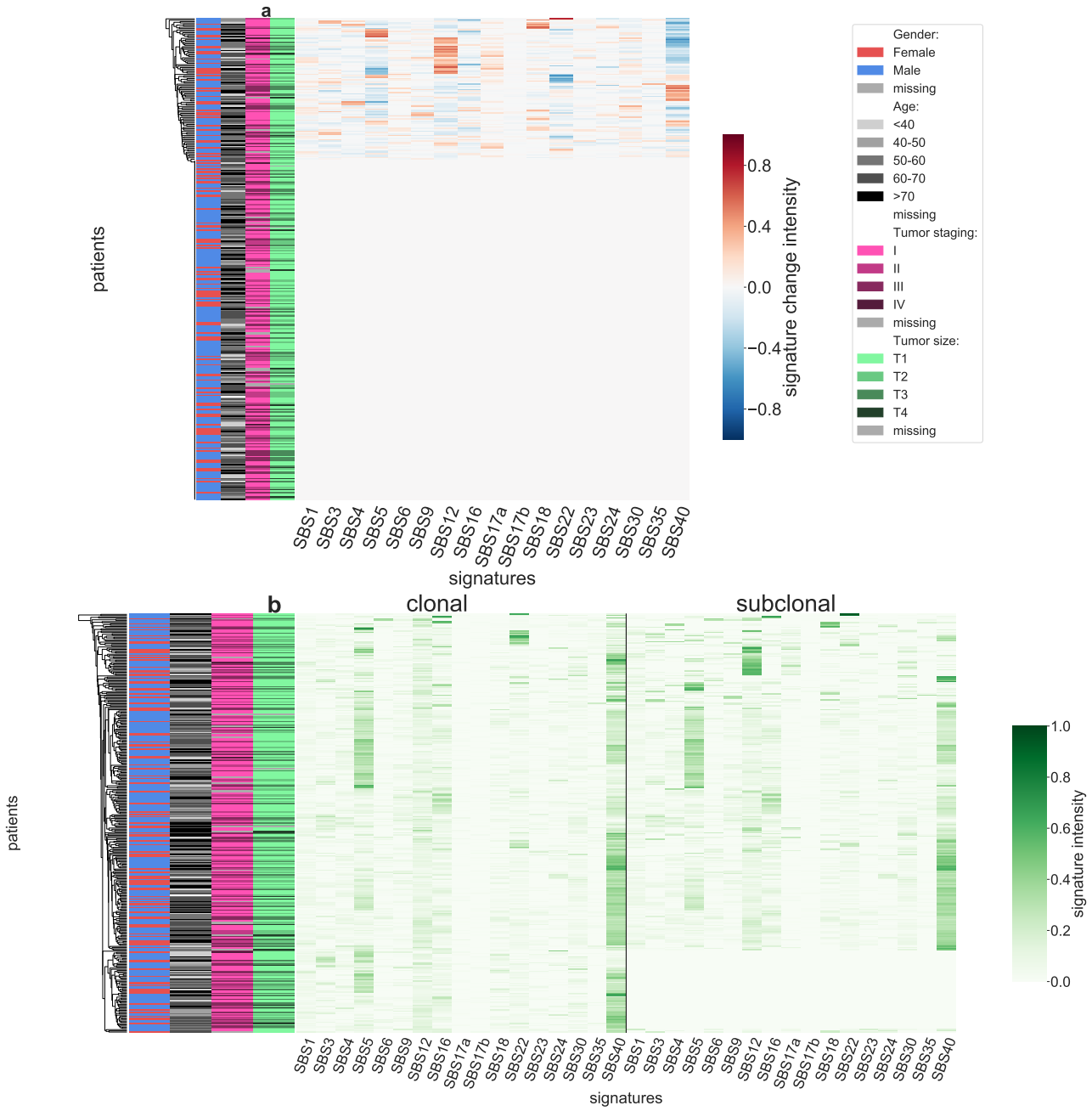


Figure S40: Panel a: Stratification of patients depending on their pattern of signature change for LIHC patients (347 patients, including 102 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

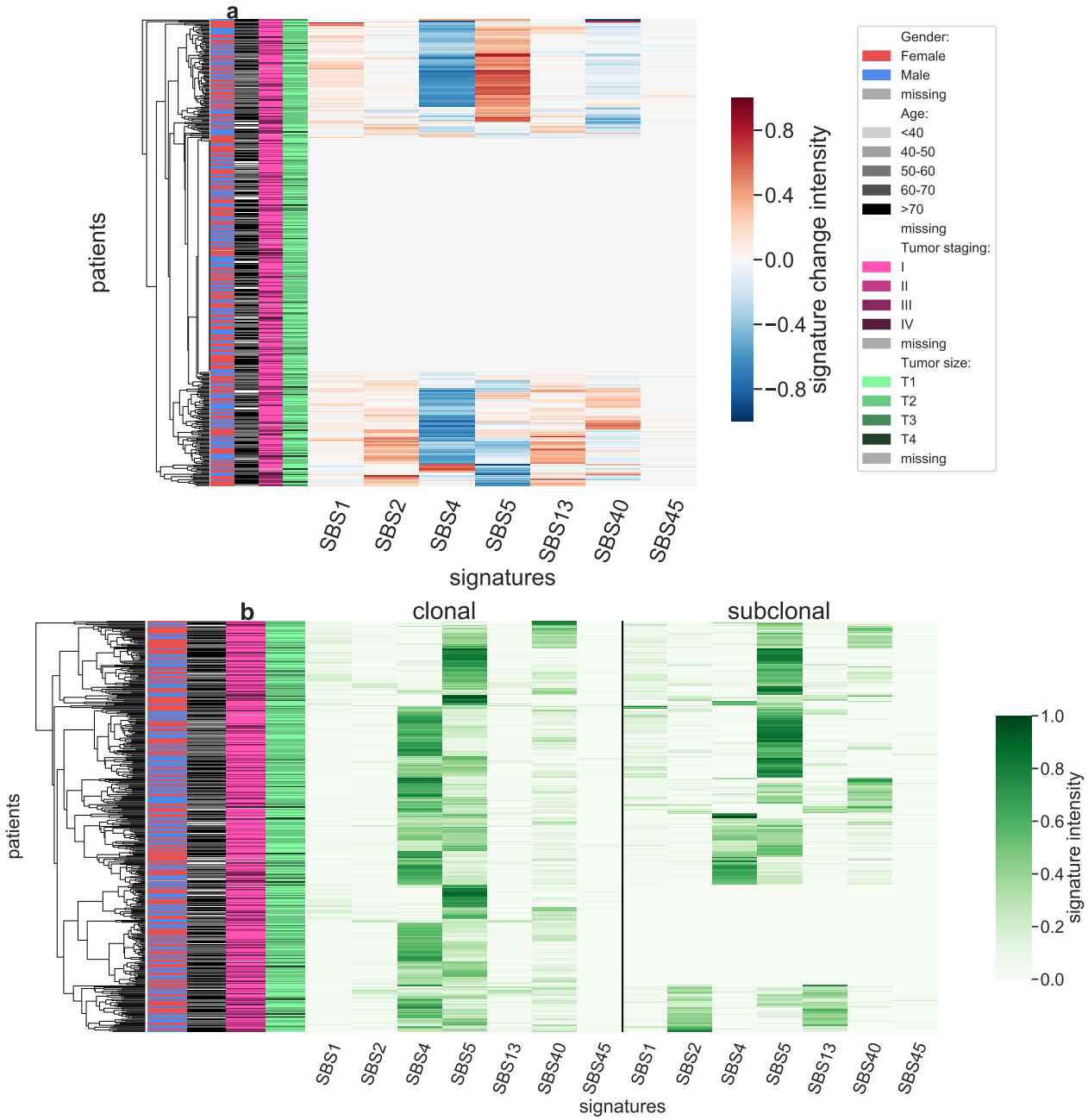


Figure S41: Panel a: Stratification of patients depending on their pattern of signature change for LUAD patients (433 patients, including 217 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

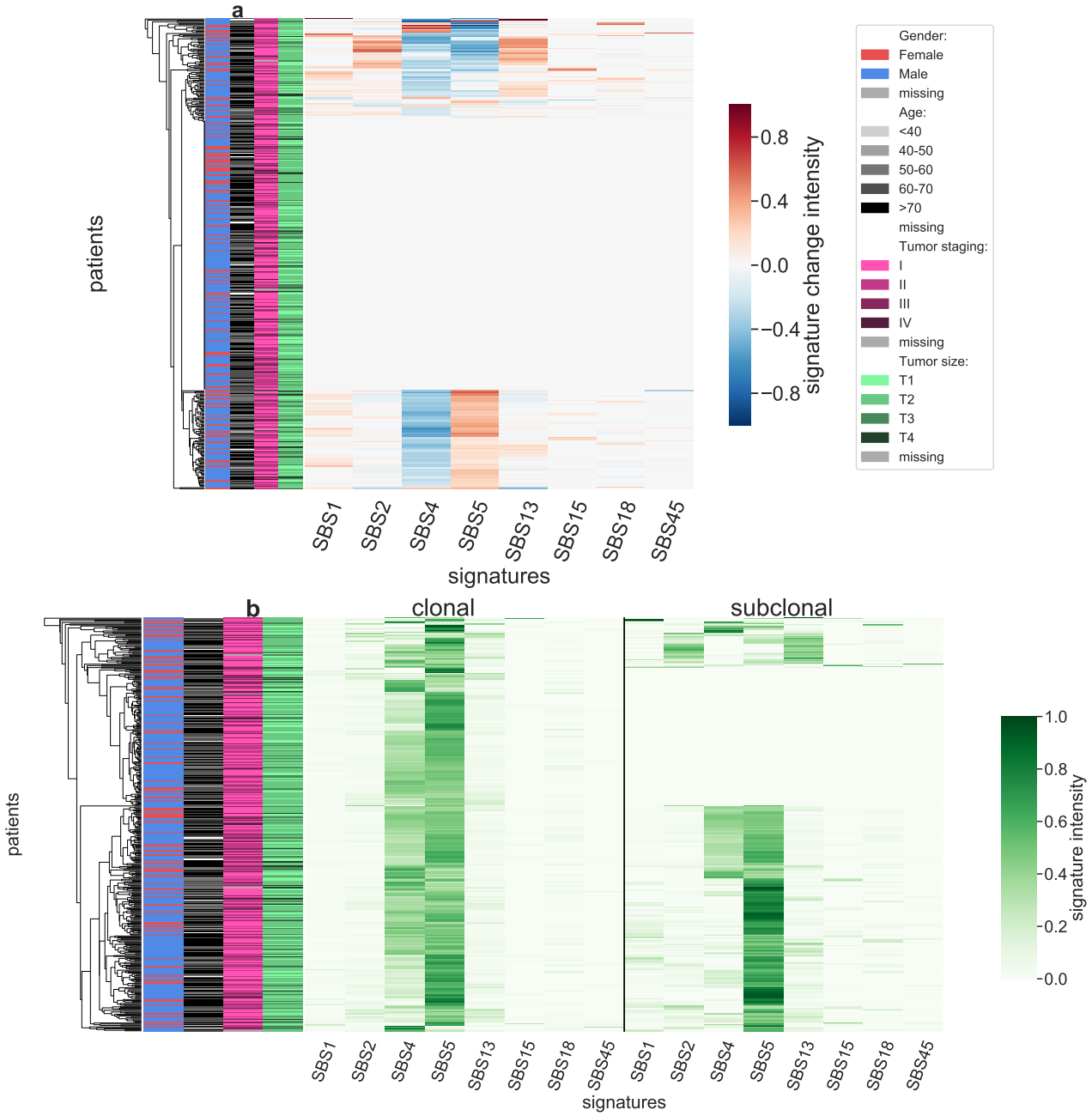


Figure S42: Panel a: Stratification of patients depending on their pattern of signature change for LUSC patients (423 patients, including 180 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

