

Chapter 4

Mutational process signatures: an application for the Hierarchical Dirichlet Process

Somatic genome alterations stem from a variety of underlying processes, including mutagen exposure, replication error, and defective repair. As each different process generates a characteristic distribution or ‘signature’ of alteration classes, somatic mutation catalogues serve as useful records of historic and ongoing mutagenic activity. To decipher the constituent signatures, observed mutations in a sample cohort are fractionated by their co-occurrence patterns (Nik-Zainal et al., 2012). For SNVs, a subset of about twenty derived signatures have a proposed aetiology as genuine mutational processes or known sequencing artefacts (Alexandrov et al., 2013b; Helleday et al., 2014), often confirmed through experimental data (Segovia et al., 2015; Drost et al., 2017). Cancer sample characterisation by signature exposure has important applications such as: quantifying the effect of environmental carcinogens like aristolochic acid (Poon et al., 2013; Poon et al., 2015), and revealing druggable opportunities like HR-deficiency (with or without *BRCA* loss) (Alexandrov et al., 2015a; Davies et al., 2017; Polak et al., 2017).

In this chapter, I briefly review published methods for mutational signature decomposition (Section 4.1), and describe a different statistical approach using the hierarchical Dirichlet process (Section 4.2). I illustrate the performance of this HDP method on simulated (Section 4.3) and real (Section 4.4) SNV catalogues, and then examine its ability to match new data to an existing signature library while *simultaneously* discovering novel signatures (Section 4.5).

Finally, I return to the PCAWG SV dataset and find fifteen novel rearrangement signatures defined by SV class, size, and replication timing (Section 4.6).

4.1 Existing methods for mutational signature analysis

In the first formal analysis of this kind, Nik-Zainal et al. (2012) used non-negative matrix factorization (NMF) to find five underlying signatures in 21 breast cancer genomes. Alexandrov et al. (2013a) further expounded the details of this NMF application, developing the most widespread signature analysis framework to date. This NMF method assumes each signature is a discrete probability distribution over a finite set of unordered mutation classes. For the $p \times n$ count matrix M which tallies p mutation classes in a cohort of n samples, standard NMF algorithms approximate $M \approx S \times E$, where S is the $p \times k$ matrix of k signatures (constrained to have non-negative columns sum to one), and E is the $k \times n$ sample exposure matrix recording the non-negative burden of each signature in each sample. That is, the sample *exposure* to a signature is the estimated number of mutations generated by that signature in that particular sample. Most SNV studies define mutation classes by the trinucleotide context, yielding $p = 96^a$.

To determine the number of signatures (k unknown), Alexandrov et al. (2013a) calculate NMF solutions at a range of plausible k values for a series of bootstrap-resampled M matrices, returning the consensus solution with stability across bootstraps and minimal reconstruction error by the Frobenius norm. NMF was most notably used to extract 21 validated SNV signatures from 7000 cancer samples analysed by Alexandrov et al. (2013b). However, NMF is not a formal statistical model with probabilistic interpretation, and the choice of Frobenius norm does not account for the integer nature of the input matrix.

Three alternative methods—EMu (Fischer et al., 2013), signeR (Rosales et al., 2016), and SignatureAnalyzer (Kim et al., 2016)—make similar assumptions as the NMF approach, namely that mutations derive from sample-specific mixtures of shared underlying signatures, where each signature is a discrete probability distribution over unordered mutation classes. Crucially, all three methods assume the observed mutation counts follow a Poisson distribution. Their

^aThere are six possible single base substitutions (reported from the pyrimidine side), with four possible flanking bases either side, so $4 \times 6 \times 4 = 96$ SNV classes.

major distinctions lie in the method of estimation and signature number choice. EMu uses expectation-maximisation (EM) iterations to fit signatures and sample exposures until convergence, whereas *signeR* models the signature probabilities and sample exposures with gamma priors in a Bayesian framework. Both EMu and *signeR* use the Bayesian information criterion to select a model with high likelihood and few parameters (penalising too many signatures). *SignatureAnalyzer* aims to minimise the Kullback-Leibler divergence between the NMF solution and input matrix (rather than Frobenius norm), which is equivalent to maximising the Poisson likelihood. To select the number of signatures, *SignatureAnalyzer* adopts a Bayesian shrinkage methodology (Tan and Févotte, 2013) which automatically determines the relevant components by driving some signature weights to a small lower bound (effectively zero).

Taking a different approach, the ‘*pmsignature*’ method (Shiraishi et al., 2015) does *not* consider a mutational signature to be one discrete probability distribution over a large set of mutation classes. Instead of assigning each mutation to just one classification, Shiraishi et al. (2015) model each mutation as having a *set* of observed categorical variables such as substitution type, flanking base identity, and transcriptional strand (in genic regions). In this paradigm, signatures are defined by a collection of distributions over each separate variable, under the simplifying assumption of independence between all mutation features. With this strategy, many relevant features beyond simple trinucleotide context are included within a relatively small parameter space. *Pmsignature* uses EM iterations to calculate all signature and sample exposure parameters, and selects the number of signatures that yields high likelihood without splitting into multiple components with very similar distributions.

In this chapter, I propose that the hierarchical Dirichlet process (HDP) (Teh et al., 2006) is well-suited to the problem of mutational signature decomposition, particularly in the context of multiple sample groups and/or prior signature information. The signatures defined by the HDP model match the ‘traditional’ paradigm of one discrete probability distribution per signature, as previously established by NMF and most other methods (with the notable exception of ‘*pmsignature*’). With HDP, a flexible hierarchical model borrows information across samples and groups to identify shared signatures, while also quantifying differences between samples and groups. Under the nonparametric Bayes assumption of infinitely many generating processes, HDP automatically determines the underlying signature number, and can discover novel patterns while simultaneously matching against a prior library of known signatures.

4.2 HDP method for mutational signatures

Teh et al. (2006) first developed the hierarchical Dirichlet process mixture model for the problem of topic modelling in corpora (collections of written text; concept reviewed by Blei (2012)). The HDP is a non-parametric Bayesian clustering method, and infers the number of clusters directly from the complexity of the dataset by assuming the data is drawn from some finite subset of infinitely many generative processes. In Appendix B, I describe the HDP for the novel use case of signature patterns within somatic mutation catalogues.

4.2.1 HDP overview

An overview of the HDP model is shown in Figure 4.1, as designed for multiple groups of samples. Other designs with different hierarchical levels are also possible; for example, an additional child node layer could capture multiple samples from the same individual.

In brief, mutations observed in each sample are tallied into discrete, unordered categories. I assume these mutation counts are randomly drawn from a sample-specific mixture of an infinite number of multinomial distributions (the signatures) over the set of possible mutation classes. Under the HDP model, the sample-specific signature distribution is a Dirichlet process (DP) draw from the group-specific signature distribution. A DP can be intuitively understood as taking in one probability density function, and outputting a sparser, more discretised probability function defined on the same domain^b. That is, the signature distribution in a sample is based on the parent distribution of its group, but with the probability density further concentrated at particular values/signatures. Moving up one hierarchical level, the group-specific signature distribution is itself a DP draw from the overall distribution of signatures in the dataset. At the top level, the dataset-specific signature distribution is a DP draw from the uniform probability over the infinite set of all possible signatures.

In practice, we observe the mutation catalogues at the bottom of the tree, and specify the uniform Dirichlet prior at the top of the tree, but must estimate the signatures (their identity and prevalence) at each node in-between. To

^bTo use the stick-breaking analogy, a DP draw is built from an infinite random sample from the input distribution, weighted by an infinite series of successive weights randomly broken off an imaginary ‘stick’ of unit length. A concentration parameter controlling the proportion of ‘stick’ broken off each time (rate at which the weights attenuate) controls the degree of sparsity in the output.

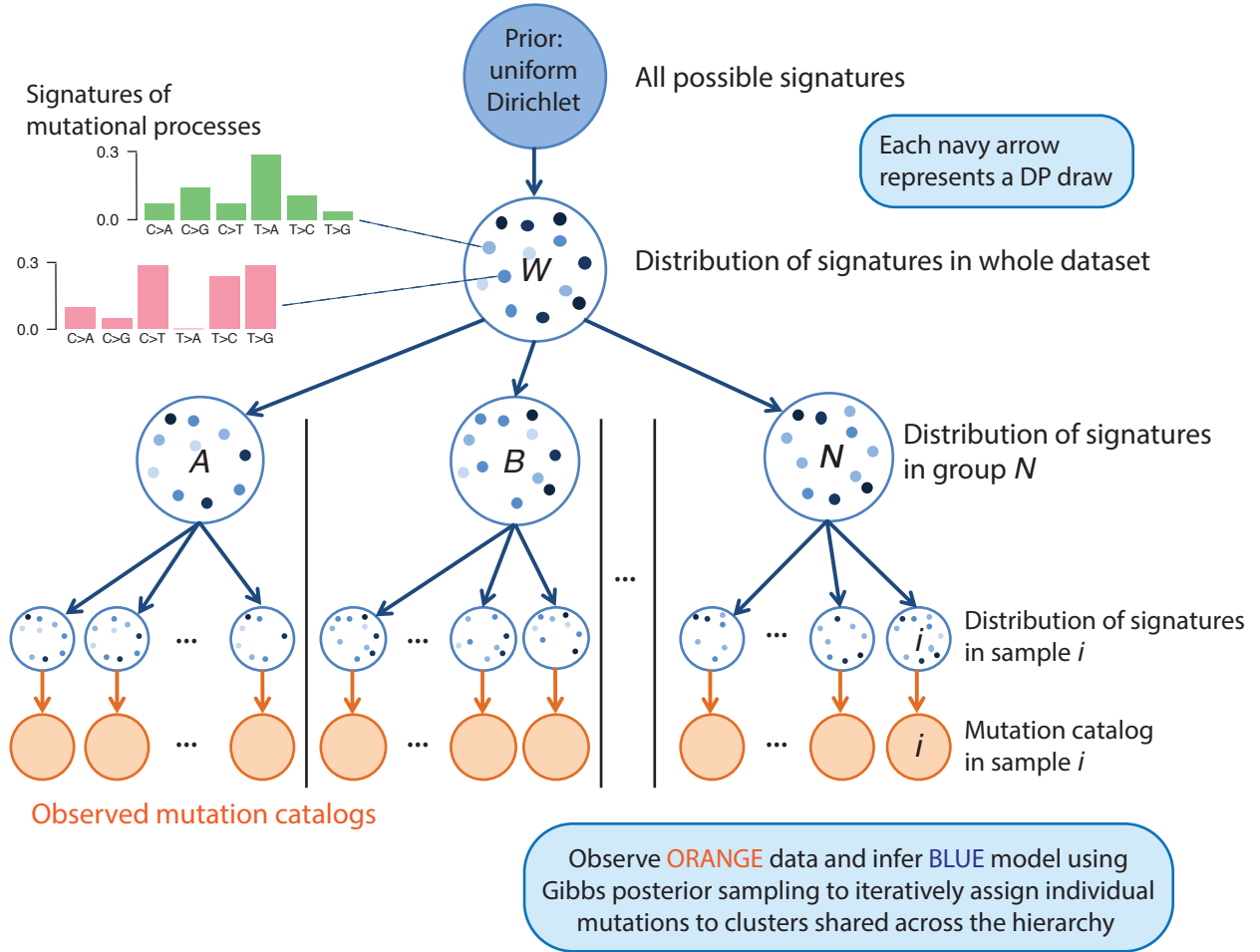


Figure 4.1: Schematic overview of the HDP model for multiple sample groups. Each blue node represents a distribution over the infinite set of all possible signatures, and is a Dirichlet process draw from its parent node. At the top of the tree, the prior distribution is uniform over all possible signatures. Each successive child node concentrates the probability density at particular signature values. The small blue dots inside each node represent particular signatures (discrete probability distributions over the mutation classes), with different shades representing the probability of that signature in the node. The two example signatures over six mutation classes are illustrative only; in practice, mutations are classified into more specific groups (e.g. 96 SNV classes in a trinucleotide context). At the bottom of the tree, the observed data are per-sample mutation catalogues (tallies of mutation classes), assumed to be drawn from sample-specific multinomial mixtures.

perform this posterior inference with the Gibbs sampling method by Teh et al. (2006), all observed mutations are initialised with a random cluster allocation (clusters of mutations define the estimated signatures, and are shared across nodes/samples). Then, the Gibbs procedure cycles through each mutation in turn and assigns an updated cluster allocation, most likely moving to a cluster with: (a) a high proportion of that same mutation class (across all samples); and/or (b) a high proportion of mutations in that sample and/or parent DP (across all mutation classes). At any iteration, there is also a small chance (controlled by the concentration parameter) that a mutation gets assigned to a brand new cluster by itself. In this way, the number of clusters fluctuates throughout the MCMC sampling chain, and there is no need to specify how many clusters should be found. Concentration parameters for each DP are sampled from a Gamma hyper-prior as one of the Gibbs iterations. More details are available in Appendix B. After a burn-in period, posterior samples taken at regular intervals provide a snapshot of possible cluster allocations that defines the space of probable signatures and their prevalence at each node.

Originally, Teh et al. (2006) implemented this Gibbs scheme for the HDP as a suite of functions written in MATLAB and C^c. To encourage the adoption of HDP in the bioinformatics community, I developed the open-source R package `hdp` as a practical front-end to the original C engine for MCMC inference (R Core Team, 2017; Roberts, 2015). In addition to providing a user-friendly package with detailed documentation and examples, I also developed a suite of post-processing functions for practical reporting across MCMC chains, and a convenient method for setting up pseudo-counts in frozen nodes to condition on prior knowledge. Although this work was motivated by mutational signatures analysis, the utility of my `hdp` package extends to any similar problem involving categorical count data, and was used by Papaemmanuil et al. (2016) to cluster co-occurring driver alterations in acute myeloid leukaemia. Given the range of applications, the package documentation refers to the generic nomenclature of ‘components’ rather than mutational signatures.

4.2.2 Extracting consensus signatures

Each posterior sample collected off an MCMC chain consists of per-mutation cluster allocations. This output is not immediately amenable to direct reporting because:

^c<http://www.stats.ox.ac.uk/~teh/research/npbayes/npbayes-r21.tgz> available as of December 2017

- the number of raw clusters varies across posterior samples;
- many raw clusters are very small, with only a few mutations assigned (because HDP assumes infinitely many underlying signatures); and
- multiple clusters can have the same data distribution (strong signatures sometimes found multiple times).

To extract meaningful output from a collection of posterior samples (ideally from multiple independent MCMC chains), I developed a new post-processing method^d to hone in on the stable set of consistently returned clusters while consolidating the smaller, transitory raw clusters into an additional component capturing noise and uncertainty. While useful for accessible interpretation, this method loses the variable posterior distribution over the number of signatures. However, as the number of raw clusters in a non-parametric Bayesian model is known to scale logarithmically with the number of observed data items (Teh and Jordan, 2009), I conjecture that the raw clusters do not provide the best biological insight by themselves, and instead propose the following approach.

Let S be the number of posterior samples collected, and $K^{[s]}$ the number of raw clusters in posterior sample s for $s \in 1, \dots, S$. Each posterior sample s assigns each individual mutation to a raw cluster $k^{[s]} \in 1, 2, \dots, K^{[s]}$. Let the maximum number of raw clusters be denoted K^m . For p mutation classes, let $\mathbf{r}_k^{[s]}$ be a p -length count vector of mutations assigned to raw cluster k in posterior sample s , and $\mathcal{R}^{[s]}$ denote the $p \times K^{[s]}$ count matrix of mutation classes in all raw clusters from that posterior sample.

My method for extracting consensus signatures is as follows.