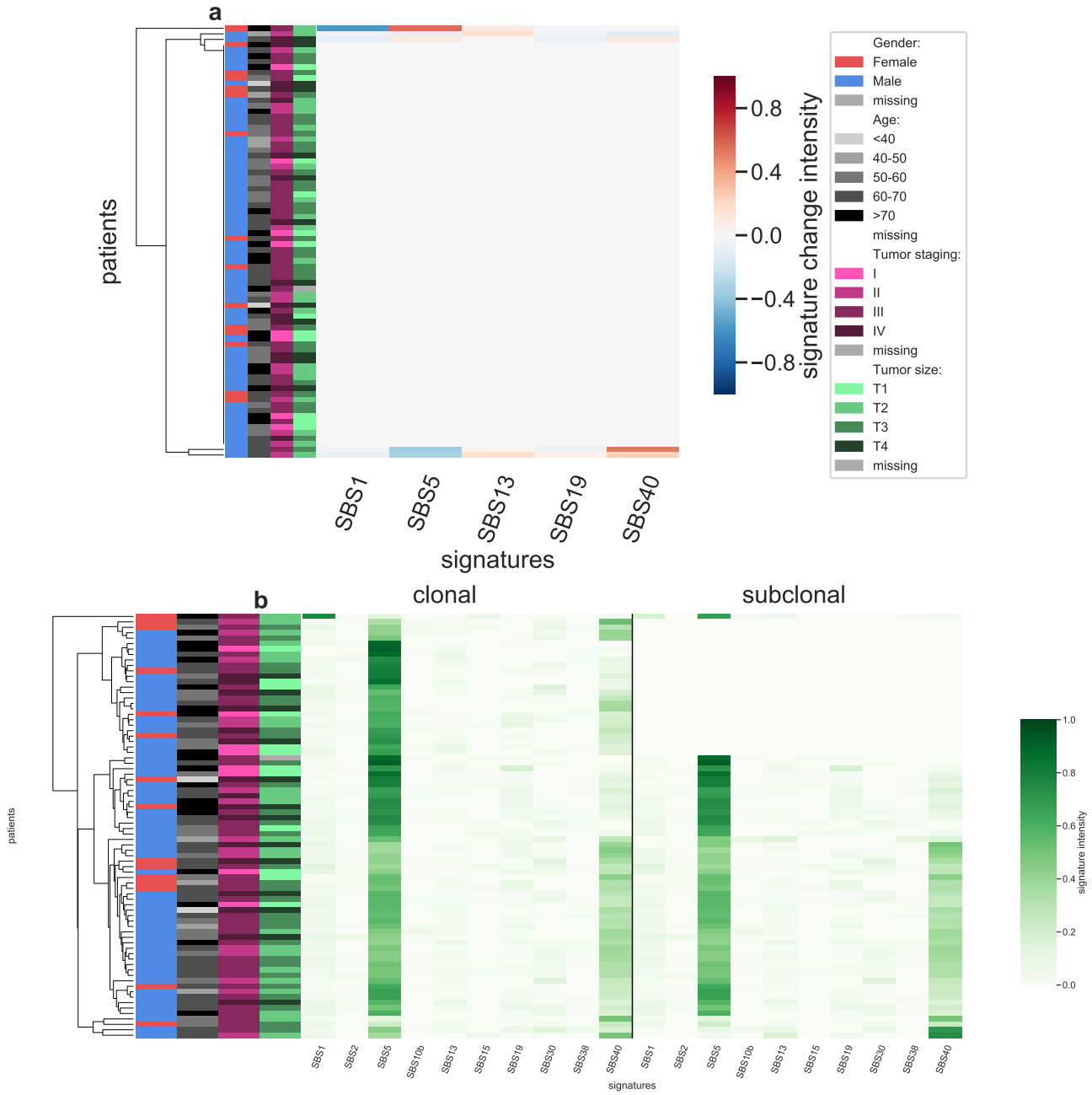


Figure S43: Panel a: Stratification of patients depending on their pattern of signature change for MESO patients (78 patients, including 5 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

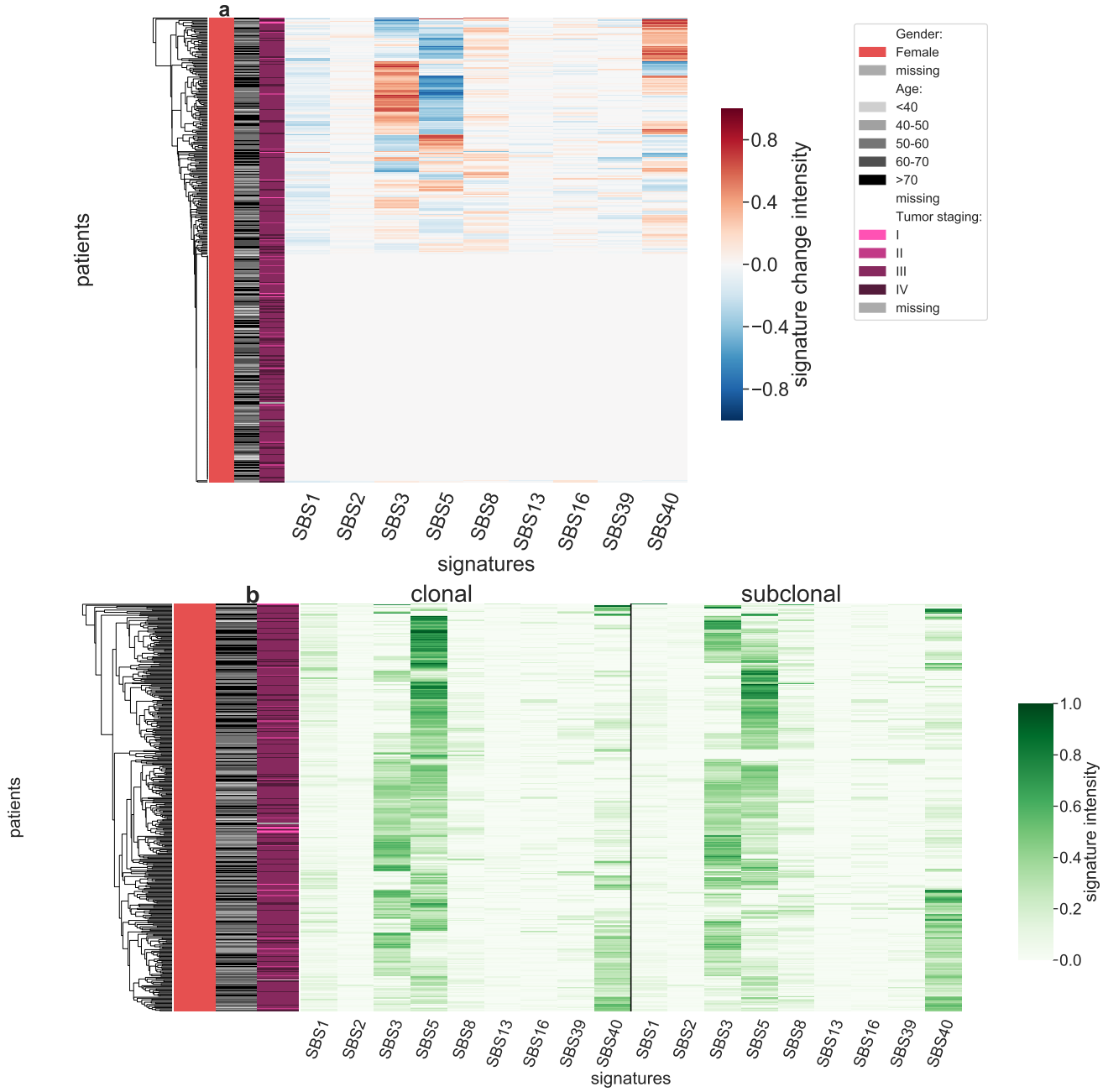


Figure S44: Panel a: Stratification of patients depending on their pattern of signature change for OV patients (390 patients, including 201 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