1. Append $K^m - K^{[s]}$ zero vectors to each $\mathcal{R}^{[s]}$ so all count matrices have dimension $p \times K^m$. That is, $\mathcal{R}^{[s]'} = \begin{bmatrix} \mathcal{R}^{[s]} & \mathbf{0}_{p \times (K^m - K^{[s]})} \end{bmatrix}$.
2. Match up raw clusters across posterior samples by K^m -centroid clustering of all $\mathbf{r}_k^{[s]'}$, minimising the Manhattan distance to the median and imposing a cannot-link constraint on raw clusters from the same posterior sample. This enforces a result of K^m super-clusters (components), each with S members all from different posterior samples. Let $\mathcal{C}_\ell = \begin{bmatrix} \mathbf{r}_\ell^{[1]'}, \mathbf{r}_\ell^{[2]'}, \dots, \mathbf{r}_\ell^{[S]'} \end{bmatrix}$ be the $p \times S$ matrix of all raw clusters (and possibly some zero vectors) assigned to component ℓ for $\ell \in 1, \dots, K^m$.
3. Merge components with very similar mutation class distributions. Let

^dAvailable in the `hdp_extract_components` function within the `hdp` package.

the average mutation class distribution for \mathcal{C}_ℓ be

$$\bar{\mathbf{c}}_\ell = \left[\sum_{s=1}^S \left(\mathbf{r}_\ell^{[s]'} / \left\| \mathbf{r}_\ell^{[s]'} \right\|_1 \right) \right] / S.$$

If cosine similarity($\bar{\mathbf{c}}_a, \bar{\mathbf{c}}_b$) ≥ 0.9 , then $\mathcal{C}_{\text{new}} = \mathcal{C}_a + \mathcal{C}_b$ ^e.

4. Assign components with no significantly non-zero mutation classes to ‘component zero’ to capture the fraction of noise/uncertainty. Let $\text{HPD}_{0.95}(\mathbf{y})$ return the highest posterior density interval containing 95% of \mathbf{y} values, so an indicator for the absence of significant mutation classes is

$$z_\ell = \begin{cases} 1 & \text{if } 0 \in \text{HPD}_{0.95}(\mathcal{C}_{\ell,i,:}) \text{ for all rows } i = 1, \dots, p, \\ 0 & \text{otherwise.} \end{cases}$$

Initialise the zero component as

$$\mathcal{C}_{\text{zero.init}} = \sum_{\{\ell | z_\ell = 1\}} \mathcal{C}_\ell,$$

removing non-significant components (with $z_\ell = 1$) from the main set.

5. Assign components with no significantly non-zero sample exposures to ‘component zero’. Where previously we have pooled samples and looked at the distribution across mutation classes (p rows), now pool mutation classes and consider the distribution across samples. For n samples (leaf nodes), let \mathcal{C}_ℓ^* be the $n \times S$ count matrix of mutations assigned to component ℓ for each sample (row) in each posterior sample (column). An indicator for the absence of significant sample exposures is

$$z_\ell^* = \begin{cases} 1 & \text{if } 0 \in \text{HPD}_{0.95}(\mathcal{C}_{\ell,i,:}^*) \text{ for all rows } i = 1, \dots, n, \\ 0 & \text{otherwise.} \end{cases}$$

Add to the zero component, such that

$$\mathcal{C}_{\text{zero}} = \mathcal{C}_{\text{zero.init}} + \sum_{\{\ell | z_\ell^* = 1\}} \mathcal{C}_\ell,$$

removing non-significant components (with $z_\ell^* = 1$) from the main set^f.

^e0.9 is the default similarity threshold for merging components, but can be changed.

^fAn optional variation is to require non-zero sample exposure in two (or more) samples, changing the z_ℓ^* indicator to one if all but one (or more) rows have credibility intervals including zero.

6. Finally, the remaining components are ranked by their prevalence (total number of mutations assigned, averaged over posterior samples) and reported as the set of consensus signatures.

This method returns a set of robust signatures with significant exposure in at least one sample and significant presence of at least one mutation class. The number of signatures is empirically determined without resorting to separate model fitting for every plausible number. A fraction of mutations are assigned to component zero, and reflect the extent of noise and uncertainty in the signature estimation method. Credibility intervals for the mutation classes in each signature, and for the level of signature exposure in each sample and group, are simply constructed as highest posterior density intervals from the set of posterior samples.

4.2.3 Conditioning on prior knowledge

Given the availability of known SNV signatures extracted from large datasets (Alexandrov et al., 2013b; Alexandrov et al., 2015b), it will often be desirable to match a new mutation catalogue to existing signatures, rather than performing *de novo* signature discovery every time. Conditioning on prior knowledge not only saves computational time and effort, but also improves accuracy for small datasets, and leverages existing signature aetiology explanations.

Matching a new dataset to an existing library of mutational signatures is already possible with several methods. For example, with NMF, any mutation tally matrix can be factored into a fixed matrix of known signatures and an unknown sample exposure matrix to be estimated. Alternatively, the ‘deconstructSigs’ R package by Rosenthal et al. (2016) matches new mutation data to existing signatures with brute-force iterations to minimise the reconstruction error. However, both these approaches are restricted to the pre-defined signature set, and will find a poor solution if the new dataset contains previously unreported signatures, either from cohort-specific mutational mechanisms or a specific profile of artefactual variant calls. Artefacts vary with DNA library preparation, sequencing platform, and bioinformatics calling pipelines, so may appear as novel signatures even in well-studied cancer types.

With its non-parametric Bayes assumption of infinitely many generating signatures, the HDP model is uniquely suited to address this problem, and can *simultaneously* match data to known signatures *and* allow for potential discovery of new signatures.

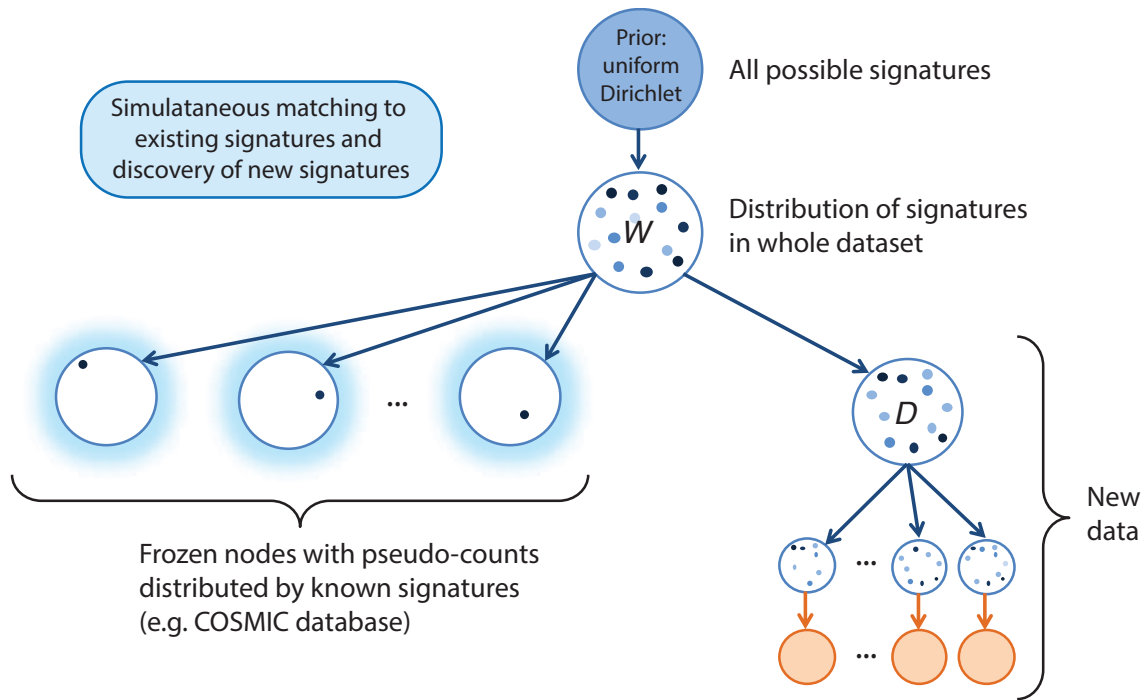


Figure 4.2: Overview of the HDP model conditioning on prior knowledge about known mutational signatures. For each known signature, a number of pseudo-counts following the expected mutation class distribution are assigned to a frozen node and allocated to one fixed cluster. The ‘frozen’ node status means the cluster allocation of these pseudo-counts is fixed throughout the MCMC posterior sampling. This forces their parent node describing the distribution of signatures in the whole dataset to always apportion some probability to these known signatures. The other nodes behave as in Figure 4.1, and observed mutations in the new dataset are free to cluster either with the fixed pseudo-counts of prior signatures, or in separate clusters describing novel signatures.

The diagram in Figure 4.2 overviews the pseudo-count strategy for conditioning on prior knowledge. When initialising the HDP structure describing the generative model for a new dataset, each known signature is assigned to a ‘frozen’ child node (DP draw) as shown in Figure 4.2. For each prior signature, the characteristic distribution over mutation classes is instantiated as a set of pseudo-counts fixed to one cluster throughout the posterior sampling process. While the pseudo-counts are held frozen in their cluster allocation, the mutations observed in the new dataset are free to cluster with either the fixed pseudo-counts of a prior signature, or in novel clusters solely composed of new data observations. Following the collection of posterior samples and extraction of consensus signatures (Section 4.2.2), the signatures are labelled by their match in the prior set or with a new label for novel discoveries.

4.3 HDP performance on simulated data

4.3.1 Simulated mutation catalogues

To assess the performance of the HDP method for mutational signatures analysis, I generated a collection of simulated SNV mutation catalogues and compared the HDP reconstruction with the known underlying signatures and sample exposures under a range of conditions.

In total, I simulated 240 separate datasets (parameters in Table 4.1) by varying the number of samples, number of underlying signatures, similarity of sample exposure patterns, number of distinct sample sub-groups, and whether or not the total mutational burdens are consistent with WES or WGS data.

Table 4.1: Parameter combinations for simulated SNV catalogues. Five independent datasets were simulated with every possible combination of parameters within each column, generating 240 simulated datasets in total.

	base combinations	different exposure	three sub-groups
samples	50, 100, 200	50, 100	50,100
generating signatures	5, 10, 15, 20	5, 10	5, 10
seq tech / burden	WES, WGS	WES, WGS	WES, WGS
exposure similarity	medium	low, high	medium
number of groups	1	1	3
replicates	5	5	5
total datasets	120	80	40

Each simulated dataset randomly sampled K underlying signatures from a set of 30 published by the COSMIC database (v74, Forbes et al. (2015)^g) after NMF analysis of 10,952 exomes and 1,048 whole genomes (Alexandrov et al., 2013b; Alexandrov et al., 2015b). Each COSMIC signature θ_k is a discrete probability distribution over 96 mutation classes (SNVs in trinucleotide context).

The number of mutations in sample j was taken to be $n_j = \min(\lfloor 10^{x_j} \rfloor, 20000)$ for $X \sim \text{Gamma}(\alpha, \beta)$, with shape and rate parameters specific to either WES or WGS data^h.

Next, the signature exposure vector ϕ_j for sample j (probability distribution over the set of K signatures) was sampled from $\phi \sim \text{Dirichlet}_K(\tau \times \eta)$, where

^g<http://cancer.sanger.ac.uk/cosmic/signatures> available as of December 2017

^hBy fitting gamma distributions to the \log_{10} -transformed per-sample mutation counts in exome and genome datasets described in Section 4.4, I obtain shape $\alpha_E = 8.23$ and rate $\beta_E = 4.55$ for WES data, and $\alpha_G = 10.02$ and $\beta_G = 3.15$ for WGS data.

the concentration parameters $\boldsymbol{\tau} \sim \text{Dirichlet}_K(\mathbf{1})$ were newly sampled for each simulated cohort, and the exposure similarity weight η was set to 10 for ‘medium’ similarity across samples, 1 for ‘low’, and 20 for ‘high’. In simulations with three sample sub-groups ($g \in [1, 2, 3]$), the sample exposures were drawn from group-specific distributions with $\boldsymbol{\tau}_g \sim \text{Dirichlet}_K(\mathbf{1} \times 0.5)$. As an additional constraint on the exposure profile, each of the K generating signatures was forced to contribute at least 2% (for $K \in 5, 10$) or 1% (for $K \in 15, 20$) of the total mutations in the cohort.

Finally, the n_j mutations in sample j were drawn from a sample-specific distribution over the 96 mutation classes as defined by $[\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K] \times \boldsymbol{\phi}_j$ (signatures mixed by sample-specific exposure proportions).

4.3.2 Posterior inference settings

For the 240 separate simulated datasets outlined in Table 4.1, I attempted to reconstruct the generating signatures and sample exposures using the HDP model with a variety of settings. Unless otherwise specified, the default HDP design is for one shared concentration parameter across all nodes, with one top parent node modelling the dataset distribution of signatures, and one child node per sample descended from the same shared parent.

The base setting was to collect 500 posterior samples (50 iterations apart) from four independent MCMC chains after 5000 burn-in iterations (2000 posterior samples total). Under the base setting, I initialised all models with 10 clusters, and set the gamma hyper-parameters for the shared concentration parameter at shape = 1 and rate = 1. All 240 datasets were put through HDP clustering with these base settings, and some were also run with additional combinations. As the generating signatures from the COSMIC set include one pair with cosine similarity just below 0.92, I set 0.92 as the similarity threshold for component merging during signature extraction.

To assess the influence of initial clustering, 60 datasets were also run with initial cluster counts of 5 or 15, holding the other settings constantⁱ.

To assess the influence of the concentration parameter, 40 datasets were also run with a shape hyper-parameter of 0.1 or 10, holding the other settings

ⁱFor datasets with 50 or 100 samples; WES or WGS burden; 5, 10, or 15 underlying signatures; medium sample exposure similarity; and one shared group.

constant^j.

Finally, to assess the influence of specifying a sub-group structure in cases where it does and does not exist, 80 datasets were also run with a three-group node hierarchy and group-specific concentration parameters, holding the other settings constant^k.

4.3.3 HDP performance on simulated data

Metrics To assess performance of the HDP method, Figures 4.3–4.8 compare the following four metrics:

- number of signatures returned (compare against number of generating signatures, indicated by colour);
- proportion of mutations explained by the fit (proportion of mutations *not* assigned to component zero, averaged over posterior samples);
- cosine similarity of returned signatures with underlying signatures; and
- cosine similarity of estimated sample exposures with the true underlying exposure vectors.

Above each plot is a p -value for the independent variable in question (either a posterior sampling setting, or a property of the simulated dataset) and its relation to the performance metric, controlling for all other variables with a Poisson regression for number of signatures returned, or a beta regression for the other three metrics (defined on a 0–1 scale).

Posterior sampling settings Overall, HDP solutions are robust to the posterior sampling settings, and do not change significantly as the number of initial clusters varies from five to fifteen (Figure 4.4), nor as the mean of the hyper-prior for the concentration parameter varies by a factor of ten (Figure 4.5). Increasing the number of independent MCMC chains from two to eight has little impact (Figure 4.3), indicating that the sampling procedure is mixing around the posterior distribution in a reasonably representative manner even within one chain.

^jFor datasets with 50 or 100 samples; WES or WGS burden; 5 or 10 underlying signatures; medium exposure similarity; and one shared group.

^kFor datasets with 50 or 100 samples; WES or WGS burden; 5 or 10 underlying signatures; medium exposure similarity; and one or three shared groups.