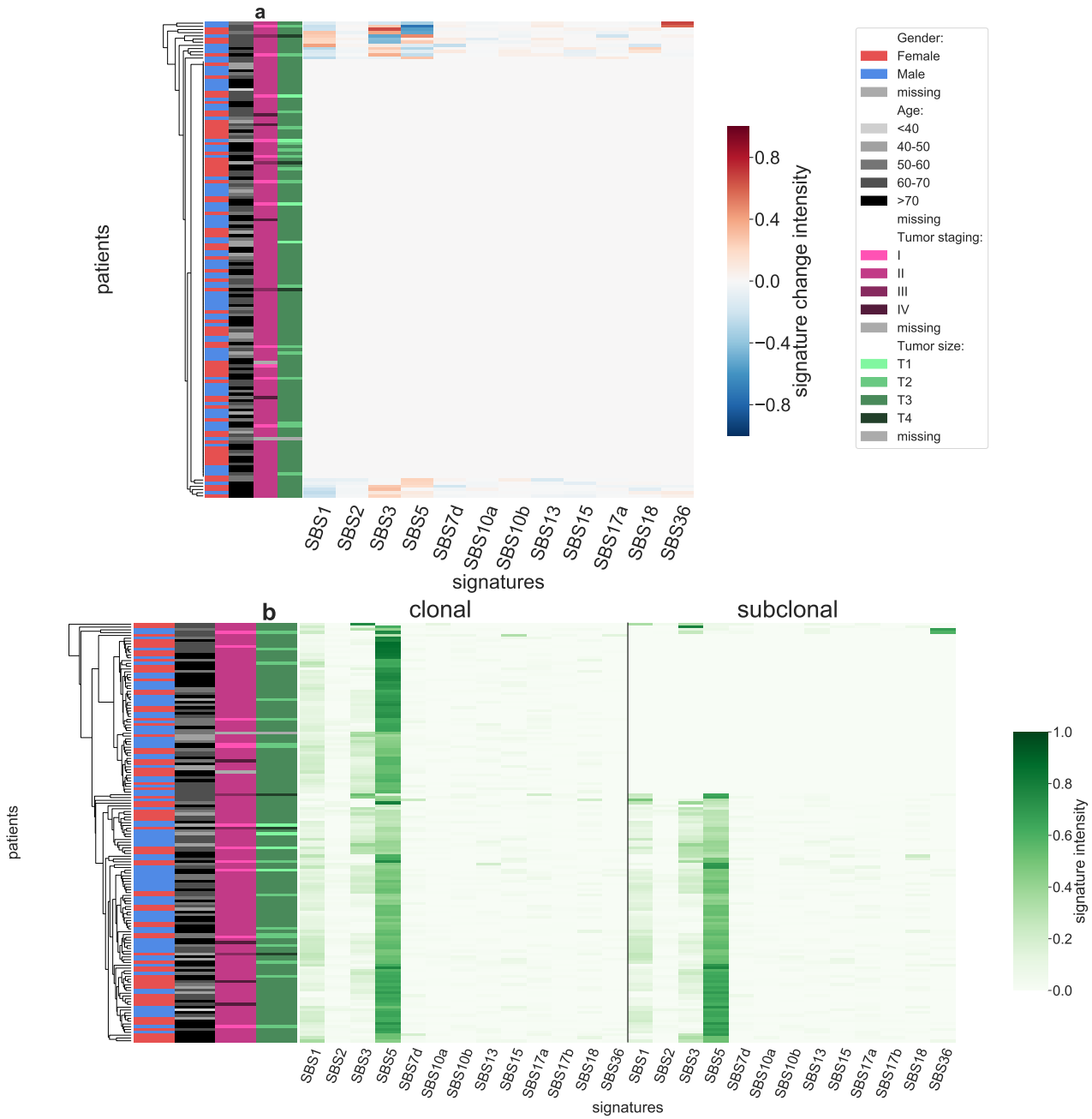
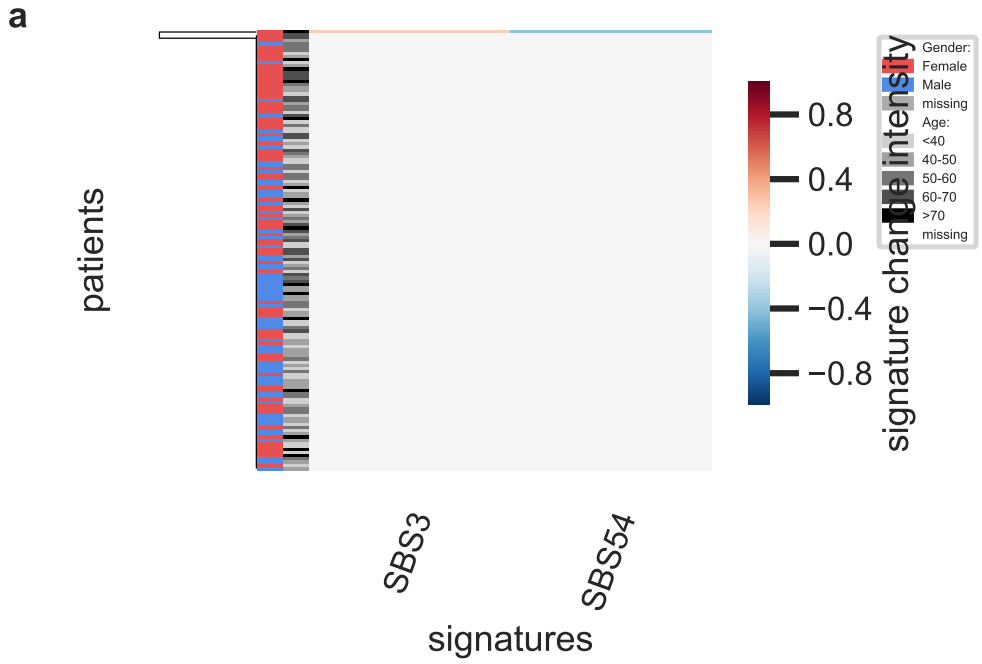


Figure S45: Panel a: Stratification of patients depending on their pattern of signature change for PAAD patients (150 patients, including 18 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).



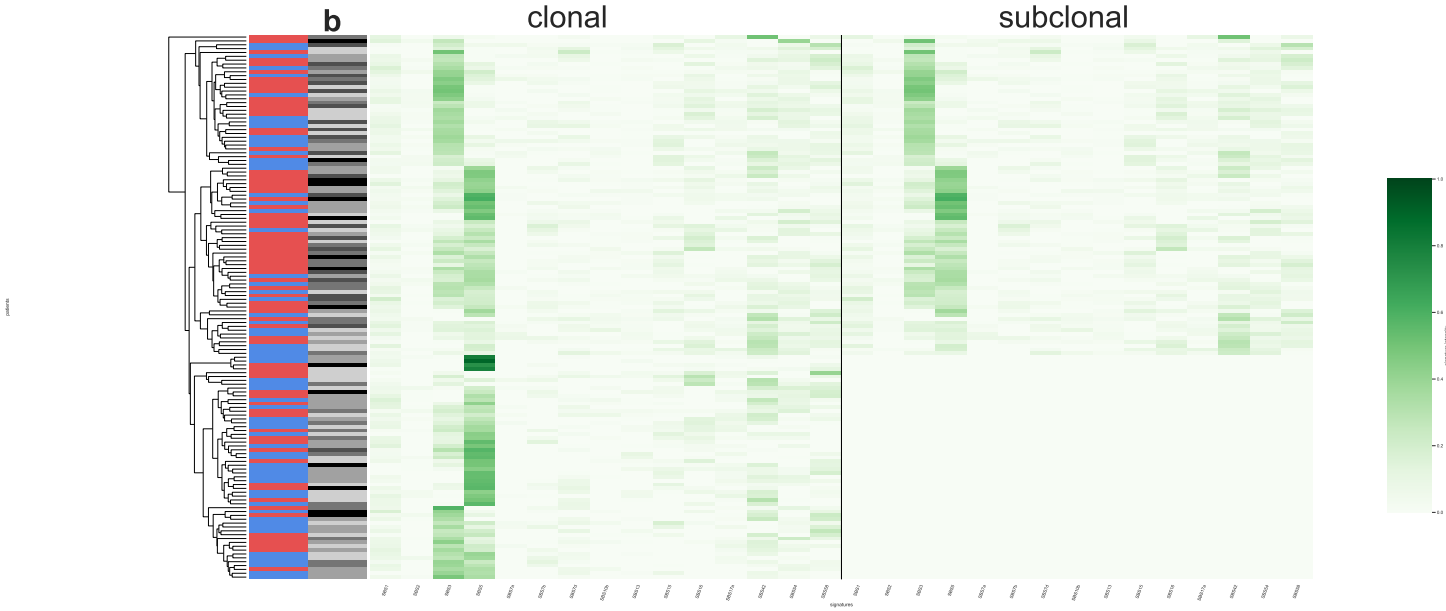


Figure S46: Panel a: Stratification of patients depending on their pattern of signature change for PCPG patients (141 patients, including 1 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

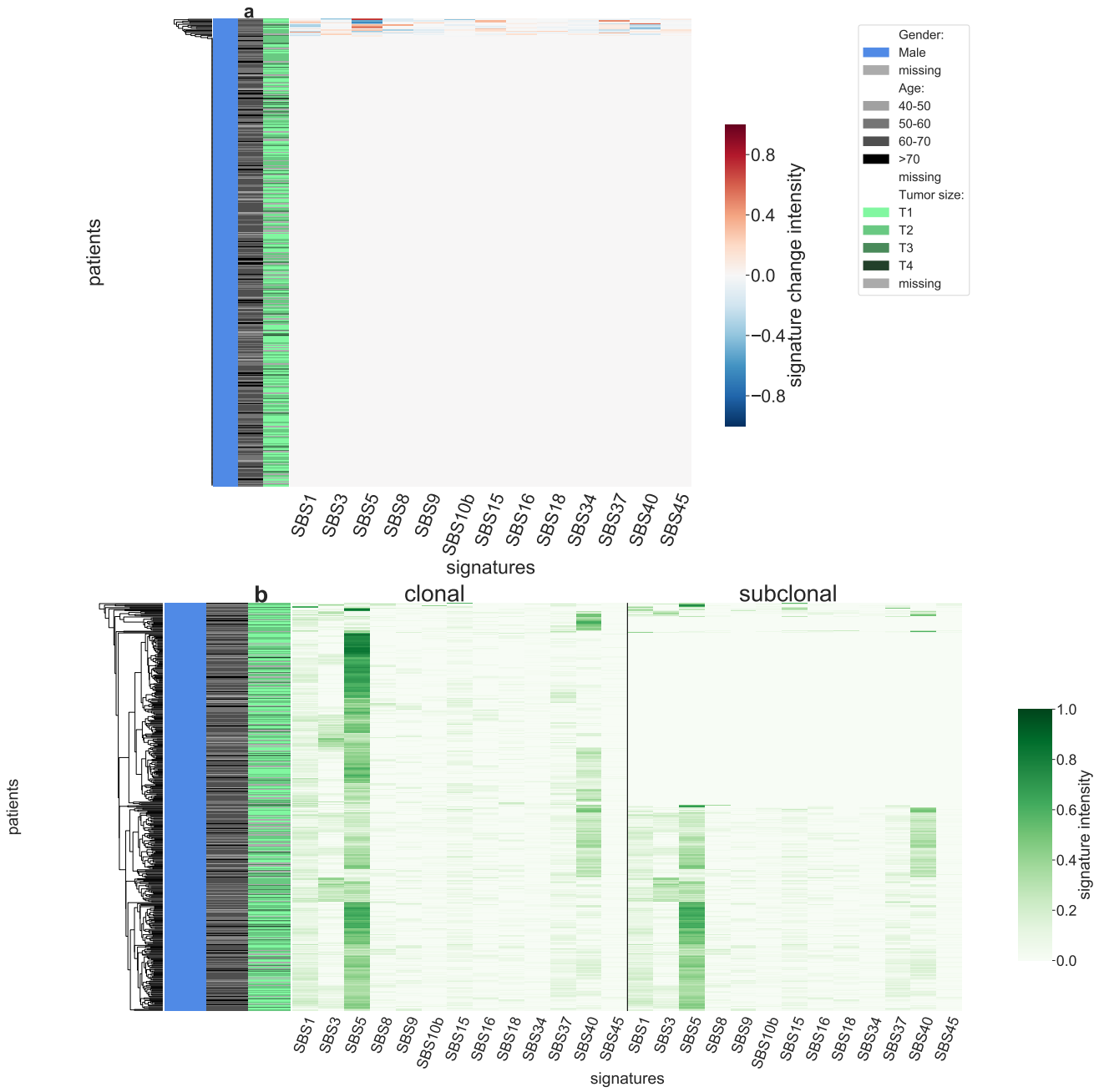


Figure S47: Panel a: Stratification of patients depending on their pattern of signature change for PRAD patients (458 patients, including 18 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

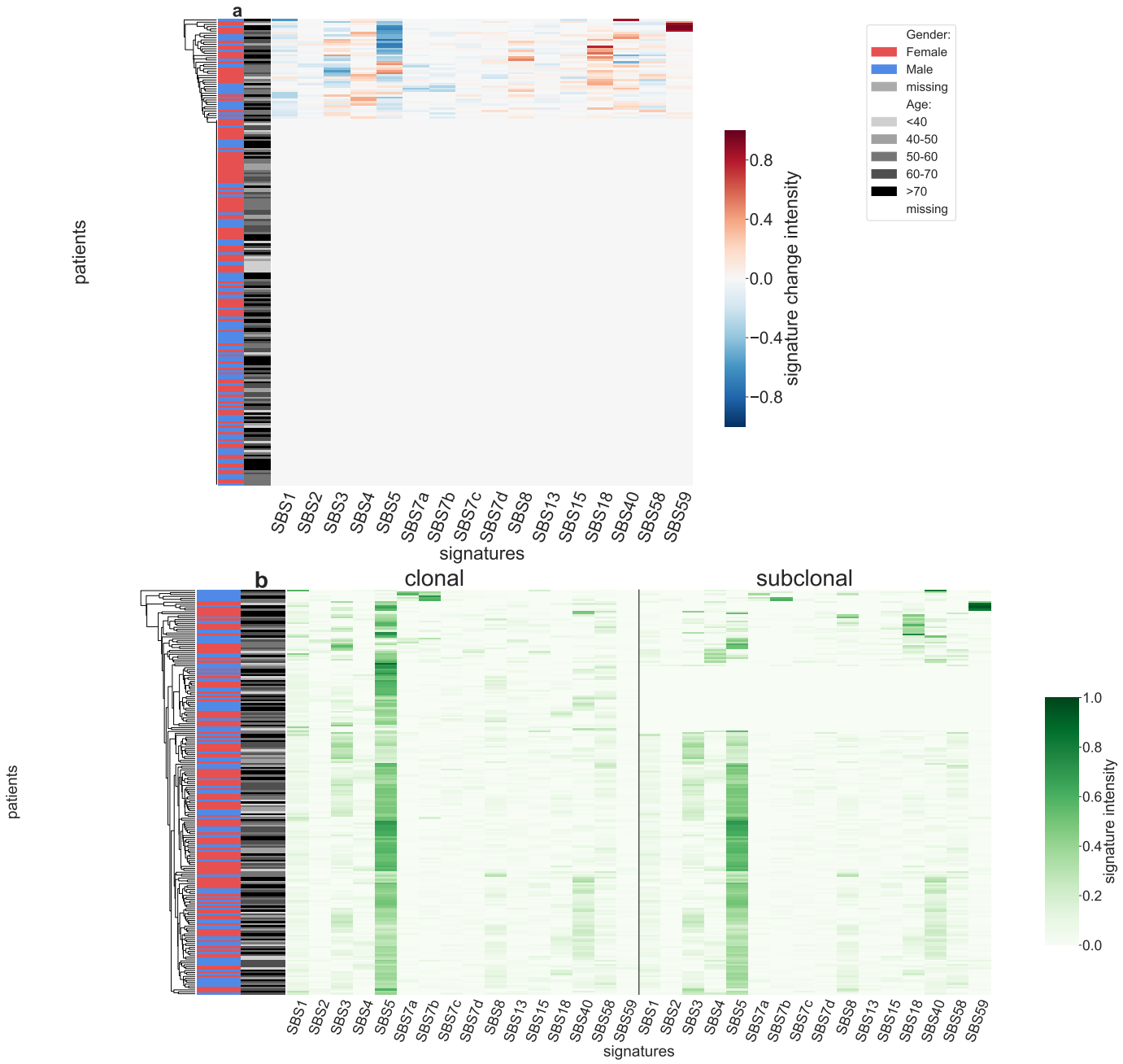


Figure S48: Panel a: Stratification of patients depending on their pattern of signature change for SARC patients (210 patients, including 45 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

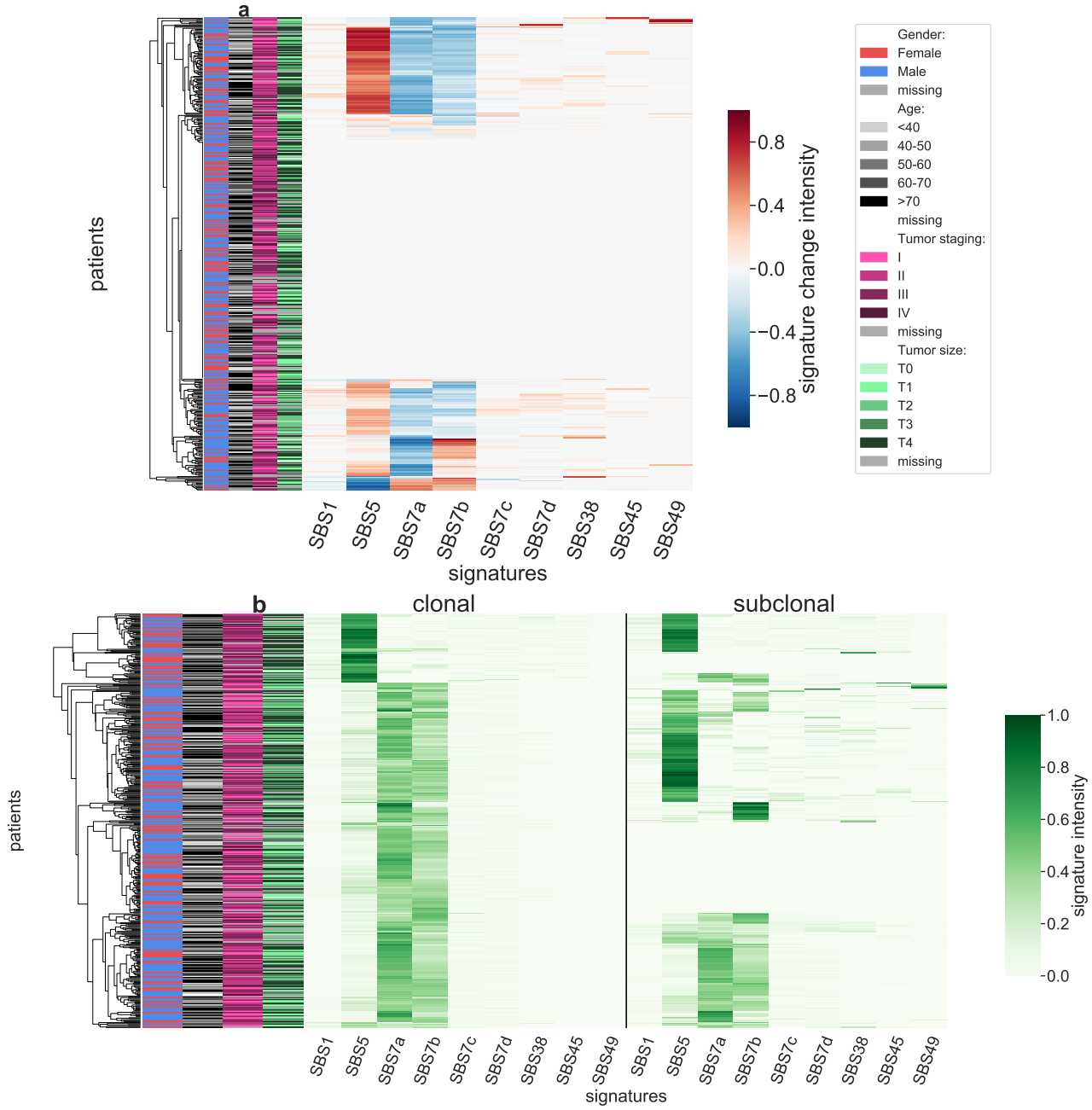


Figure S49: Panel a: Stratification of patients depending on their pattern of signature change for SKCM patients (423 patients, including 210 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