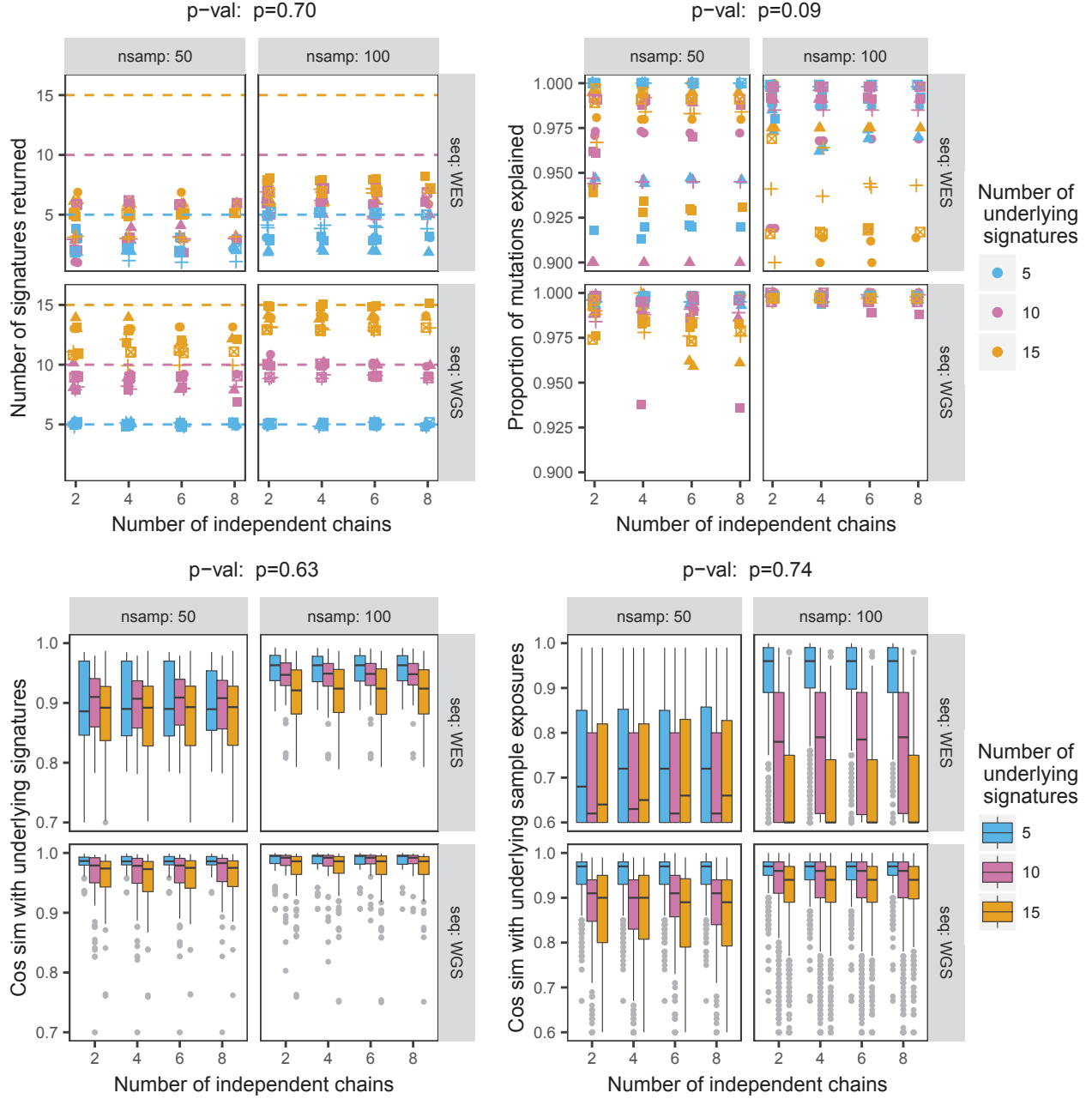


Figure 4.3: HDP performance as the number of independent MCMC chains varies. P -values above each plot are for the number of chains as a quantitative predictor of each vertical axis metric, controlling for number of samples, number of underlying signatures, and sequencing type.

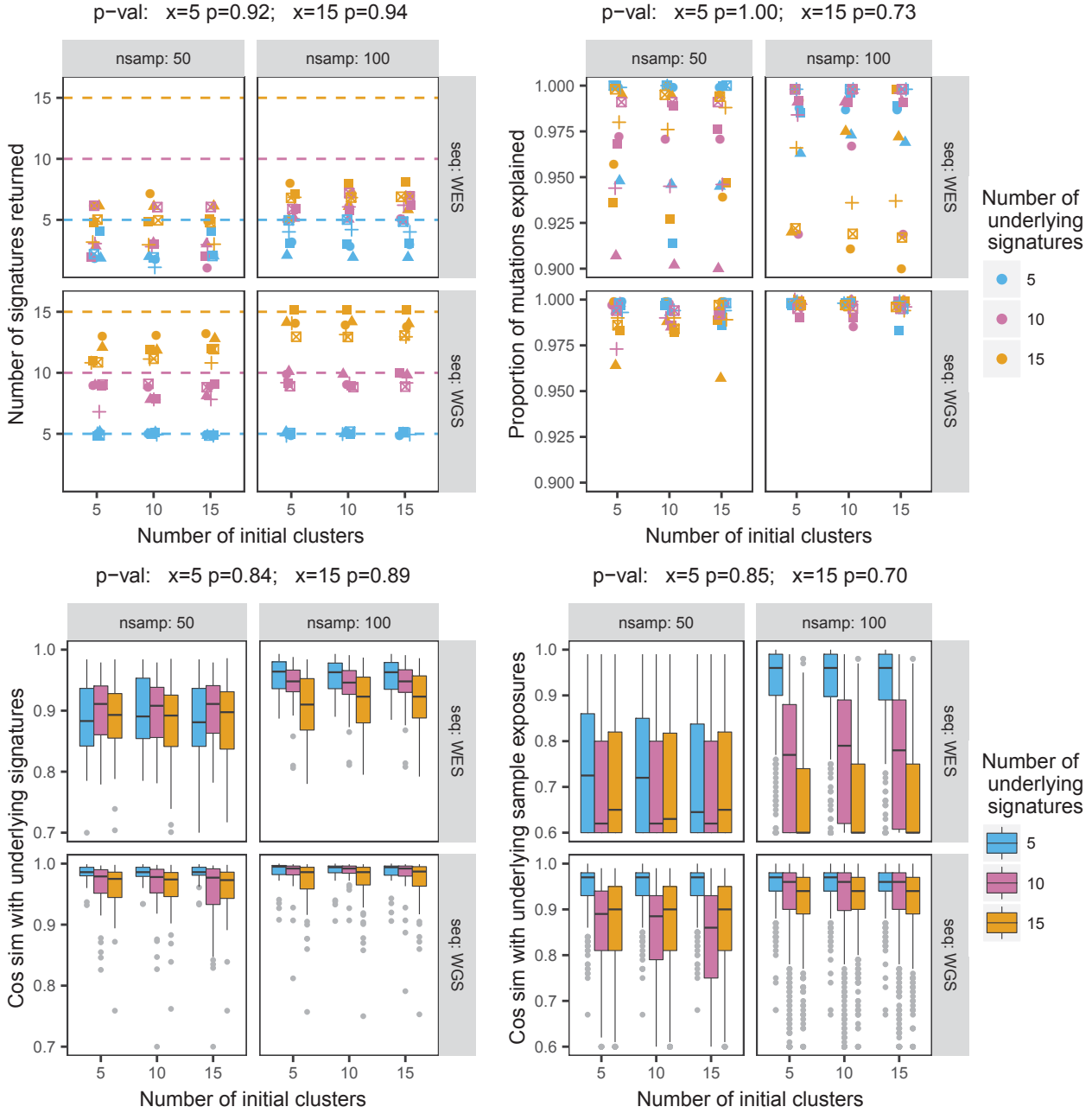


Figure 4.4: HDP performance as the number of initial clusters varies. P -values above each plot are for the number of initial clusters compared to a baseline of 10, controlling for number of samples, number of underlying signatures, and sequencing type.

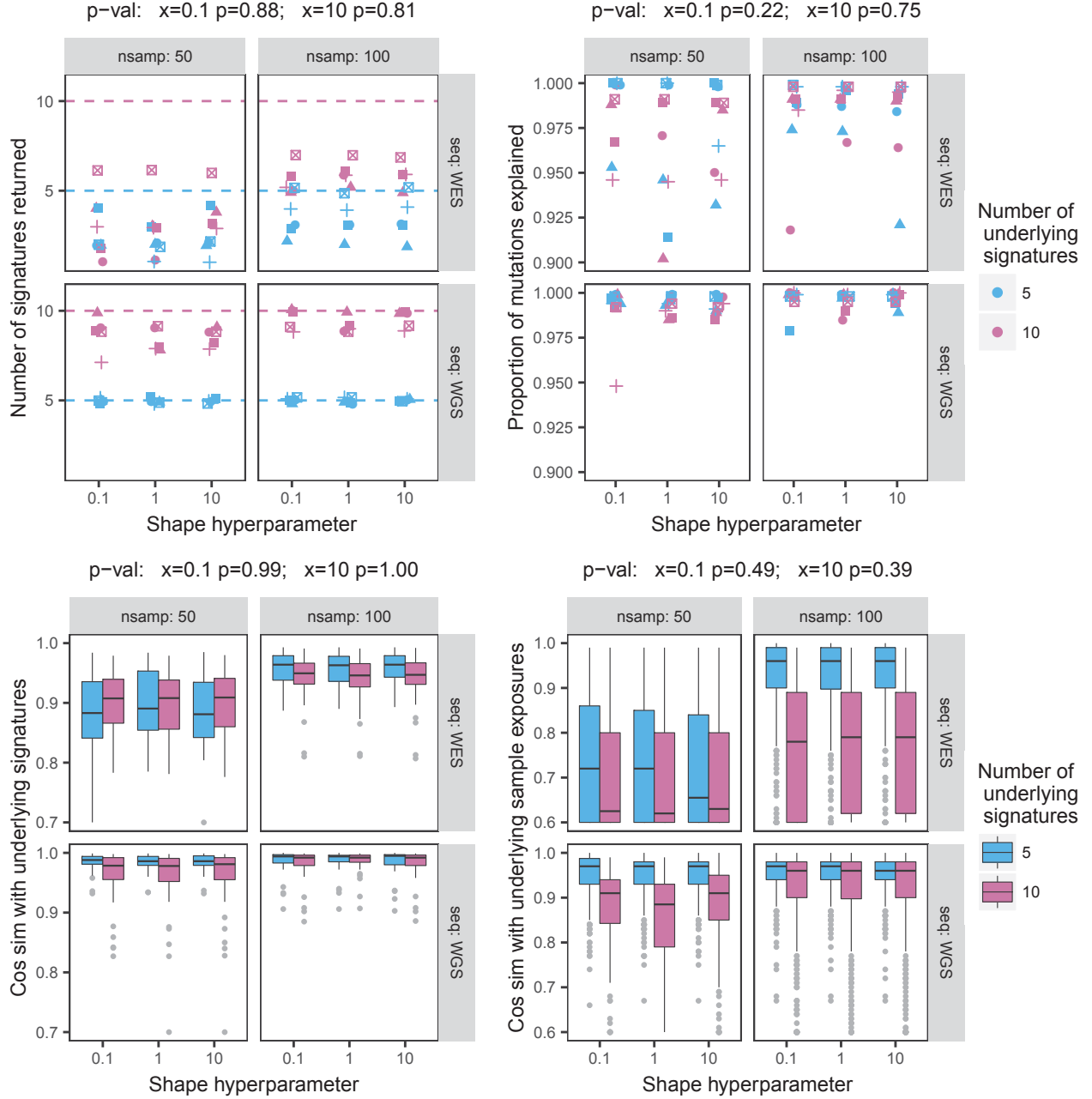


Figure 4.5: HDP performance as the shape hyper-parameter for the DP concentration parameter varies. P -values above each plot are for the shape hyperparameter levels compared to a baseline of 1, controlling for number of samples, number of underlying signatures, and sequencing type.

Dataset properties The greatest determinants of accurate signature reconstruction are dataset size and sample exposure similarity. As expected, HDP performance improves in larger datasets with more samples and more observed mutations (WGS better than WES, Figure 4.6). Under the simulation conditions established here, five mutational signatures are reliably reconstructed with about 200 exomes or 50 whole genomes. For datasets generated with 15 or 20 underlying signatures, 200 exomes will only reconstruct about 10 of these, whereas 200 whole genomes can accurately return all 15, and almost all 20 signatures. However, these guidelines are heavily dependent on the sample exposure patterns. As shown in Figure 4.7, accurate signature reconstruction requires much less data when samples have very different signature exposures, as the co-occurrence profile is more distinct for variably assorting signatures.

Sub-group structure Finally, for these simulations, the HDP results are broadly similar whether or not the samples' sub-group structure is accounted for (Figure 4.8). Although modelling the sub-group structure has no discernible influence on signature estimation, it does significantly improve the sample exposure estimates when there *is* a genuine underlying difference, and is of no detriment when the sub-group division is erroneous.

Factors influencing accuracy In all the HDP model fits on simulated data, some signatures and sample exposures are more reliably reconstructed than others. To investigate factors influencing reconstruction accuracy, I considered the subset of base setting simulations with 50 or 100 WGS samples with 5 or 10 underlying signatures and medium exposure similarity. Pooling observations across simulated cohorts, I fitted a beta regression for the outcome variable of cosine similarity between the estimation and underlying truth, with predictor variables as indicated in Figures 4.9 and 4.10. This exercise shows that sample exposure recall (Figure 4.9) is more likely to be inaccurate if the number of extracted signatures is incorrect, and/or the sample has: similar contributions from most signatures (low standard deviation); low mutation count; or a higher proportion of mutations in rare signatures. For the signatures (but not the exposures), I subset to the models returning the *correct* number of underlying signatures (eliminating poor reconstruction due to incorrect number). Signature recall (Figure 4.10) is more likely to be inaccurate if the dataset is small, or if the signature in question is: rare in the cohort (contributes a low proportion of total mutations); close to uniform across mutation classes (low standard deviation); or roughly similar to another generating signature in the cohort.

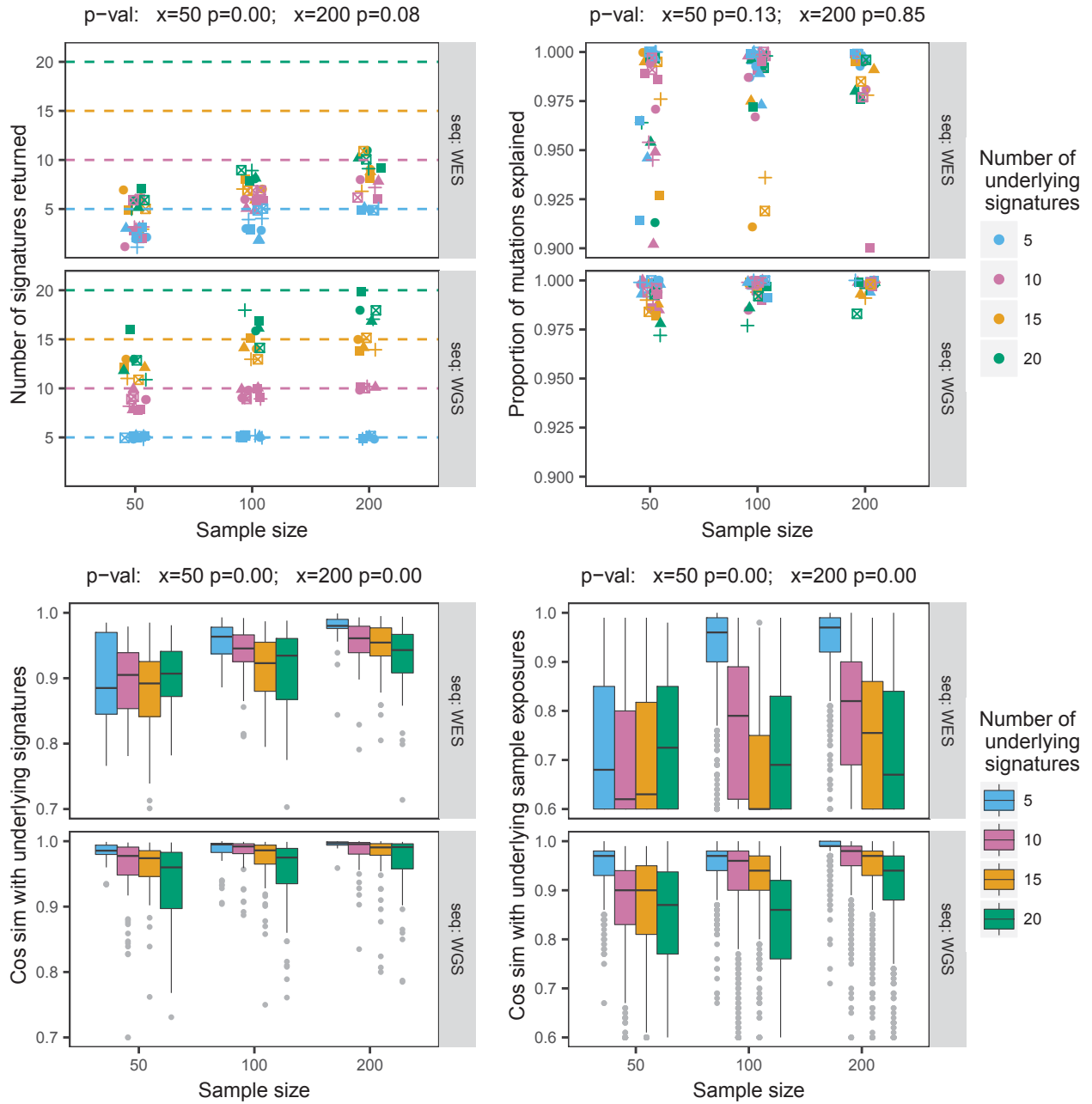


Figure 4.6: HDP performance as the number of samples varies. P -values above each plot are for the number of samples compared to a baseline of 100, controlling for number of underlying signatures, and sequencing type.