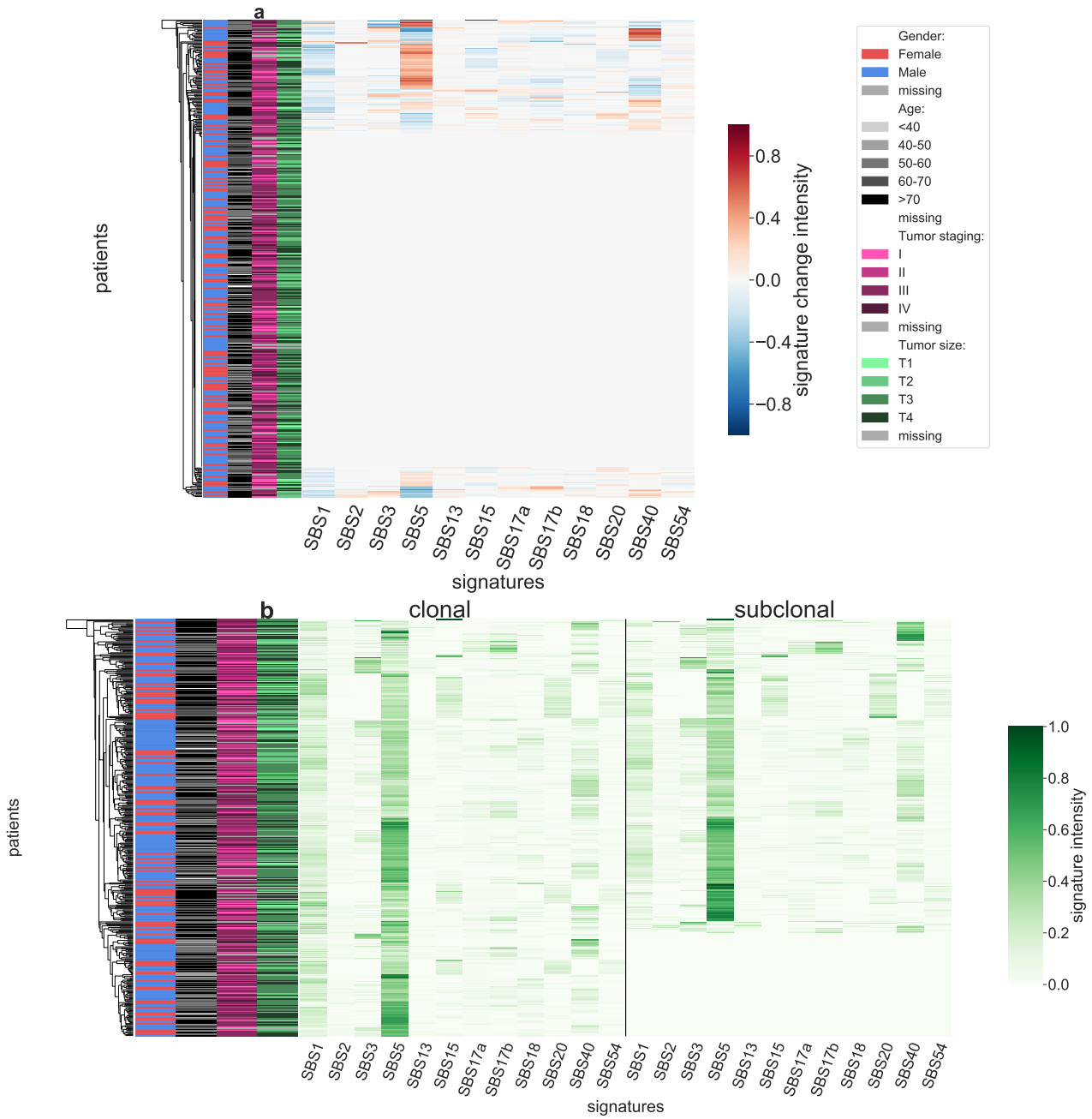


Figure S50: Panel a: Stratification of patients depending on their pattern of signature change for STAD patients (418 patients, including 127 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

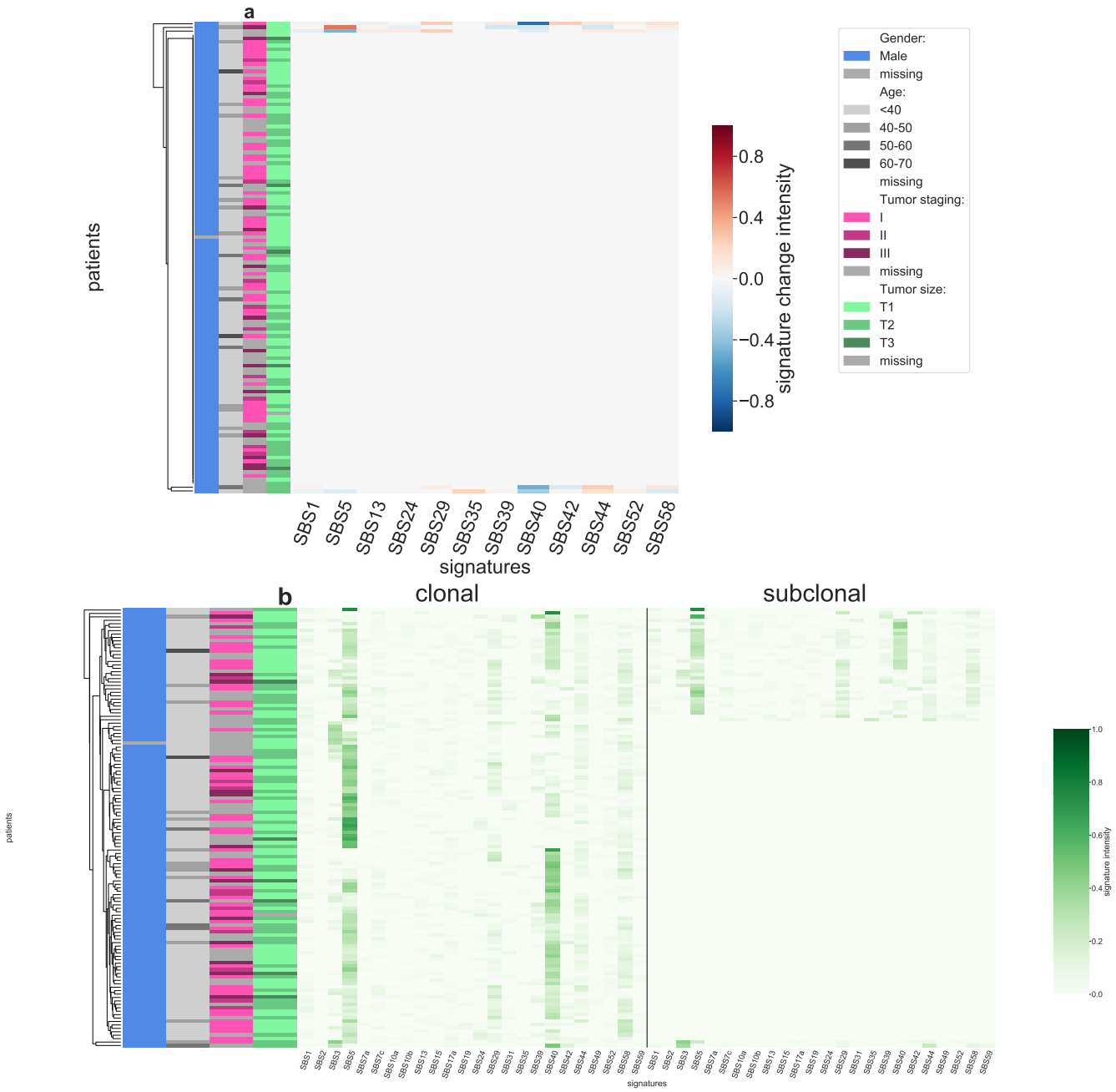


Figure S51: Panel a: Stratification of patients depending on their pattern of signature change for TGCT patients (128 patients, including 5 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

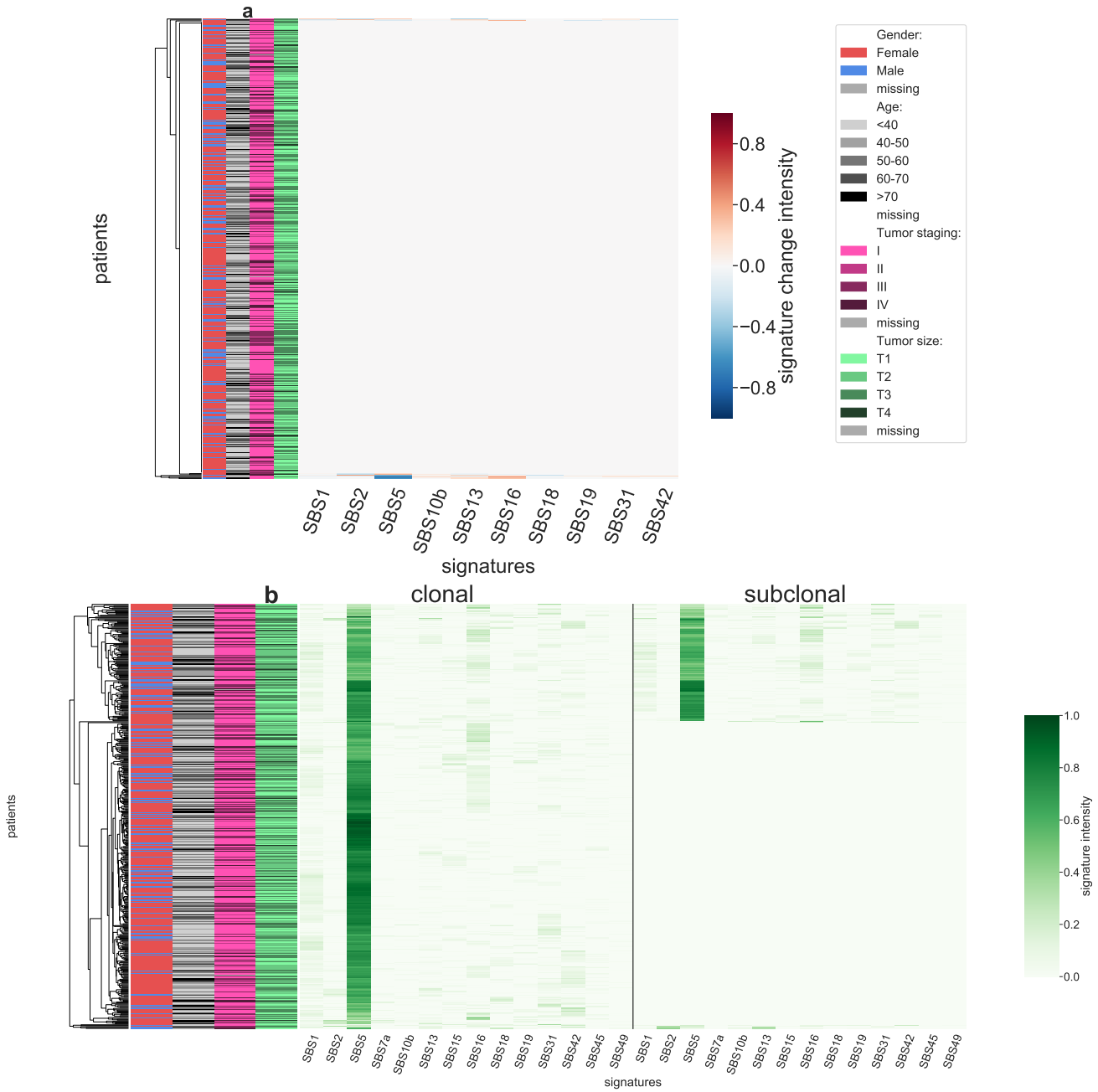


Figure S52: Panel a: Stratification of patients depending on their pattern of signature change for THCA patients (467 patients, including 9 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