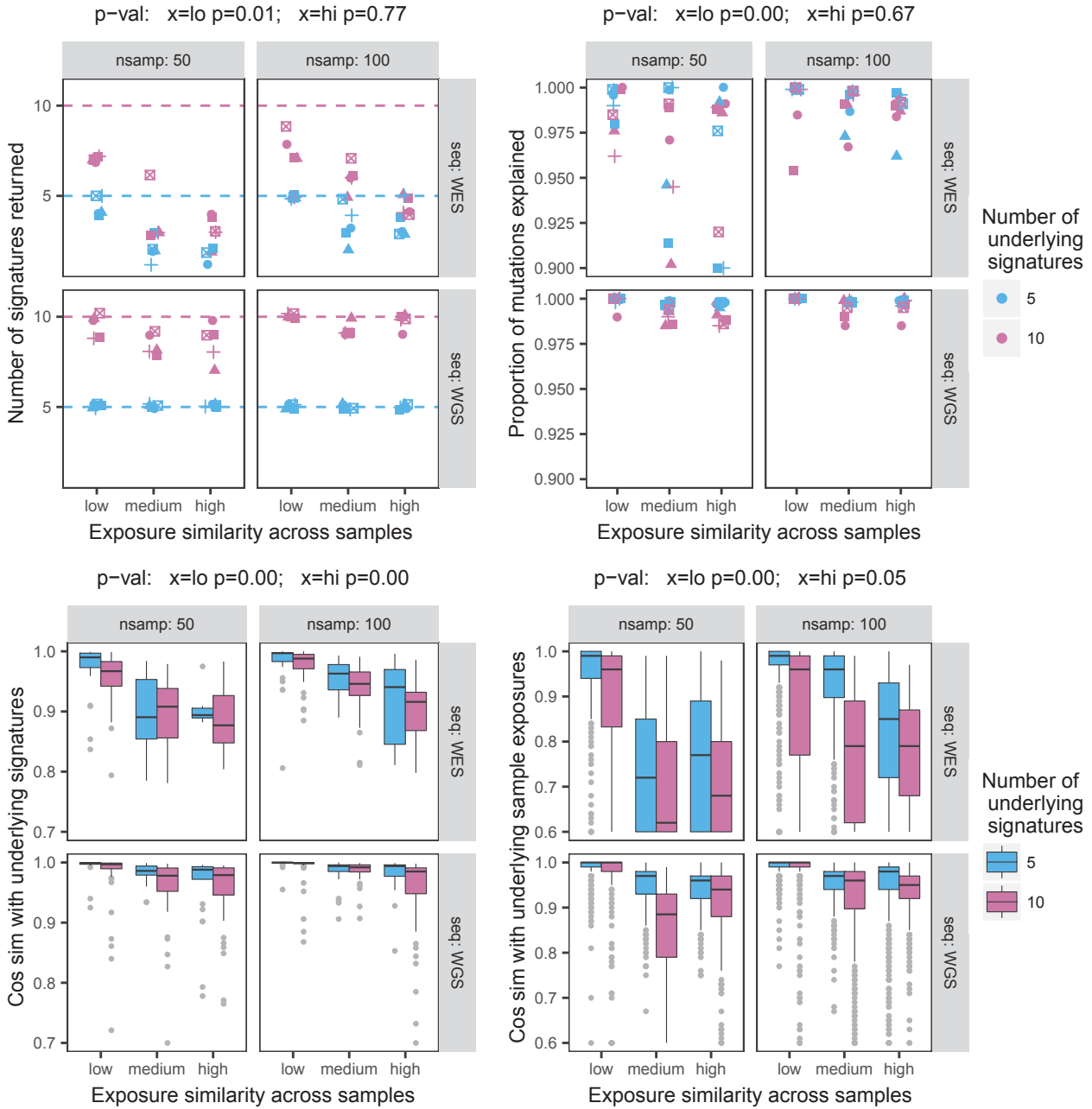


Figure 4.7: HDP performance as the level of signature exposure similarity across samples varies. P -values above each plot are for the level of exposure similarity compared to a ‘medium’ baseline, controlling for number of samples, number of underlying signatures, and sequencing type. Exposure similarity was controlled by a weight on Dirichlet concentration parameters for the sample exposure vectors when generating the simulated data (Section 4.3.1).

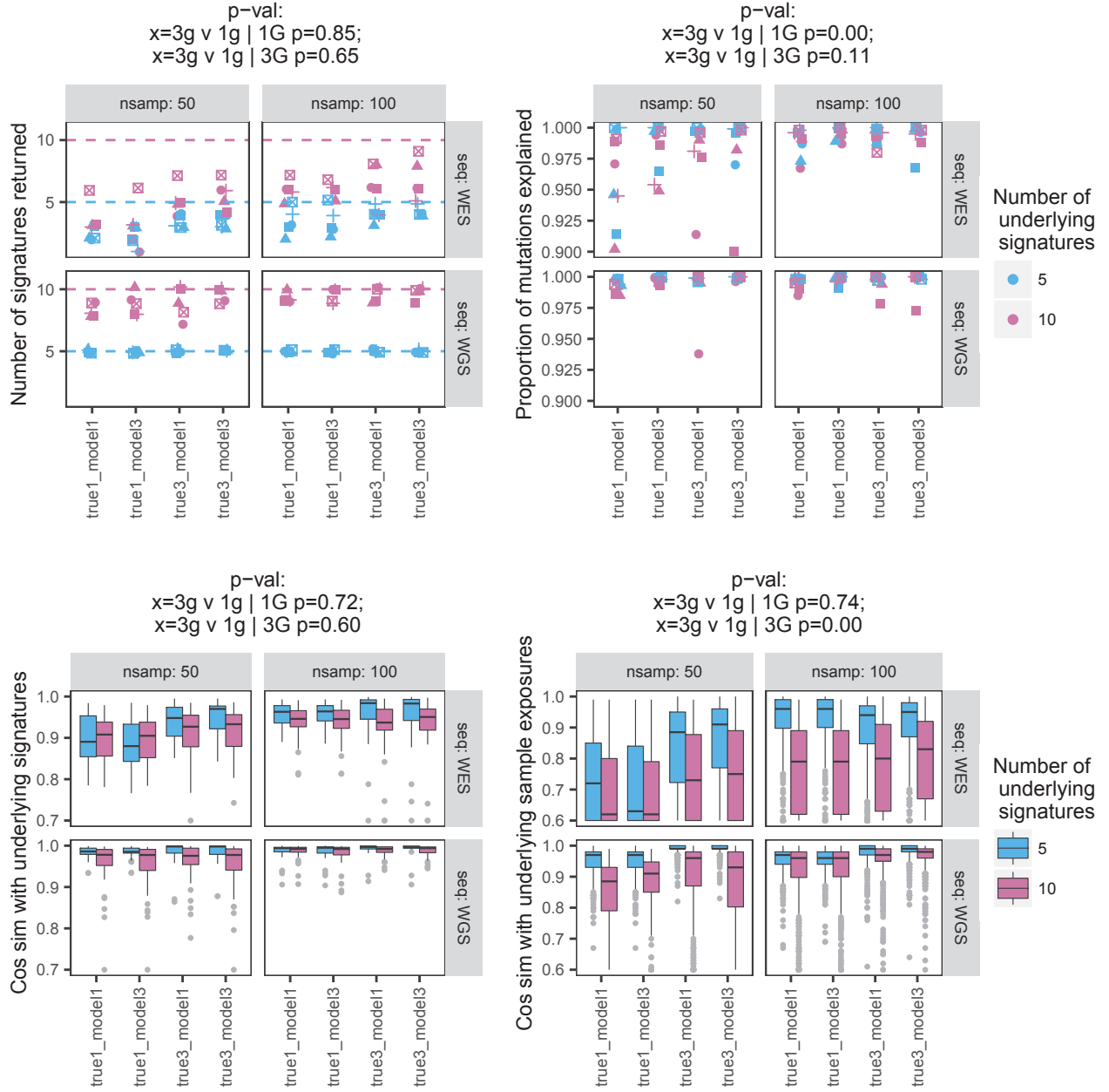


Figure 4.8: HDP performance when modelling the sub-group structure of samples, in cases where this sub-group structure genuinely existed (true3) and in cases where it did not (true1). P -values above each plot compare the 3-group model with the 1-group model, given that the dataset was simulated from one true group or from three true groups. Regression tests controlled for number of samples, number of underlying signatures, and sequencing type.

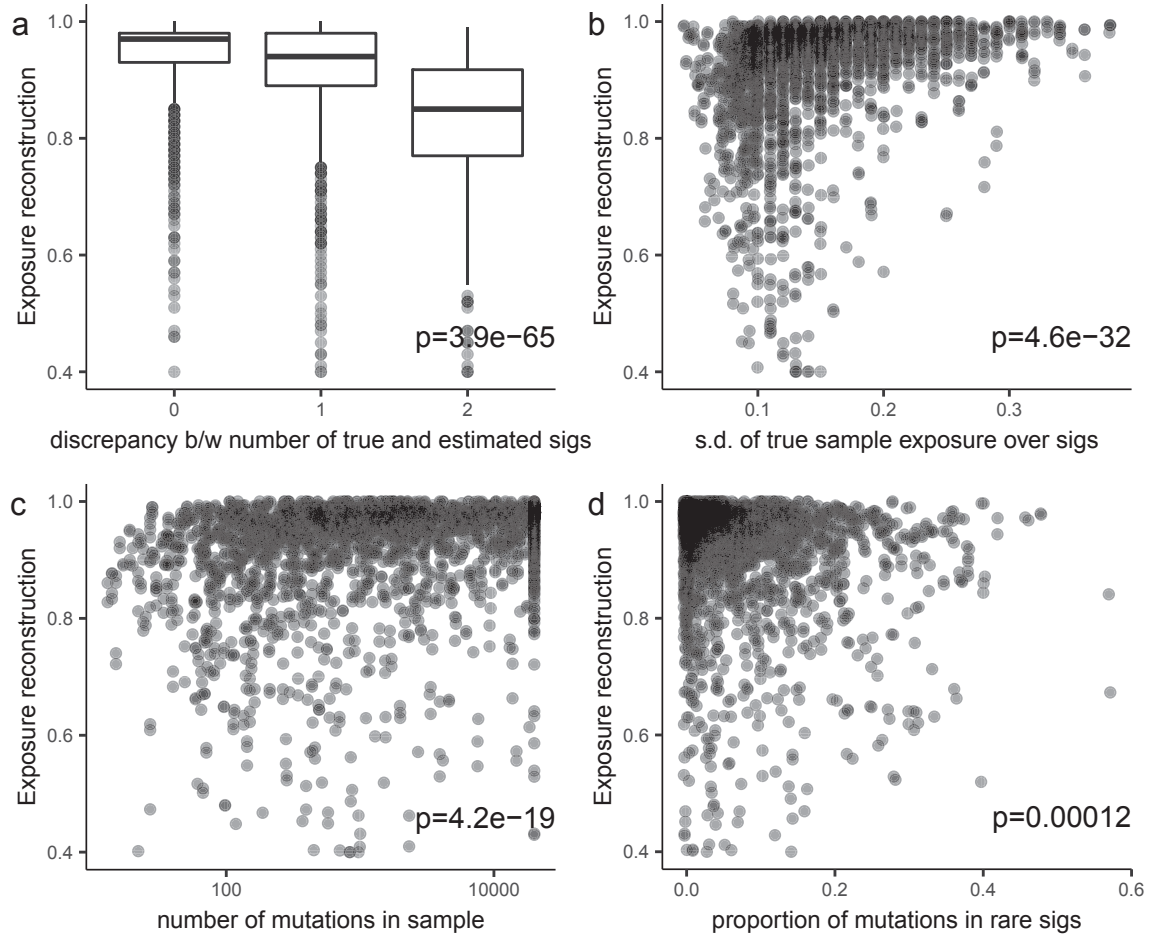


Figure 4.9: Factors influencing cosine similarity between true and estimated sample exposures (vertical axis). P -values shown are from a multivariate beta regression fit on the four predictor variables: (a) discrepancy between the number of underlying signatures and the number of signatures returned by HDP for the cohort; (b) standard deviation of the true exposure values for a sample; (c) number of mutations in a sample; and (d) proportion of mutations in a sample from ‘rare’ signatures (defined as the maximal subset which cumulatively contribute less than 10% of total cohort mutations).

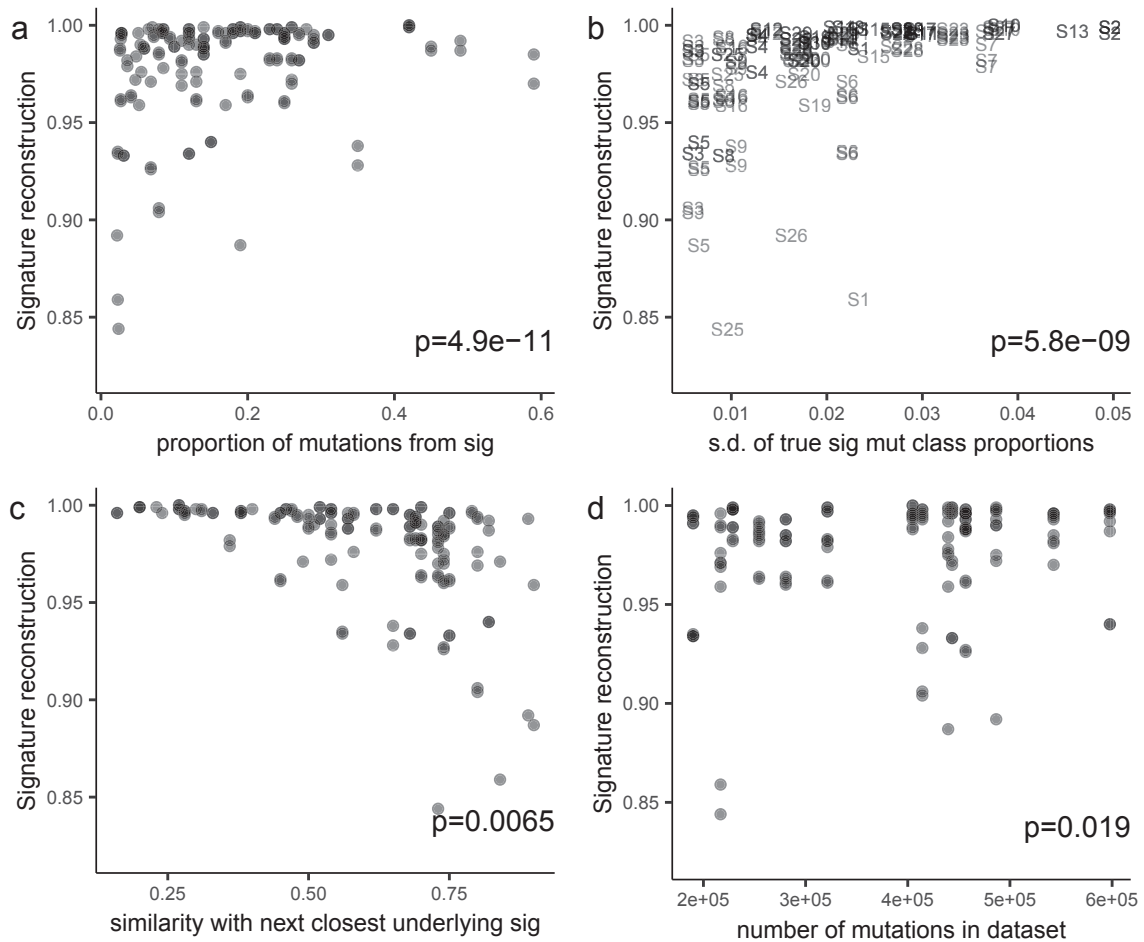


Figure 4.10: Factors influencing cosine similarity between true and estimated mutational signatures (vertical axis). P -values shown are from a multivariate beta regression fit on the four predictor variables: (a) proportion of mutations in the cohort from that signature; (b) standard deviation of the true mutation class probabilities in that signature; (c) cosine similarity with the most similar generating signature in the cohort; and (d) total number of mutations in the cohort. For panel (b), the underlying signatures are marked with their identifier in the COSMIC database (horizontal position for underlying signature s.d. is constant across datasets).

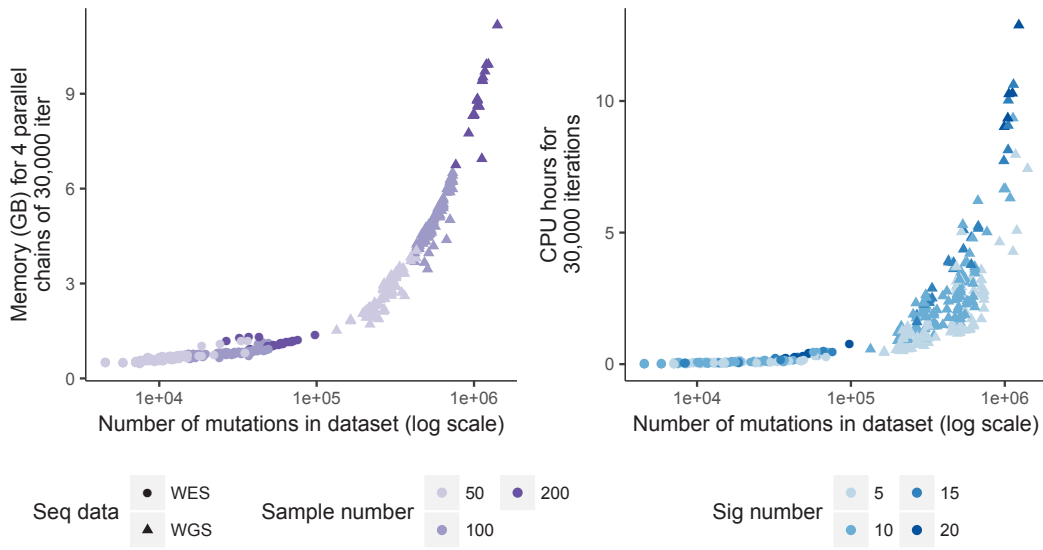


Figure 4.11: Computational resources for HDP posterior inference (MCMC approach) on simulated datasets.

Computational resources The computational time and memory required for HDP inference with MCMC (Figure 4.11) scales with the number of mutations that must be iterated over and tracked through successive cluster allocations. The CPU time also increases with the complexity of the data (number of underlying signatures), as the volume of calculations at each step relates to the number of clusters. The easiest way to reduce computational cost is to sub-sample the mutation set in hypermutators, thereby reducing the number of data items. The memory requirements are also reduced by collecting fewer posterior samples, and time in human hours (rather than CPU hours) is reduced by running more chains in parallel, particularly after the burn-in period.

4.4 Application to SNVs in original signature discovery dataset

In the first major effort to describe mutational signatures in a large pan-cancer somatic SNV dataset, Alexandrov et al. (2013b) applied NMF to almost 5 million mutations from over 7000 samples (mostly exomes) representing 30 different cancer types. By focusing on 96 SNV classes in a trinucleotide context, the original report presented 27 consensus signatures, including: 22 which validated (including two versions of the CpG deamination ‘signature 1’); 3 confirmed artefacts; and 2 unable to be validated. The COSMIC database (Forbes et al.,

2015) has since released an updated set of 30 signatures (numbers 22–30 not reported in the 2013 paper) extracted with the same methods from an updated set of more than 10,000 samples (Alexandrov et al., 2015b). In this section, I return to the original signature discovery cohort of ~ 7000 samples (summarised in Table E.3), and compare HDP results in a practical real-world dataset.

4.4.1 Model design, combining exomes and genomes

One obstacle to combining exome and genome data in signatures analysis is the difference in background trinucleotide frequency (Figure D.16). To take the extreme examples, **ATA** has a trinucleotide frequency of 4.1% in the whole genome but 2.7% in the exome (ratio 1.5)¹, while **GCG** has frequency 0.47% in the genome but 1.3% in the exome (ratio 0.36)¹. The upshot of this discrepancy is that the same underlying mutational process will present with different mutation class proportions in exome or genome data. In their original signatures analysis paper, Alexandrov et al. (2013b) ran NMF in separate sample groups divided by cancer type and sequencing type (exome or genome), then matched the signatures post hoc, adjusting for exome biases on the signature distributions at this stage^m.

In contrast, I choose to pool the exome and genome data, and fit the HDP signatures model to all samples simultaneously, grouping cancer types by parent nodes as illustrated in Figure 4.1. This approach empowers the clustering method to share information across cancer type boundaries, while upholding the prior expectation of significant differences between groups. However, mutation class tallies in the exome samples require adjustment to reflect mutational signatures on a comparable background.

For an exome sample j with observed 96-length mutation class count vector $\boldsymbol{\mu}_j$ and total SNV count of $\|\boldsymbol{\mu}_j\|_1 = m_j$, the adjusted mutation class counts are

$$\boldsymbol{\mu}'_j = \left\lfloor \left\{ (\boldsymbol{\gamma} \odot \boldsymbol{\mu}_j m_j^{-1}) / \|\boldsymbol{\gamma} \odot \boldsymbol{\mu}_j m_j^{-1}\|_1 \right\} \times m_j \right\rfloor ,$$

¹For trinucleotide frequency in the whole genome, I only include the callable genome regions defined in Section 3.1.1. For the exome, I include all protein-coding exons plus 100 bp flanks as variant calls are often made in flanks and off-target regions.

^mI follow Alexandrov et al. (2013b) in reporting mutation class signature probabilities as their expected relative frequency in a (human) genome-wide landscape, without normalising by background trinucleotide frequency. That is, the reported probabilities inherently account for how rare (e.g. **ACG** or **TCG**) or common (e.g. **TTT**) the context is. If the genome composition was adjusted for (maybe useful to generate species-agnostic signatures), the signature probabilities would increase for the rare contexts, and decrease for the common contexts.

where \odot denotes element-wise multiplication, γ is the genome-to-exome ratio of the trinucleotide context for each mutation class, and $\lfloor \dots \rfloor$ is shorthand for integer rounding using a modified procedure guaranteed to preserve m_j total SNV count for sample j .

Using these adjusted mutation tallies for any exome sample, and down-sampling hypermutator samples to a maximum of 20,000 SNVs eachⁿ, I allocated each sample to a leaf node using the HDP design for multiple cancer type groups as in Figure 4.1.

In the first instance, I ran four independent burn-in chains for 15,000 iterations, each separately initialised with 30 random clusters. Picking up from the end of each initial chain, I started another four independent MCMC chains for a further 10,000 burn-in iterations and then collected 50 posterior samples at intervals of 300 iterations (800 total samples from 16 separate chains).

4.4.2 Sampling chain diagnostics

Theoretically, an infinitely long MCMC chain would sample all possible cluster allocations in proportion to their likelihood. In practice, we aim to have a finite posterior sample set that approximates the true random sampling space without strong biases imparted by the initialisation state or by slow mixing between successive iterations. The diagnostic plots in Figure 4.12 show no strong trends in the likelihood or number of raw clusters across the MCMC chains which might indicate poor sampling. In future method development, it would be beneficial to include more formal convergence diagnostics.

4.4.3 HDP signature and exposure estimates

Using the method outlined in Section 4.2.2^o, the HDP model returned 54 consensus mutational signatures and assigned 19.8% of mutations (on average) to the zero component for noise and uncertainty.

For each HDP-estimated mutational signature ('HSig'; all presented in Figure D.17), I matched the mean mutation class distribution with its closest

ⁿ30 hypermutator samples downsampled.

^oI set the similarity threshold for signature merging to 0.92 as the COSMIC set includes one pair with this level of similarity. Also, I required every reported signature to have significant exposure (95% credibility interval above zero) in at least two samples.

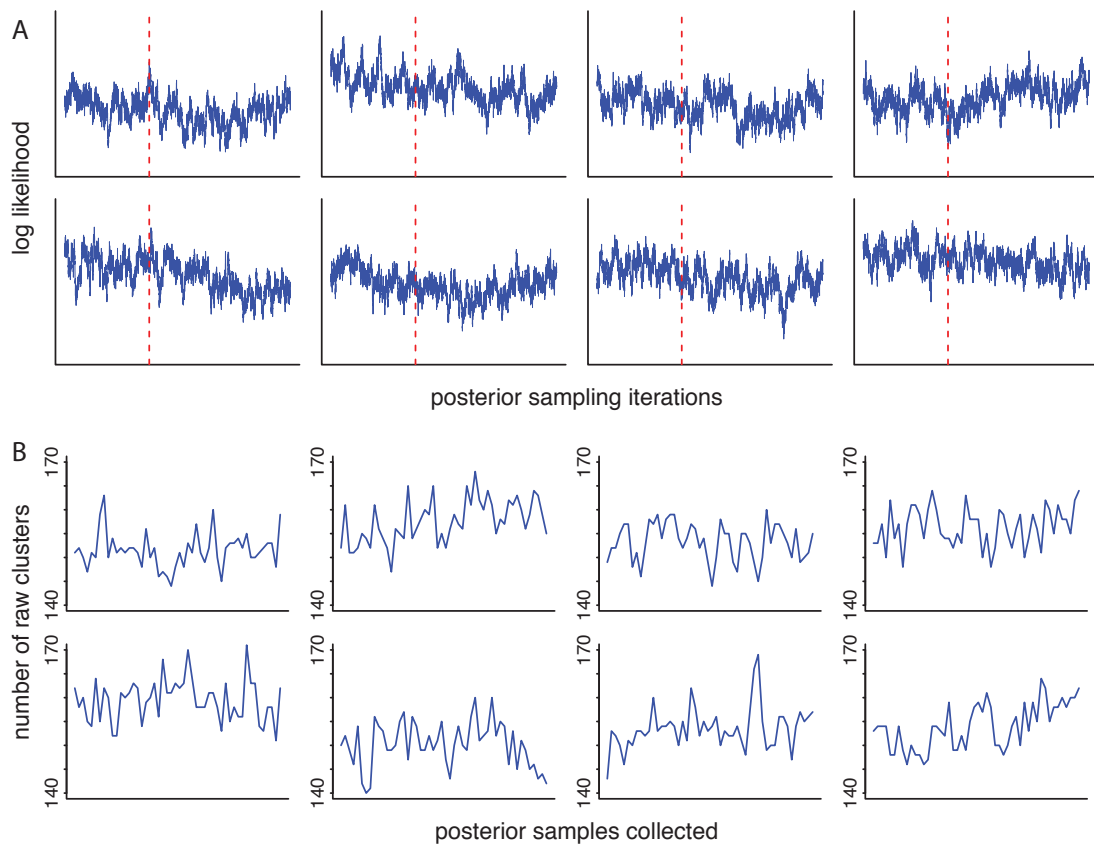


Figure 4.12: Diagnostic plots to assess HDP sampling chains for the signature discovery dataset (8 of 16 chains shown). (a) Log-likelihood of HDP state at each iteration, showing the end of the burn-in period up to the red dashed line, with posterior samples collected at regular intervals thereafter. (b) Number of raw data clusters at each collected posterior sample (50 per chain).

match in the current COSMIC set ('CSig'; including the artefact and un-validated signatures from Alexandrov et al. (2013b)), down to 0.9 cosine similarity.

Of the artefactual and un-validated signatures described by Alexandrov et al. (2013b), HDP only recovered the three artefacts R1–R3. Of the 21 validated signatures (CSig1–CSig21), HDP recovered all but four. The four missing signatures were CSig3, CSig5, CSig6, and CSig19. COSMIC annotates CSig3 as a HR-deficiency signature, and CSig5 as a 'clock-like' process associated with age (Alexandrov et al., 2015b). CSig3 and CSig5 have relatively uniform mutation class profiles which HDP may struggle to differentiate in exome data, presumably apportioning many of these mutations to the zero component with uncertain allocation. Part of the HR-deficiency signature is probably captured by HSig7, with a 0.88 similarity to CSig3 and frequent contribution to the breast cancer cohort. CSig6 is annotated as defective DNA mismatch repair, often co-occurring with the other mismatch repair signatures CSig15 and CSig20. Of the HDP-estimated signatures, HSig30 matches CSig15 extremely closely (0.98 similarity) while HSig10 matches CSig20 quite roughly (0.91 similarity). Further investigation reveals that HSig10 is a much closer match to a blend of CSig20 and CSig6^P, so it seems that HDP does not distinguish between these aspects of defective mismatch repair. The missing CSig19 is solely identified in pilocytic astrocytoma (Alexandrov et al., 2013b), and is not apparent in the HDP output. Considering the signature-tissue overview presented in Figure 4.13, the mutations originally attributed to CSig19 are presumably subsumed by the zero component.

Of the nine validated signatures subsequently added to the COSMIC database after analysis of more data (Alexandrov et al., 2015b), HDP recovered CSig22 (aristolochic acid) and CSig28 (mostly T>G in NTT) without requiring the extra samples. This suggests the HDP method may have greater sensitivity for detecting some genuine signatures.

The NMF analysis of this dataset recovered two versions of the common CSig1 CpG deamination signature (Alexandrov et al., 2013b). Similarly, HDP outputs three signatures resembling CSig1: HSig17 as a very pure distribution of C>T in NCG (even stronger peaks than the current COSMIC estimate); and HSig13 and HSig9 as relatively 'muddled' versions (Figure D.17) with particular prominence in esophageal and breast cancers respectively. It seems likely that these latter estimates mix different underlying processes. Two other COSMIC signatures are also represented multiple times in the HDP output. The CSig7 UV radiation

^P0.97 cosine similarity between HSig10 and a 60:40 mixture of CSig20 and CSig6.

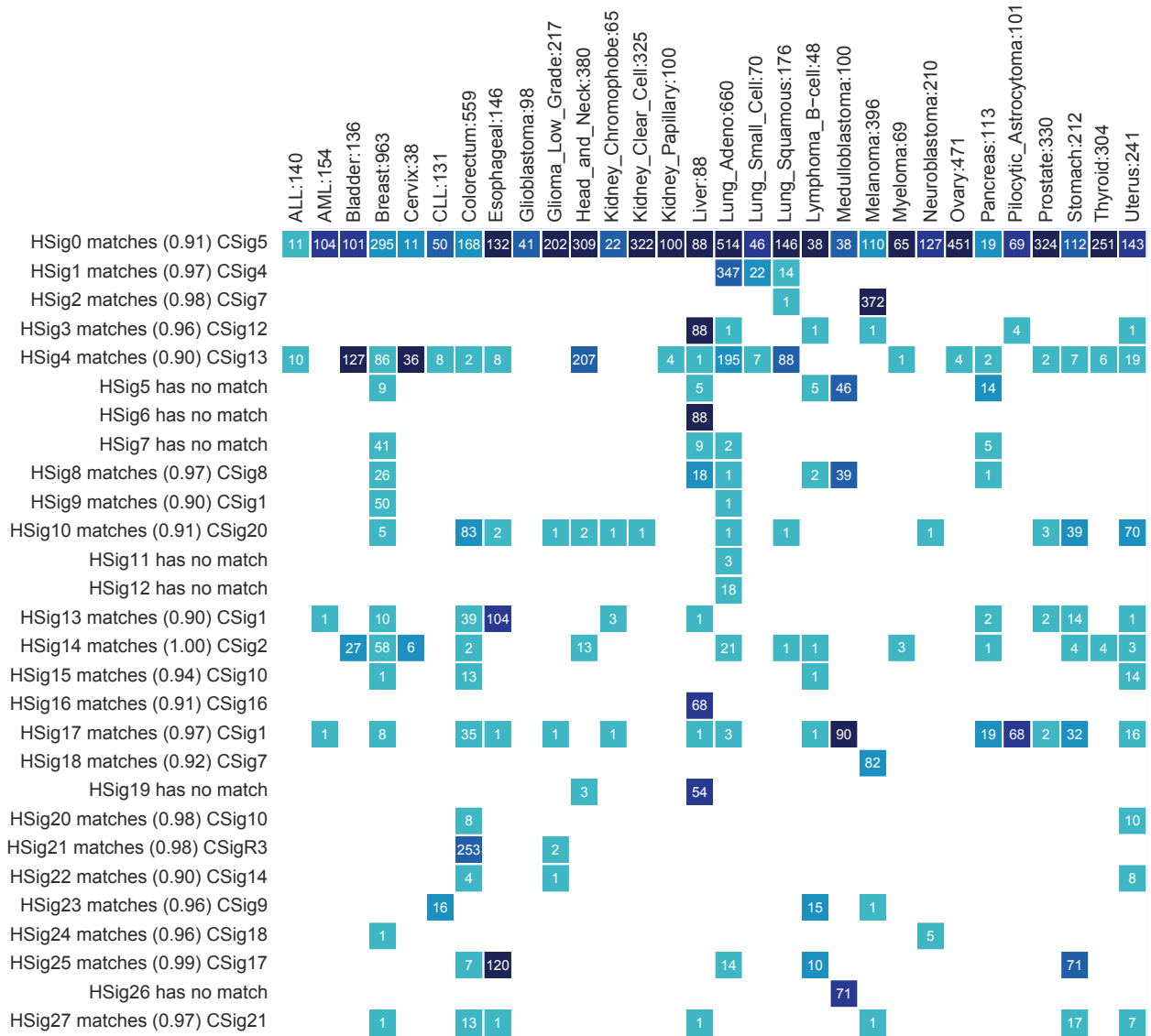


Figure 4.13: Number of samples with significant exposure to HDP-extracted signatures (95% credibility interval above zero). HDP signatures (‘HSig’) are labelled with their closest match in the COSMIC signature library (‘CSig’), with known artefacts prefixed ‘R’. ‘HSig0’ denotes the zero component for noise and uncertainty. The number of samples considered is indicated in the column label for each cancer type. Figure continues on the next page.