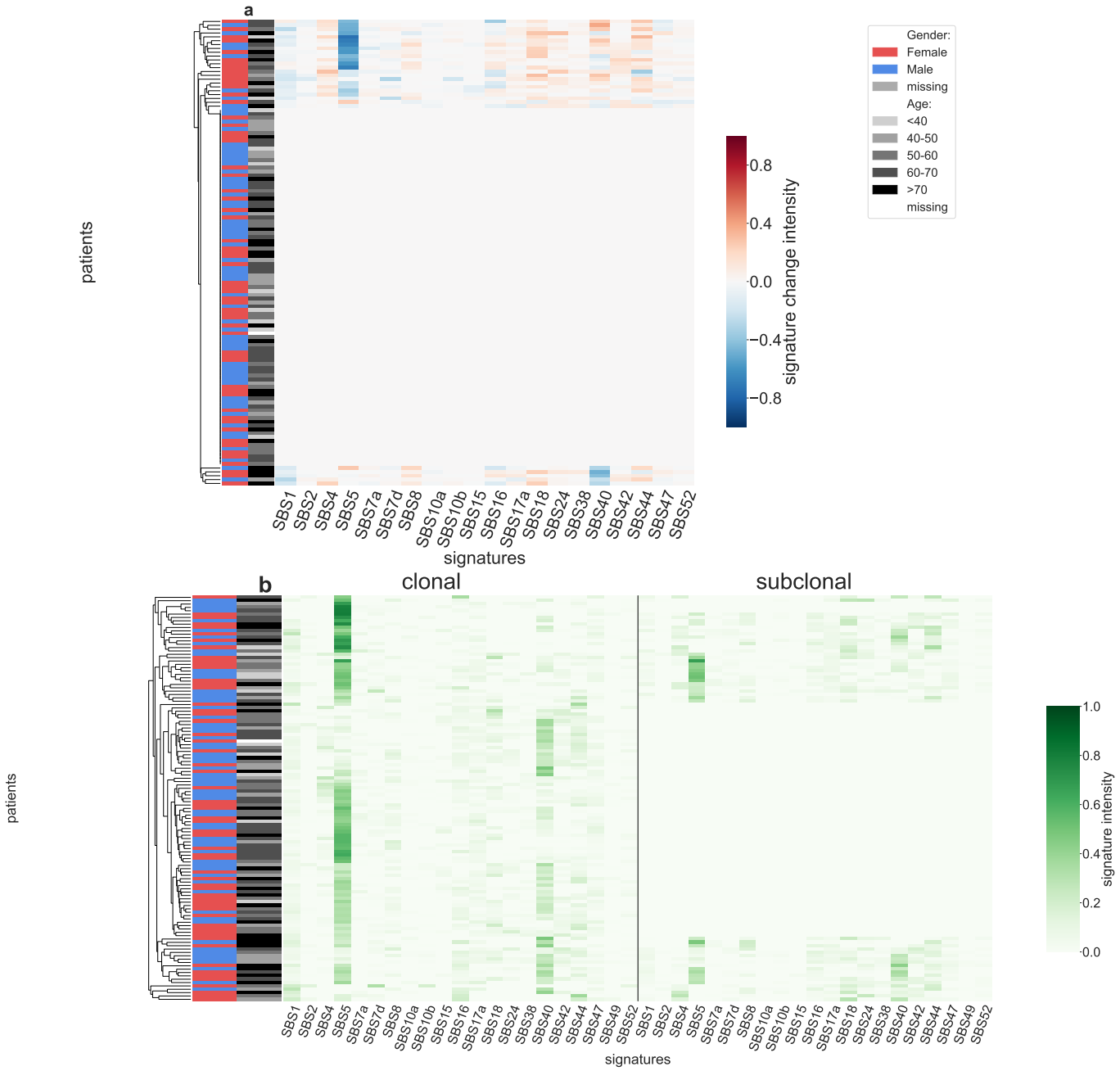


Figure S53: Panel a: Stratification of patients depending on their pattern of signature change for THYM patients (121 patients, including 28 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

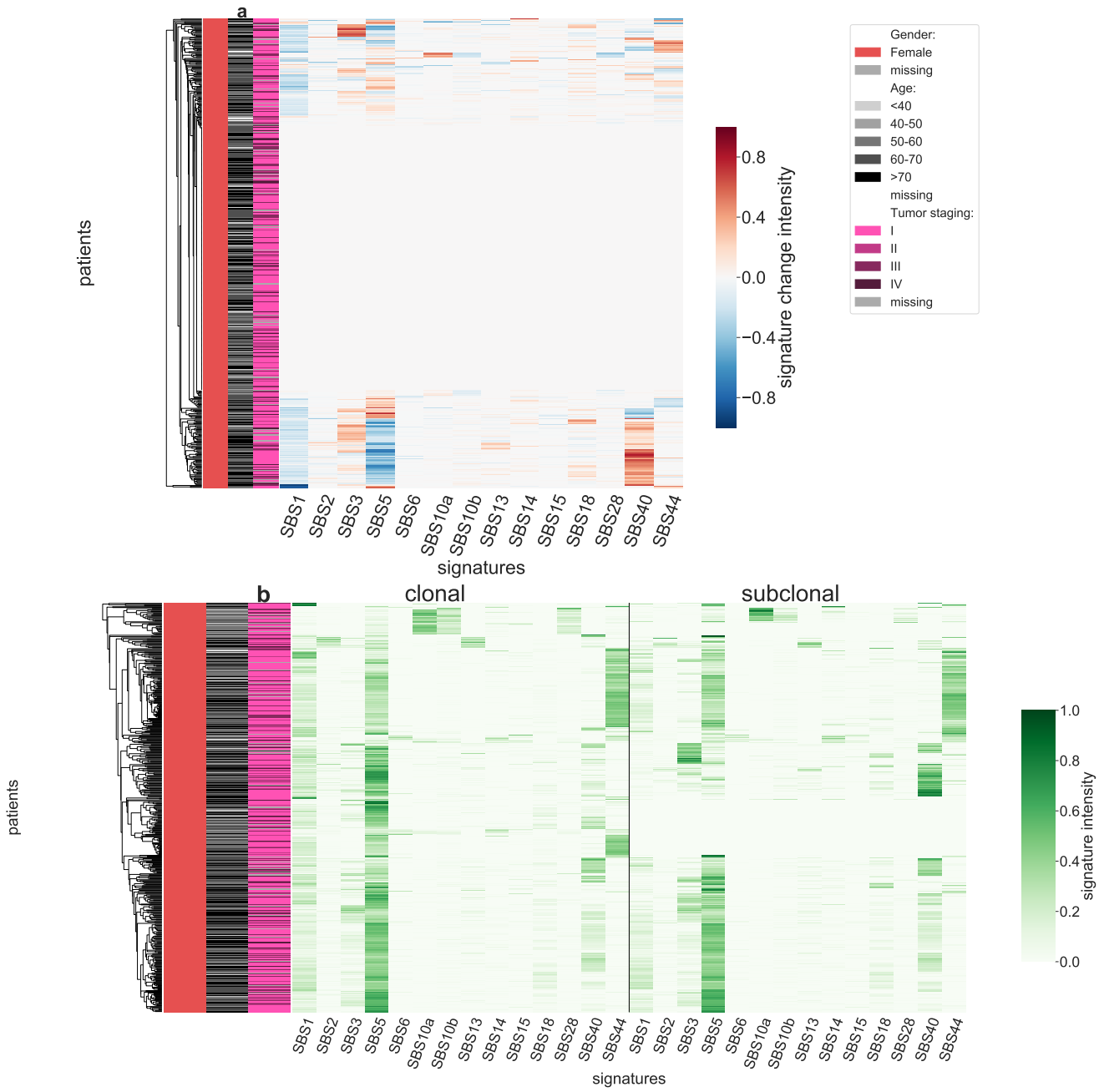


Figure S54: Panel a: Stratification of patients depending on their pattern of signature change for UCEC patients (487 patients, including 213 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