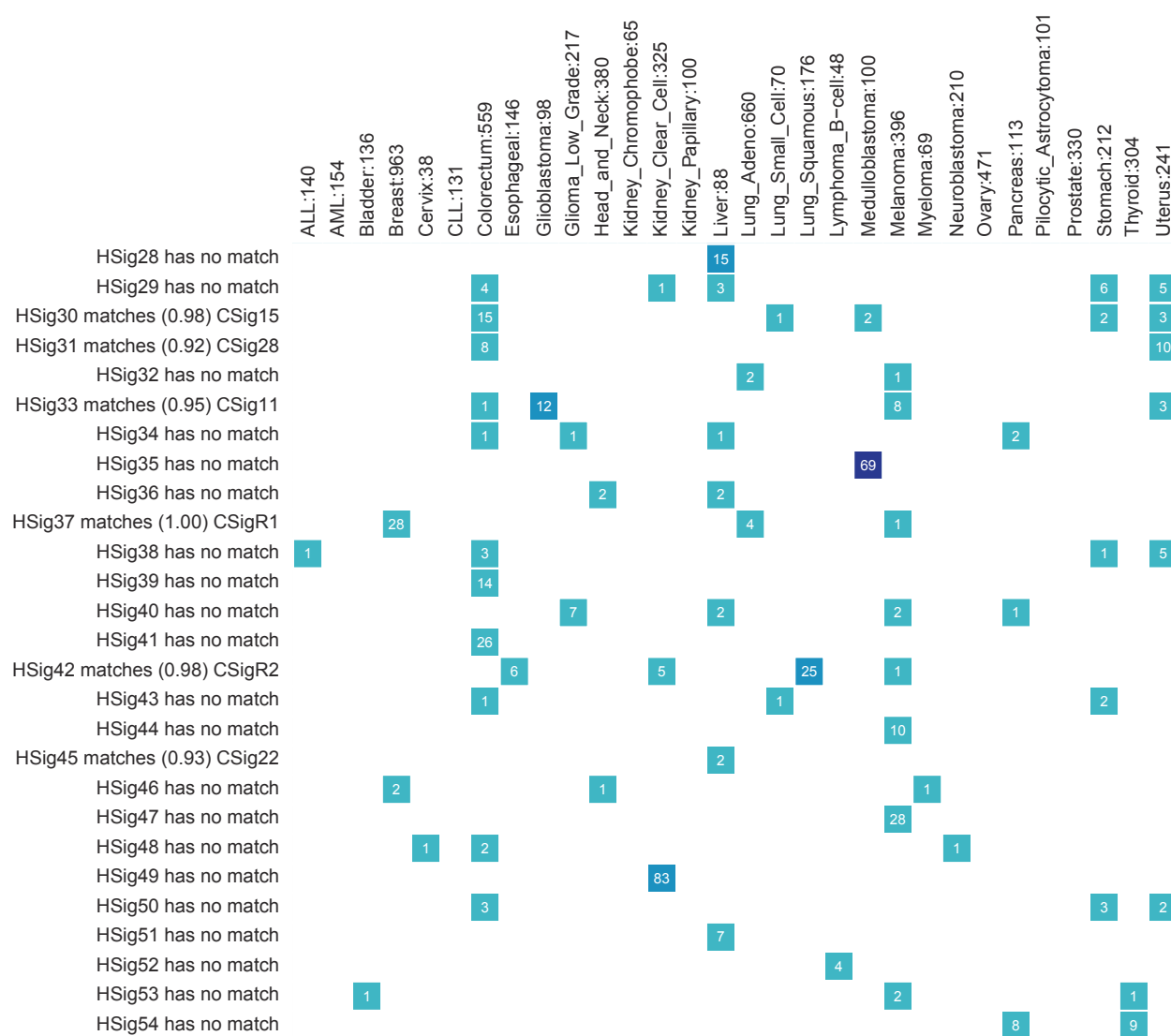


Figure 4.13: Number of samples with significant exposure to HDP-extracted signatures (95% credibility interval above zero)—continued from previous page.

signature is matched by both HSig2 and HSig18, both with strong exposure in the melanoma cohort (Figure 4.13). It remains unclear whether this is a genuine biological difference in the UV profile affecting some samples, or a false positive signature split caused by over-fitting. Finally, the CSig10 signature of mutant POLE activity is matched by both HSig15 and HSig20, differentiated by the relative probabilities of C>A in TCT and C>T in TCG (Figure D.17). This signature split may plausibly reflect biological variation in the POLE effect of different mutant protein residues (Rayner et al., 2016; Campbell et al., 2017a).

Of the 28 HDP signature estimates with no match in the COSMIC database, some are close to uniform with similar patterning to other known signatures (centre of Figure 4.14), whereas others have unique, distinctive peaks (edges of Figure 4.14). Some novel signatures with particularly clear patterns (see all in Figure D.17) include:

- HSig19 in 54 liver cancers (T>C in ATN, possibly a cleaner extraction of CSig16/HSig16);
- HSig28 in 15 liver cancers (C>T in TCT);
- HSig29 in 19 various samples including stomach, uterus (T>C in NTG);
- HSig40 in 12 various samples, including 7 gliomas (C>G in WCW);
- HSig44 in 10 melanomas (T>C in GTT);
- HSig47 in 28 melanomas (T>N in TTT);
- HSig49 in 83 kidney clear cell cancers (T>C in NTY);
- HSig51 in 7 liver cancers (T>G in GTN); and
- HSig53 in rare bladder, melanoma and thyroid samples (C>G in CCN).

The tendency for many novel signature estimates to group similar mutation classes supports their biological validity, as the model regards all 96 SNV categories as equally independent. Furthermore, the novel melanoma signatures in a TpT context are consistent with the known modality of UV radiation causing thymine dimer lesions. However, validating these signature estimates as sequencing artefacts, genuine mutational processes, or over-fitting to random data correlations, is beyond the scope of this work.

In contrast to the simulated data (Section 4.3) for which HDP typically consigned one or two percent to the zero component, this real-world example grouped almost 20% of total mutations in the zero component for noise and uncertainty.

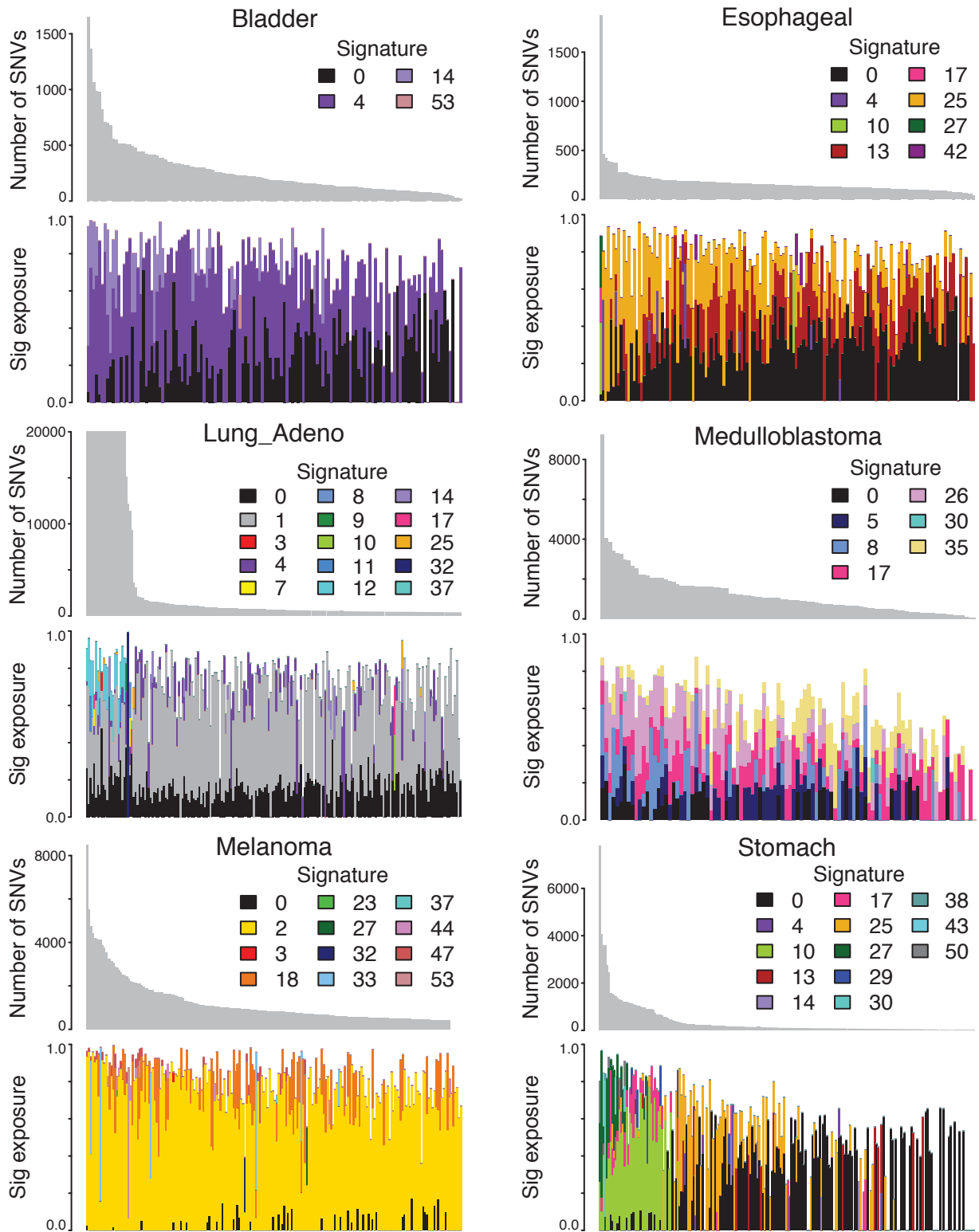


Figure 4.15: Average estimated sample exposures to HDP-extracted signatures for six cancer types with samples sorted by observed mutational burden, capped at 20,000 SNV. Large cohorts are subset to a maximum of 200 samples for presentation. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain signature allocation.

4.5 Simultaneous signature matching and discovery

One of the key theoretical advantages of the HDP method for mutational signatures analysis is the ability (outlined in Section 4.2.3) to match a new dataset to an existing library of known signatures while simultaneously empowering any novel mutational signatures in the dataset to emerge as separate clusters.

Data and methods

To briefly illustrate this approach on real data from the PCAWG cohort, I selected somatic SNV calls from the pancreatic endocrine cancer group (81 samples; 252,930 SNVs) and prostate adenocarcinoma group (198 samples; 635,688 SNVs). Following the HDP design illustrated in Figure 4.2, the 30 known mutational signatures in the current COSMIC database⁹ were each represented by 500 pseudo-count mutations in a frozen node. Analysing each cohort separately, I randomly initialised the real mutations into 35 clusters—30 linked with a prior signature, and five others solely comprised of observed mutations from the new dataset. After running four burn-in chains for 10,000 iterations, two chains bifurcated from the end of each burn-in and ran a further 2000 burn-in iterations before collecting 125 posterior samples separated by 100 iterations (1000 total posterior samples from eight independent MCMC chains).

Results

For the pancreatic endocrine cohort, HDP identifies four additional signatures (shown in Figure 4.16) while simultaneously matching to the set of known signatures. The newly extracted signature N1 is characterised by a consistent distribution of C>A mutations, with high-confidence peaks in TCT and TCA contexts. Scarpa et al. (2017) recently attributed this pancreatic neuroendocrine signature to *MUTYH* loss and consequent deficiency in base excision repair. Seven samples have significant exposure to this *MUTYH* signature, including three with high overall burden (> 7500 SNV) and > 92% of mutations attributed to signature N1 (Figure 4.17). Signatures N2 and N3 have greater uncertainty about their mutation class distribution, and have significant exposure in three and six samples respectively, with estimated exposure proportions as high as

⁹<http://cancer.sanger.ac.uk/cosmic/signatures> available as of December 2017

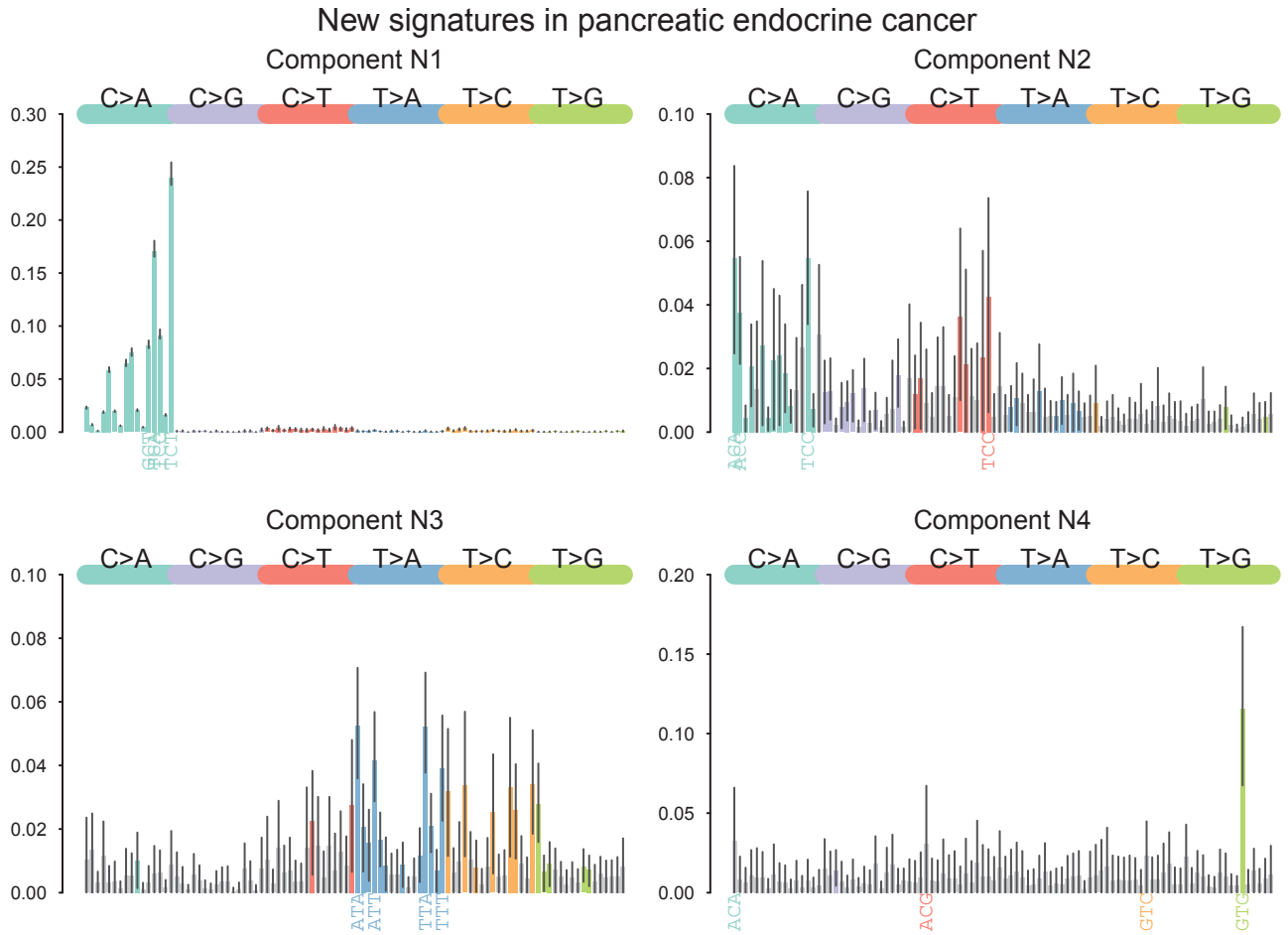


Figure 4.16: Newly discovered mutational signatures in pancreatic endocrine cancer WGS cohort (mean and 95% credibility interval from HDP posterior samples, with non-significant mutation classes in grey and four major peaks labelled with trinucleotide context). Component N4 is similar to the known artefact signature R1.

24% and 27%. Signature N4 has one dominant peak of T>G in GTG and probably corresponds to the known artefact signature ‘R1’ (Alexandrov et al. (2013b), artefact signatures not included as priors). Across the cohort, samples also have significant exposure to many prior COSMIC signatures, including age-related signatures 1 and 5 (Alexandrov et al., 2015b), APOBEC signatures 2 and 13, and signature 8 (low C>A peaks, with CC>AA double nucleotide substitutions). One unusual sample has 66% of 1463 SNVs attributed to signature 12 (peaks in T>C), previously described in liver cancer only (Alexandrov et al., 2013b).

For the prostate cohort, HDP identifies six novel signatures shown in Figure 4.18. Signature N1 is quite common in the cohort (Figure 4.19), with significant exposure in 30 samples including several with over 3000 SNVs attributed to this novel signature. In contrast, signature N6 has significant exposure in

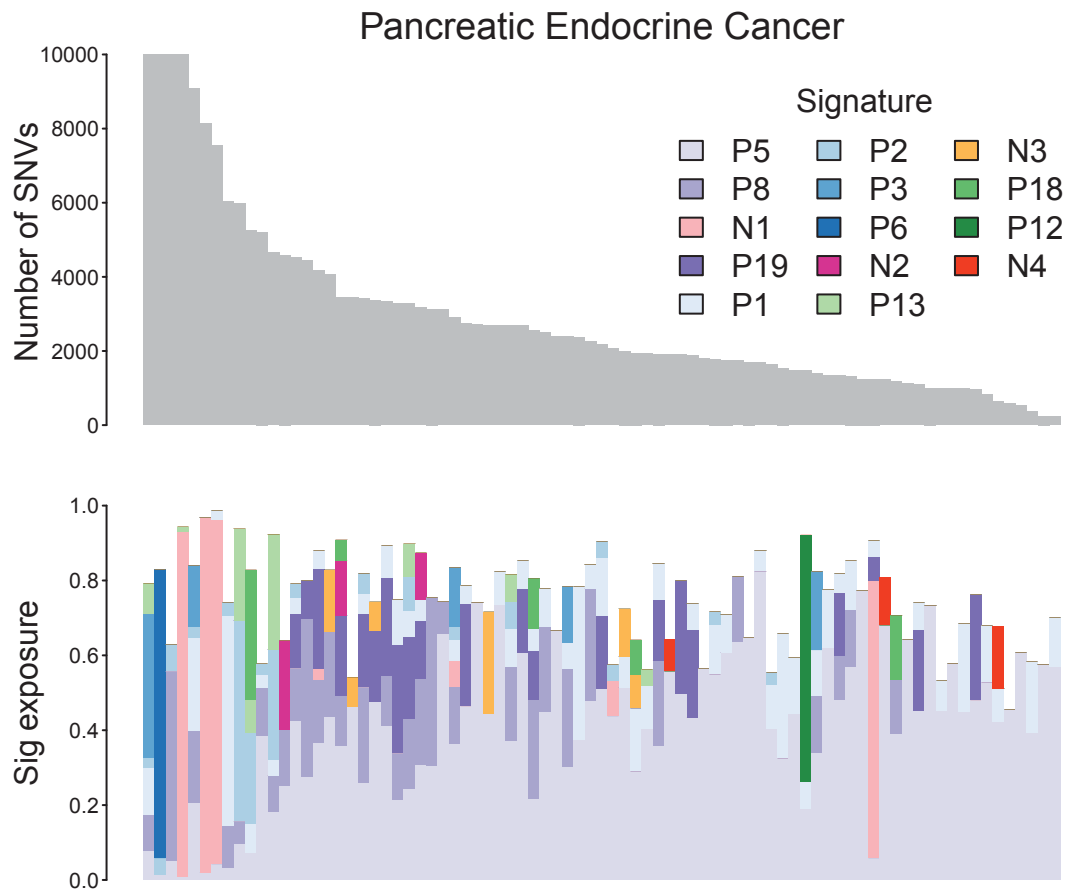


Figure 4.17: Pancreatic endocrine cancer sample exposures (average from HDP posterior samples) to a library of known signatures (labelled ‘P’ for prior) and newly discovered signatures (labelled ‘N’ for new) with samples sorted by observed mutational burden, capped at 10,000 SNVs. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain signature allocation.

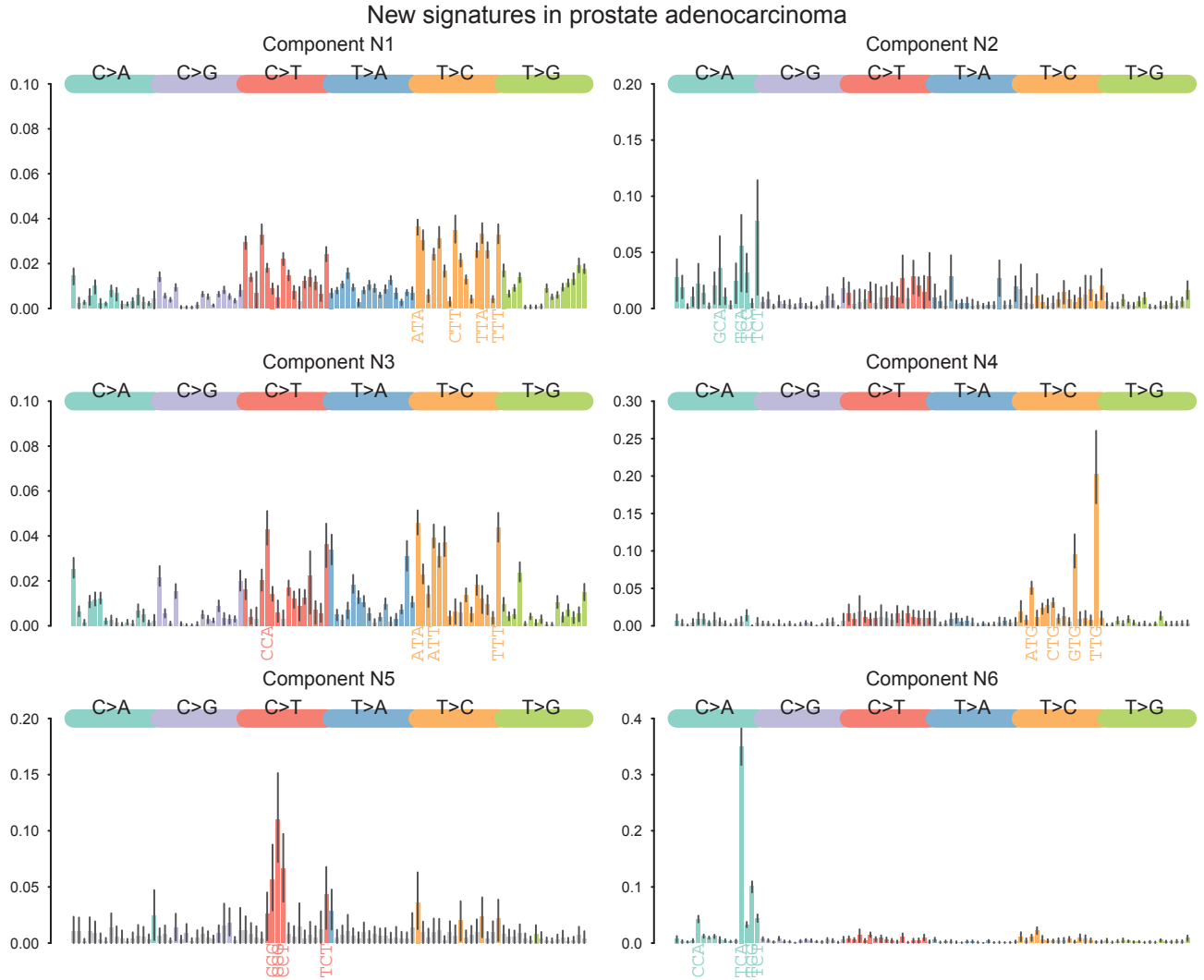


Figure 4.18: Newly discovered mutational signatures in prostate adenocarcinoma WGS cohort (mean and 95% credibility interval from HDP posterior samples, with non-significant mutation classes in grey and four major peaks labelled with trinucleotide context).

just one sample, contributing an estimated 70% of its 2775 SNV calls, with a huge peak of C>A mutations in a TCA context. Prostate samples also have significant exposure to some prior COSMIC signatures, including 1, 5, and 8. Interestingly, one sample in the prostate cancer cohort has almost 1000 SNVs attributed to COSMIC signature 9, thought to be the mark of polymerase η activity and previously identified in CLL and B-cell lymphomas only (cells with AID hypermutation) (Alexandrov et al., 2013b). This HDP finding implicates rare polymerase η activity under other conditions in prostate tissue.

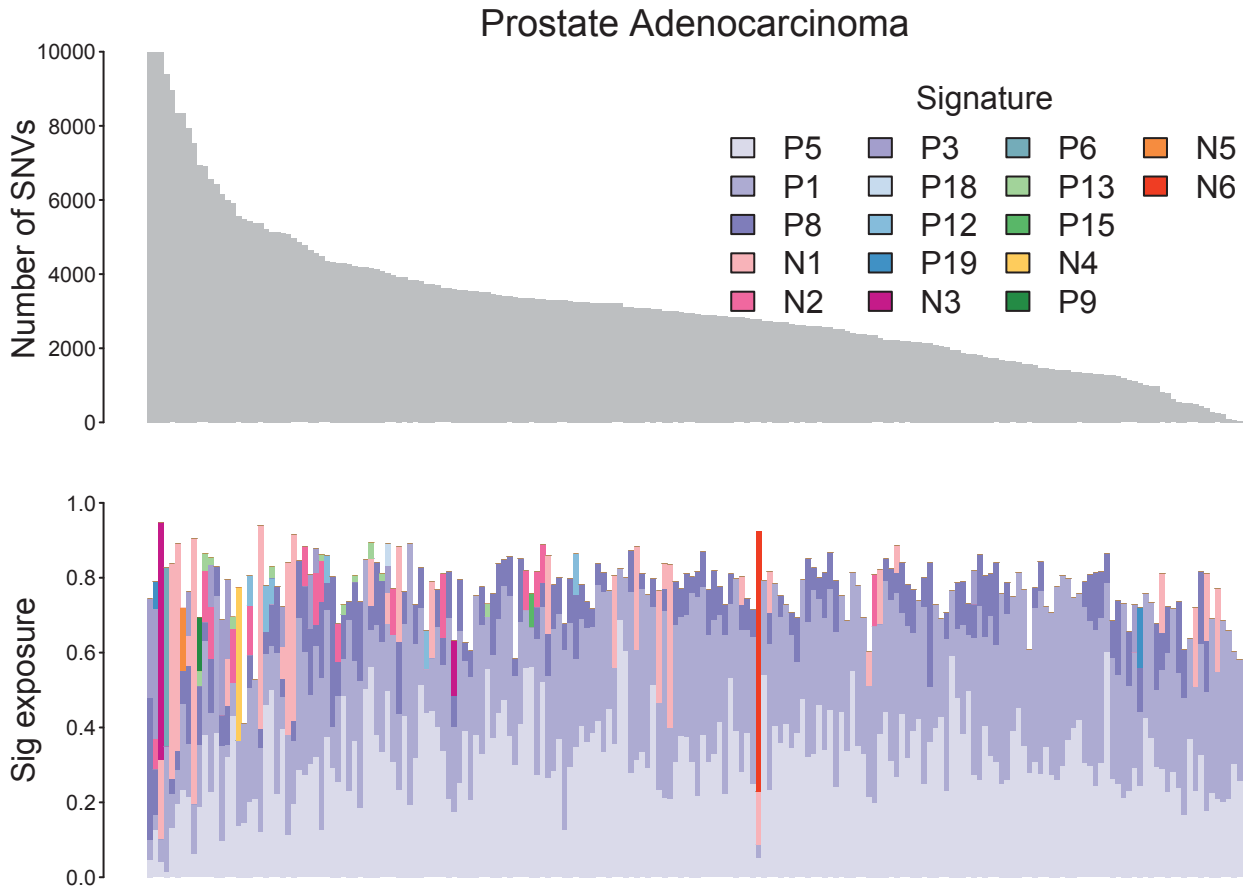


Figure 4.19: Prostate cancer sample exposures (average from HDP posterior samples) to a library of known signatures (labelled ‘P’ for prior) and newly discovered signatures (labelled ‘N’ for new) with samples sorted by observed mutational burden, capped at 10,000 SNVs. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain signature allocation.

Discussion

Any newly discovered pattern of co-occurring mutation classes may: reflect a genuine mutational process previously missed by signature analysis; be a variant form of a previously described signature (genuine biological variant, or different form due to calling bias); or be a specific profile of artefact calls from contamination, sequencing errors, etc. Validating the new signatures described in pancreas and prostate as artefact or genuine is beyond the scope of this thesis.

While these results demonstrate the efficacy of this HDP approach for discovering novel mutational signatures and quantifying the uncertainty in their distribution with credibility intervals, interpretation of the prior signature matching has possible pitfalls.

First, when observed mutations from the new dataset cluster with fixed pseudo-counts corresponding to a prior signature, the distribution over mutation classes in that cluster can sometimes deviate substantially from the prior signature as primed by the pseudo-counts. The signature extraction method attempts to resolve this problem when reporting results, and will de-couple a cluster from the prior identity of its pseudo-counts if the overall pattern has diverged from the original intended signature. With the current implementation, the final estimated signature will still be labelled with its closest match in the prior set down to a threshold of 0.85 cosine similarity. As a result, reported exposure to a prior signature may sometimes indicate a rough match only. For example, the cosine similarity between the reference version of COSMIC signature 19 and the version reported in the pancreatic endocrine cancers is 0.92, and it remains unclear whether or not this represents the same underlying process.

Second, it may be the case that more uniform signatures (low variation over mutation class proportions) are particularly difficult to distinguish when conditioning on prior knowledge. Whereas the roughly uniform COSMIC signature 5 has previously been reported in all cancer types, another roughly uniform signature 8 has only been reported in breast and medulloblastoma (Alexandrov et al., 2013b) and yet was estimated to have a significant presence in most of the pancreatic and prostate samples analysed here. It seems plausible that this signature 8 exposure stems from mis-clustering of genuine signature 5 mutations, primed by pseudo-counts with similar spread over all mutation classes. The COSMIC database reports that signature 8 has a weak transcription strand bias for C>A mutations and a tendency for double nucleotide CC>AA

substitutions, so it would be interesting to check in future research whether or not the purported signature 8 exposures in pancreas and prostate have similar distinguishing properties.

Finally, conditioning on a large number of prior signatures may increase the likelihood of a small subset of mutations consistently clustering with a known prior by chance, introducing small false positive signature exposures not adequately sampled away by the finite MCMC sample collection. For example, previous analysis with NMF reported only three mutational signatures with significant exposure in 520 prostate cancer exomes (COSMIC signatures 1, 5 and 6, Alexandrov et al. (2015b)), whereas my HDP analysis of 198 prostate genomes found significant exposure to eleven known COSMIC signatures and a further six novel signatures. This discrepancy could partially result from greater detection power in genomes rather than exomes; greater sensitivity of the HDP approach (particularly when conditioning on prior signatures); and possible false positive matching to prior signatures.

To assess how the inclusion of prior signature information impacts results, it would be informative to compare against *de novo* HDP signature extraction on these same cohorts, and see whether a pattern like COSMIC signature 8 emerges separately to COSMIC signature 5, and whether the low frequency signature exposures are still reported. The performance of HDP matching to prior signatures could also be investigated with simulation studies under a range of conditions, with particular interest in close-to-uniform signatures, and possible false-positive clustering with fixed pseudo-counts.