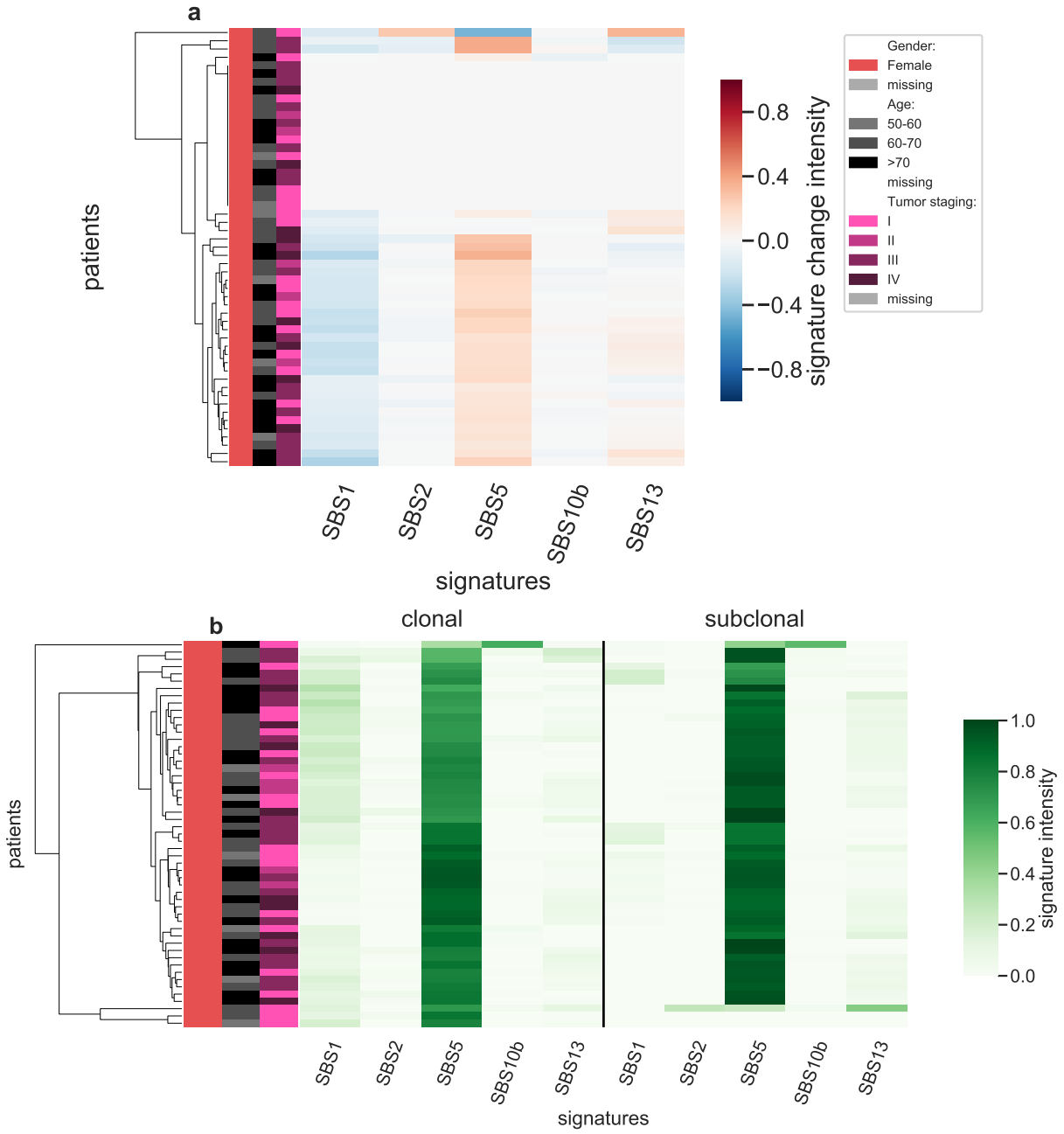


Figure S55: Panel a: Stratification of patients depending on their pattern of signature change for UCS patients (53 patients, including 35 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

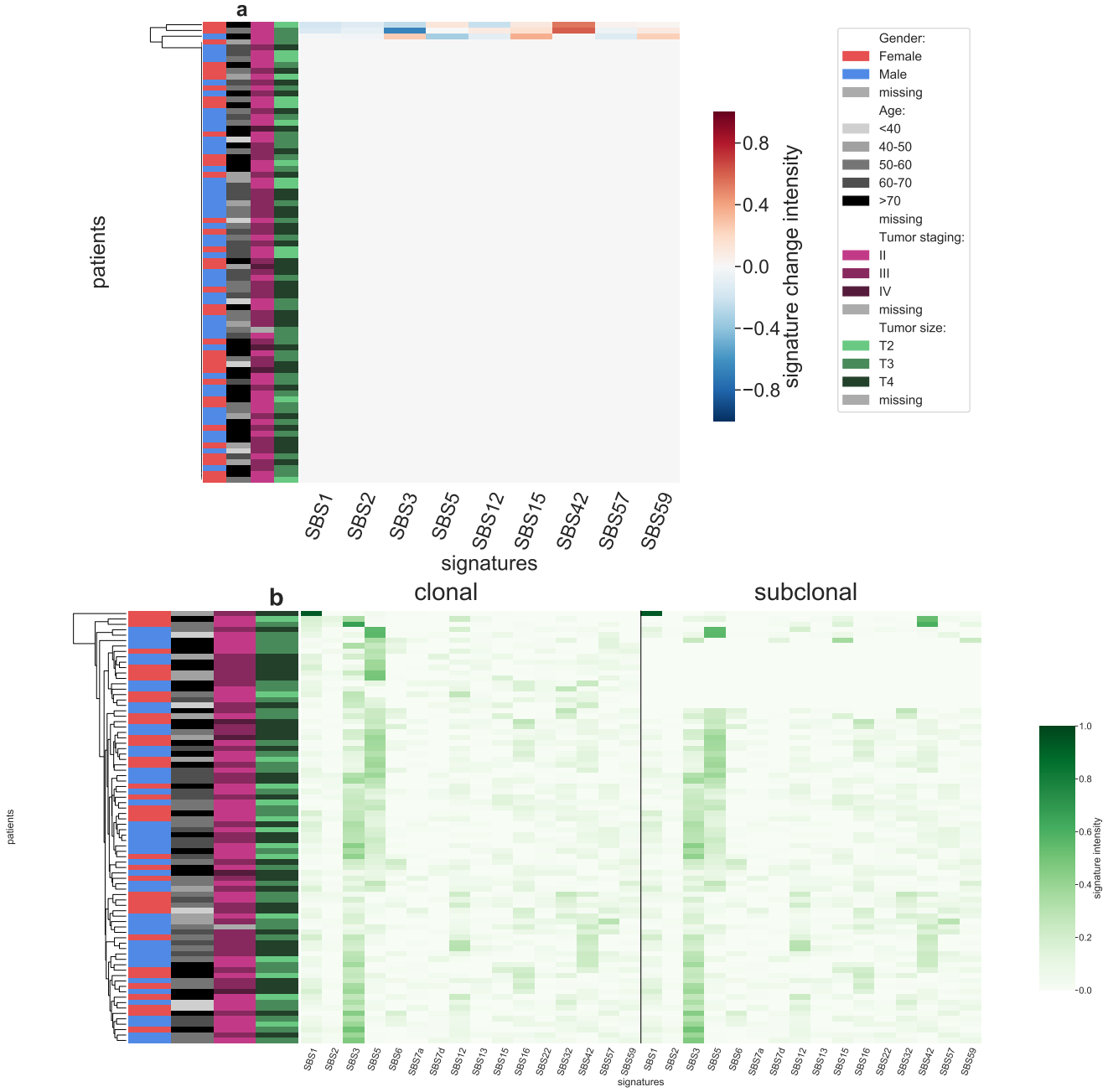


Figure S56: Panel a: Stratification of patients depending on their pattern of signature change for UVM patients (80 patients, including 3 with a significant signature change). The heatmap represents the difference between the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).Panel b: Stratification of patients depending on their complete pattern of signature exposure. The heatmap represents the signature activity in the largest subclone (in terms of number of mutations) and the clonal mutations (defined as belonging to the clone of highest CCF).

Supplementary figures

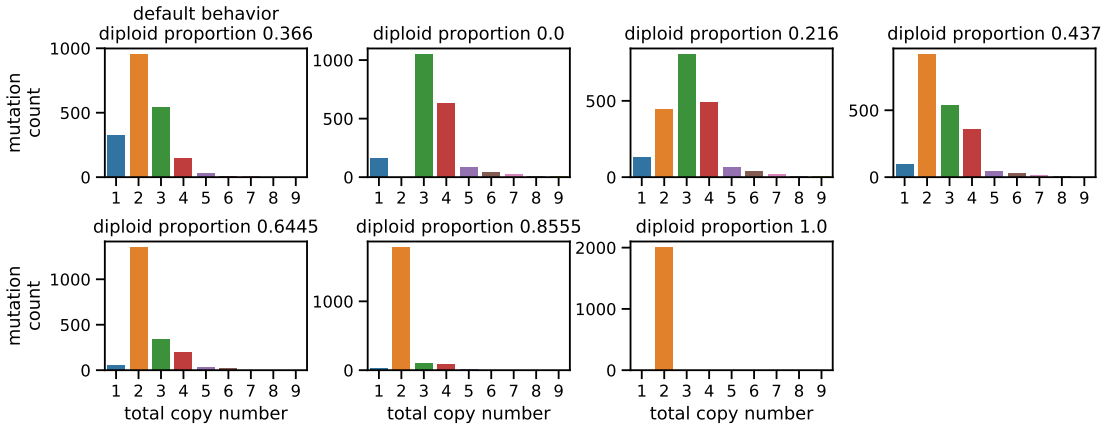


Figure S57: An example of empirical distribution of the total copy number for samples with 2000 mutations. In the first panel, labeled "default behavior", the user does not specify the percentage of genome that is diploid, and the total copy number values are drawn as specified in the Methods section. On the other panels, the user specifies a desired percentage of genome that is diploid (0, 0.2, 0.4, 0.6, 0.8, 1) respectively for the cases shown. The distribution is slightly different from the default behavior.

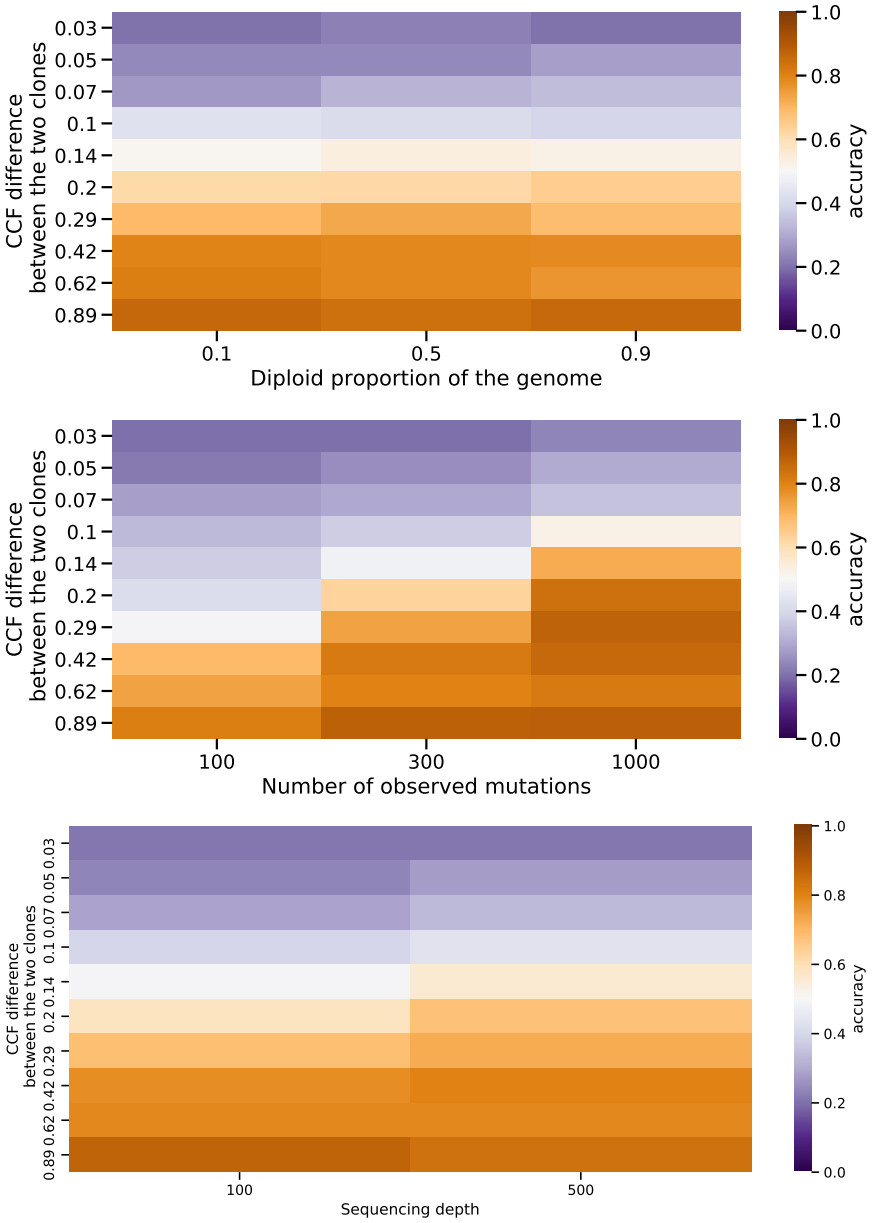


Figure S58: CloneSig's ability to distinguish 2 clones depending on the CCF distance between the two clones, and other relevant variables: number of mutations, and percentage of diploid genome, and sequencing depth

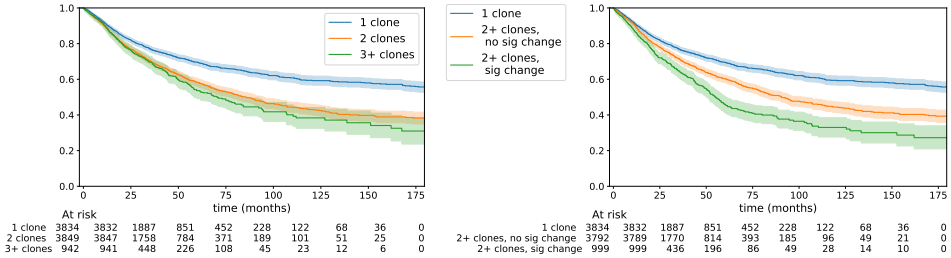


Figure S59: Kaplan-Meier curves for all TCGA samples (8625) distinguishing tumors only along the number of clones (left) or along the number of clones and the presence of a significant change in signatures along tumor evolution (right) using the public input mutation sets. A multivariate Cox model was fitted in both cases, and indicates for 2 clones, hazard ratio (HR) of 1.38 (95% confidence interval (CI): $[1.27, 1.49]$, $p = 2.62\text{e} - 15$), and 3 clones (HR= 1.54, CI= $[1.36, 1.74]$, $p = 4.53\text{e} - 12$) (left). For the distinction based on signature change, without signature change (HR= 1.32, CI= $[1.22, 1.43]$, $p = 7.55\text{e} - 12$), and with signature change (HR= 1.80, CI= $[1.60, 2.02]$, $p = 4.89\text{e} - 23$) (right)

Supplementary tables

Cancer type	Mean nb mutations (protected)	Mean nb mutations (public)	Standard deviation nb mutations (protected)	Standard deviation nb mutations (public)	Number of samples	Median followup (months)	Number of events
ACC	324.34	110.58	467.43	240.88	77	39.22	27
BLCA	732.18	350.38	886.65	424.23	354	17.21	153
BRCA	472.91	121.88	981.68	372.65	931	27.00	123
CESC	921.07	366.51	2869.13	1323.24	275	21.42	67
CHOL	358.97	100.97	505.33	225.34	35	12.65	15
COADREAD	2085.67	621.14	6247.55	1708.77	458	21.42	93
DLBC	568.30	203.68	276.72	123.15	37	24.67	5
ESCA	707.39	247.19	560.34	251.02	180	13.02	75
GBM	790.05	245.66	2583.80	1191.70	327	11.27	246
HNSC	454.21	201.73	543.69	271.16	445	20.96	185
KICH	209.32	50.12	264.68	142.42	60	85.66	8
KIRC	330.32	73.08	338.69	52.35	271	36.33	64
KIRP	280.86	82.57	130.50	37.90	242	25.13	37
LGG	212.73	76.89	1462.32	770.54	455	20.04	115
LIHC	511.53	157.00	445.93	174.27	347	19.25	117
LUAD	892.89	381.00	985.35	393.91	433	22.01	146
LUSC	909.66	382.83	686.09	289.58	423	22.27	169
MESO	203.23	47.08	114.19	48.74	78	NA	0
OV	593.79	160.15	562.30	152.24	390	31.34	227
PAAD	560.23	203.05	3780.28	1828.93	150	15.14	83
PCPG	78.70	14.09	16.71	7.43	141	NA	0
PRAD	171.30	61.78	824.76	467.50	458	30.80	8
SARC	424.27	130.72	723.66	309.04	210	30.96	81
SKCM	1876.97	886.84	2621.06	1204.90	423	35.28	184
STAD	989.96	455.07	1822.53	909.56	418	14.01	163
TGCT	148.88	22.16	33.90	11.64	128	43.05	3
THCA	120.98	16.22	95.88	14.59	467	31.01	12
THYM	253.69	36.40	175.18	83.19	121	39.17	8
UCEC	4647.22	1791.01	12341.73	4832.10	487	30.12	80
UCS	680.43	198.00	1743.77	738.28	53	NA	0
UVM	88.54	24.60	96.34	62.57	80	27.52	11

Table S3: Characteristics of the TCGA cohort used in this study.

tcga_name	ACC	BLCA	BRCA	CESC	CHOL	COADREAD	DLBC	ESCA	GBM	HNSC	KICH	KIRC	KIRP	LGG	LIHC	LUAD	LUSC	MESO	OV	PAAD	PCPG	PRAD	SARC	SKCM	STAD	TGCT	THCA	THYM	UCEC	UCS	UVM
SBS1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SBS2	1	1	1	1	1	1	1	1	0	1	1	1[21]	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1
SBS3	1	1	1	0	0	0	0	1	0	1[46]	0	1[21]	0	0	1	0	0	0	1	1 [2]	1	1 [2]	1	0	1[45]	1	0	0	1 [2]	0	1
SBS4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1[40]	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0
SBS5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
SBS6	0	0	1[39]	1	1[41]	1[20]	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1[26]	0	1[43]
SBS7a	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	1	1	1	0	0	1
SBS7b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0
SBS7c	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0
SBS7d	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	1	0	0	0	1	0	0	1
SBS8	0	0	1[39]	0	1[41]	0	0	0	1[2]	0	1	0	0	0	0	0	0	0	1[2]	0	0	1[19]	1	0	0	0	0	1	0	0	0
SBS9	0	0	0	0	0	0	1[42]	0	0	0	0	1[21]	0	0	1	0	0	0	0	0	0	1[19]	0	0	0	0	0	0	0	0	0
SBS10a	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	1	1	0	0
SBS10b	1	0	1	1	0	0	0	0	0	0	1	0	0	1	0	0	0	1	0	1	1	1	0	0	0	1	1	1	1	1	0
SBS11	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS12	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1[40]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[43]
SBS13	1	1	1	1	1	1	1	1	0	1	1[2]	1[21]	1	1	0	1	1	1	1	1	1	0	1	0	1	1	1	0	1	1	1
SBS14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SBS15	0	0	0	0	0	1	0	0	0	0	1	1	0	1	0	0	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1
SBS16	0	0	0	0	1[41]	0	0	1	0	0	0	0	0	0	1	0	0	0	1[44]	0	1	1[19]	0	0	0	0	1	1	0	0	1[43]
SBS17a	0	0	1[39]	0	1[41]	1	1	1	0	0	1	1[21]	0	0	1[40]	0	0	0	0	1 [2]	1	0	0	0	1	1	0	1	0	0	0
SBS17b	0	0	1[39]	0	1[41]	1	1	1	0	0	0	1[21]	0	0	1[40]	0	0	0	0	1 [2]	0	0	0	0	1	0	0	0	0	0	0
SBS18	0	0	1[2]	0	0	1[2]	1[42]	1	0	1[2]	0	0	0	0	1[2]	0	1[2]	0	0	1 [2]	0	1 [2]	1	0	1[2]	0	1 [2]	1	1 [2]	0	0
SBS19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0
SBS20	0	0	1[39]	0	1[41]	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
SBS21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS22	0	0	0	0	1[41]	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SBS23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[40]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
SBS25	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS26	0	0	1[39]	0	0	1[20]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SBS29	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SBS30	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
SBS32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SBS33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1 [2]	0	0	0	0	0	0	0	0	0
SBS35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1[2]	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SBS36	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1 [2]	0	0	0	0	0	0	0	0	0	0	0
SBS37	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1 [2]	0	0	0	0	0	0	0	0	0
SBS38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0
SBS39	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1[2]	0	0	0	0	0	0	1	0	0	0	0	0
SBS40	0	0	1[2]	0	1	1	1	0	1	0	1	1	1	0	1	1	0	1	1	0	0	1	1	0	1	1	0	1	1	0	0
SBS41	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS42	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	0	1
SBS43	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS44	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0
SBS45	0	1	0	0	0	0	0	0	1	1	0	1	1	1	0	1	1	0	0	0	0	1	0	1	0	0	1	0	0	0	0
SBS46	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS47	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SBS48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS49	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0	0
SBS50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS51	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS52	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
SBS53	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS54	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
SBS55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS56	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SBS57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SBS58	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0
SBS59	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	1
SBS60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table S4: Table of presence of signatures in the different cancer types. A 1 indicates the presence of the signature, and a 0 an absence. The background color indicates whether the presence of the signature comes from SigProfiler detection in the TCGA (lavender) or from the literature (green), in which case, the matching paper is referenced in the table cell.