In practice, including *all* previously reported signatures as equally weighted priors may be naive and possibly confounding, particularly as the number of known mutational signatures will continue to grow. I conjecture that a better approach would be to weight known priors by their reported prevalence in related cancer types, even to the point of excluding priors only described in completely unrelated cancer types.

If future studies come to model mutational signatures in more detail than a simple categorical distribution over 96 SNV classes—for example, including strand bias, double nucleotide substitutions, replication timing bias—then methods to match new data to previously reported signatures will have more diverse evidence to draw from, likely improving results.

4.6 Signatures of genome rearrangement

In contrast to the detailed analyses of somatic SNV signatures (Alexandrov et al., 2013b; Helleday et al., 2014; Alexandrov et al., 2015b), few publications have attempted a similar decomposition of somatic rearrangement signatures. In 560 breast cancer genomes, Nik-Zainal et al. (2016) applied NMF to a set of BPJ calls classified by their orientation, size, and presence in clustered or isolated SVs. This yielded six rearrangement signatures, broadly defined by: large or small tandem duplications (two separate signatures); small deletions; unbalanced translocations with large deletions and inversions; and intra- or inter-chromosomal complex rearrangements (two separate signatures). With essentially the same SV classification and NMF signature pipeline, Hillman et al. (2018) extracted five rearrangement signatures from 80 ovarian cancers, finding similar results with basic separation by SV class and size for isolated BPJ.

In this section, I return to the SV dataset of approximately 2500 PCAWG samples (introduced in Chapter 2) and leverage our detailed BPJ classifications (Section 2.1.3) to calculate signatures of co-occurring rearrangement patterns.

4.6.1 Generating the SV tally matrix

Using the BPJ classification in Table 2.2 as a starting point, I defined 76 SV categories to use as input alteration classes for a HDP signature analysis.

For deletion and tandem duplication, I first separated the fragile site events with both breakpoints inside one of the 21 FS regions defined in Table E.2 (Section 3.4). Then, I classified the remaining deletions and tandem duplications by both size (breaks at 50 kb, 500 kb, and 5 Mb) and replication timing (at the event mid-point; early > 60 , late < 30 , or in-between, using the definition in Section 3.1.2). Events larger than 5 Mb were not sub-categorised by replication timing.

For classified SV like translocation (unbalanced and reciprocal), 2-jumps (local and distant), chromoplexy (cycles and chains), and templated insertion (cycles, chains and bridges), I tallied the counts per-event rather than per-BPJ. Local 2-jumps and reciprocal inversions (two BPJ per event count) were additionally separated by size categories split at 100 kb (measuring the total event span). Templated insertions were divided by the size of the insert fragment (split at 5 kb), taking the median insert size for multi-insert events where necessary.

Finally, of the $\sim 150,000$ BPJ in complex unexplained clusters, I included categories for the subset of individual local footprints with a recurrent BPJ pattern present at least one hundred times in the cohort (tallied once per-footprint, not per-BPJ). In the category labels shown in Figure 4.20, the complex footprint patterns are annotated with an alphabetical segment notation, using + for the 3' end, - for the 5' end, carets for BPJ joins, and a forward slash for adjacent breakpoints in separate BPJ. The many complex unexplained BPJ in rarer, more convoluted local footprints were excluded from the signatures analysis.

After removing samples with less than three counted SV events, the final matrix tallied 147,508 SV events across 2050 samples. Of the 76 SV categories, the most common were: a single translocation breakpoint within a complex cluster (12,753) and deletions smaller than 50 kb with mid-range replication timing (11,289). The least common categories were: dup-trp-dup local 2-jumps smaller than 100 kb (31) and local+distant 2-jumps of translocation with subsequent tandem duplication (88).

4.6.2 HDP model for SV signatures

Following the HDP design for multiple cancer type groups as in Figure 4.1, I allocated each sample to a leaf node, using a separate concentration parameter for each group of child nodes and the set of all parent nodes, using gamma hyper-priors with shape = 1 and rate = 1.

I ran eight independent burn-in chains—each separately initialised with ten clusters—for 40,000 iterations, and then collected 125 posterior samples at intervals of 300 iterations (1000 total samples from 8 separate chains).

4.6.3 Estimated SV signatures

Using the method outlined in Section 4.2.2, the HDP model returned 15 consensus rearrangement signatures and assigned just 0.3% of SV events (on average) to the zero component for noise and uncertainty—a far lower proportion of uncertain clustering than for the SNV signatures in Section 4.4.

Figure 4.20 presents the fifteen PCAWG rearrangement signatures (sorted by structure, not frequency) with an inverse normalisation such that event class proportions (across signatures) sum to one. This is a different interpretation to

the standard plot showing individual signatures as proper probability distributions integrating to one. Given the extreme differences in SV class frequency, this inverted visualisation allows rare SV classes to be seen alongside common structures. However, the values shown *within* each signature need careful interpretation, as the rearrangement process will generate common SV classes far more frequently than rare SV classes at the same plotted height.

The complex SV footprints are mostly split across two signatures: one generic ‘Complex’ group, and one ‘Complex+Chromoplexy’ group co-occurring with chromoplexy cycle events. Fragile site deletions almost exclusively assort to their own ‘Fragile Site’ signature, which also includes about half the FS tandem duplications, and a range of other deletions enriched in late-replicating regions. Other deletions separate into four different signatures:

- ‘Small Deletion’, co-occurring with several other classes including: small reciprocal inversion, small insertion bridge, and reciprocal translocation;
- ‘Mid Deletion’ with few other SV classes;
- ‘Large Deletion’, co-occurring with large reciprocal inversions and reciprocal inversions within complex clusters; and
- ‘Late Deletion’ of late-replicating events at any size, also co-occurring with a small fraction of reciprocal inversion events at any size.

Tandem duplications mostly assort over five signatures:

- ‘Early Small TD’, co-occurring with templated insertions (particularly small insertion cycles) and translocation plus tandem duplication events;
- ‘Late Small TD’, co-occurring with small dup–inv–dup 2-jumps;
- ‘Early Mid TD’, co-occurring with large insertion cycles and chains;
- ‘Late Mid TD’, co-occurring with large dup–inv–dup 2 jumps; and
- ‘Large TD’ with few other SV classes.

‘Unbalanced Translocation’ forms a largely separate signature, co-occurring with a small fraction of chromoplexy chains. The ‘Reciprocal Sv’ signature pairs reciprocal translocations with other balanced events like chromoplexy cycles and some reciprocal inversions. Finally, the miscellaneous ‘Break+Ligate’ signature groups foldback rearrangements with extremely large deletions and duplications, as well as local+distant 2-jumps, chromoplexy chains, and some complex SV footprints involving foldback BPJs.

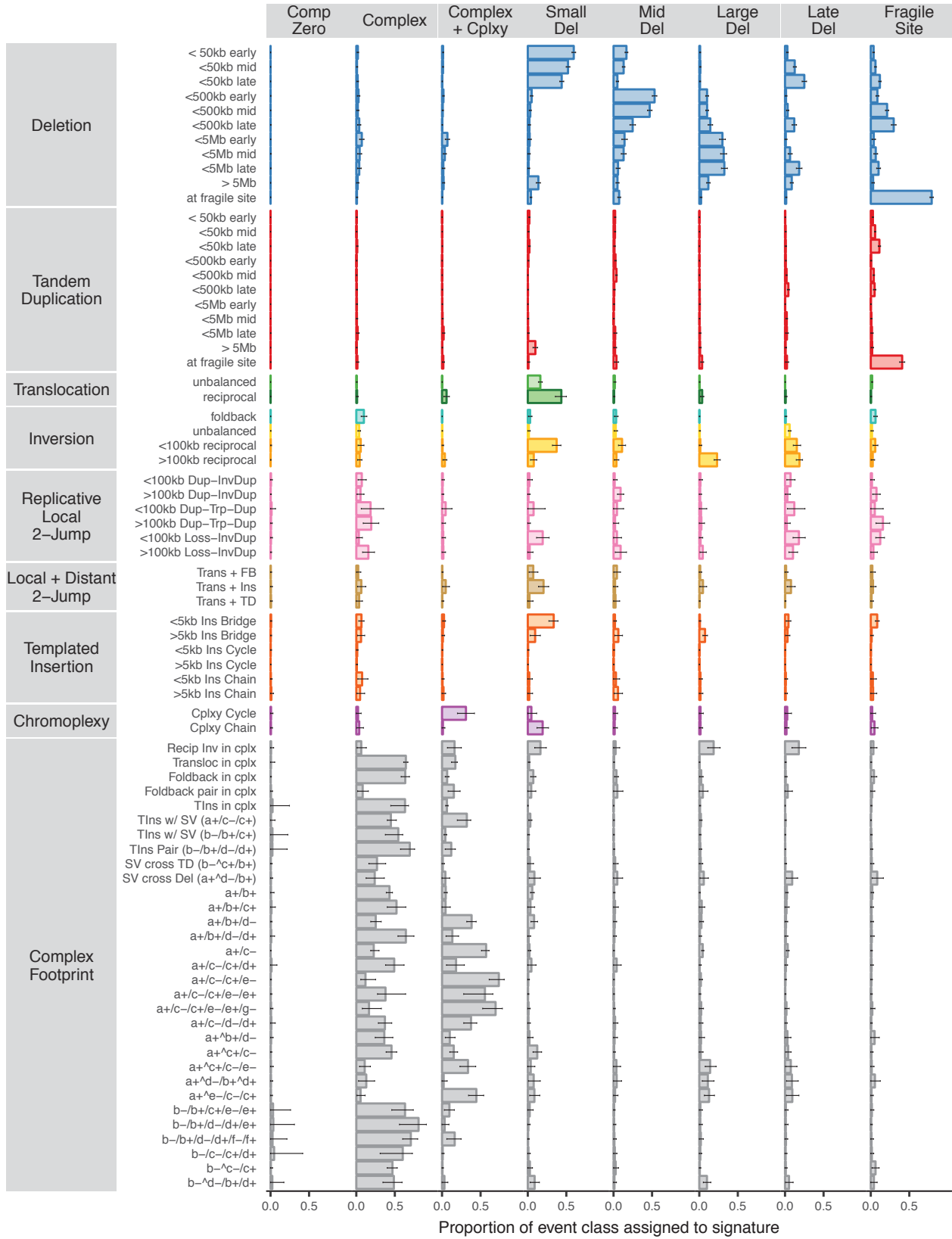


Figure 4.20: SV signatures and 95% credibility intervals, normalised by event class fraction (rows—not columns—sum to one, including the figure continuation on the next page).

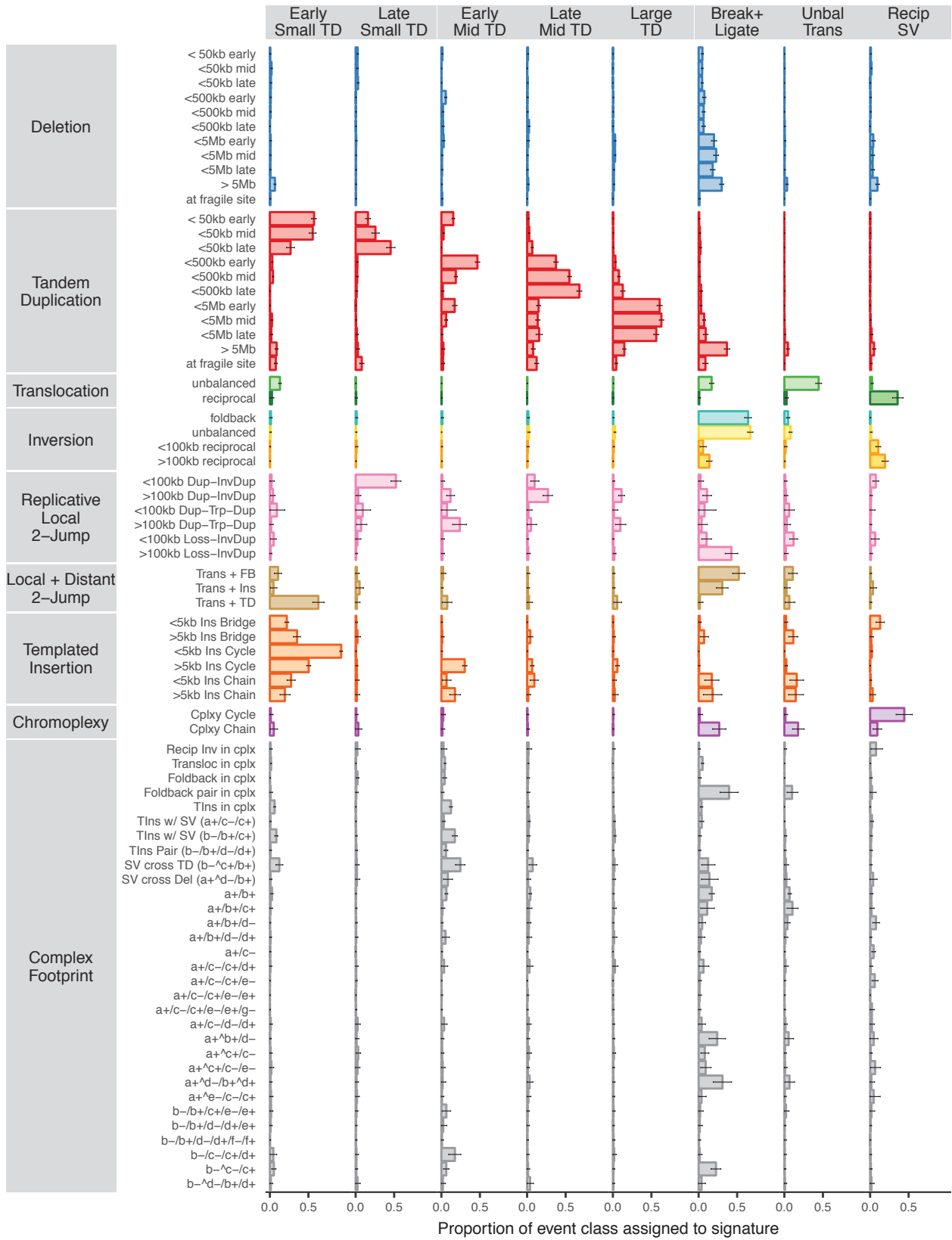


Figure 4.20: SV signatures and 95% credibility intervals, normalised by event class fraction (rows—not columns—sum to one, including the figure continuation on the previous page).

Figures 4.21 and D.18 show a subset of the estimated sample exposures, recapitulating the basic BPJ census presented in Figures 2.10 and D.2 with the enhanced signature context of size, replication timing, and SV group. Prostate cancer is particularly enriched for the late-replicating deletion and complex chromoplexy signatures; this latter exposure indicates that many BPJ in chromoplexy-associated events are found in the complex unexplained bin. Other cancer types with particularly high exposure to certain rearrangement signatures include: bladder cancer with large deletion and ‘break+ligate’ SVs; osteosarcoma with complex SVs; medulloblastoma with the unbalanced translocation signature; and colorectal cancer with the fragile site signature. Not all SV events in the fragile site signature are confined to the annotated FS regions, as the other deletions in the signature are more common than their relative values in Figure 4.20 suggest (because of the inverted normalisation to visualise common and rare SV classes concurrently). The exposure patterns in breast, liver^r and uterus highlight the different replication timing skews of tandem duplication by sample, extending the results of Section 3.2.3.

4.6.4 Discussion

The fifteen rearrangement signatures presented in this section are an apotheosis of the results presented in Chapters 2 and 3, combining SV class, size, and location (as represented by replication timing) into one decomposition of underlying rearrangement processes with characteristic structural readouts and varying activity levels across samples and groups.

The co-occurrence patterns indicate the same underlying condition may generate different structural forms with similar properties of size and/or location. For example, deletions coincide with reciprocal inversions of a similar size range, presumably mediated by break and ligate repair of DSBs at consistent intervals. Similarly, tandem duplications in late-replicating regions coincide with dup-inv-dup 2-jumps of a similar size range, presumably mediated by template and replicate repair of invading strands with consistent processivity (mechanisms reviewed in Section 1.4.1). The conditions generating fragile site deletions also foster FS tandem duplications and a range of other late-replicating deletions, possibly present in un-annotated FS regions.

Compared to previous reports of five or six relatively simplistic SV signatures

^rFor liver cancer, the outlying high-burden sample with large duplications was previously illustrated in Figure D.6.

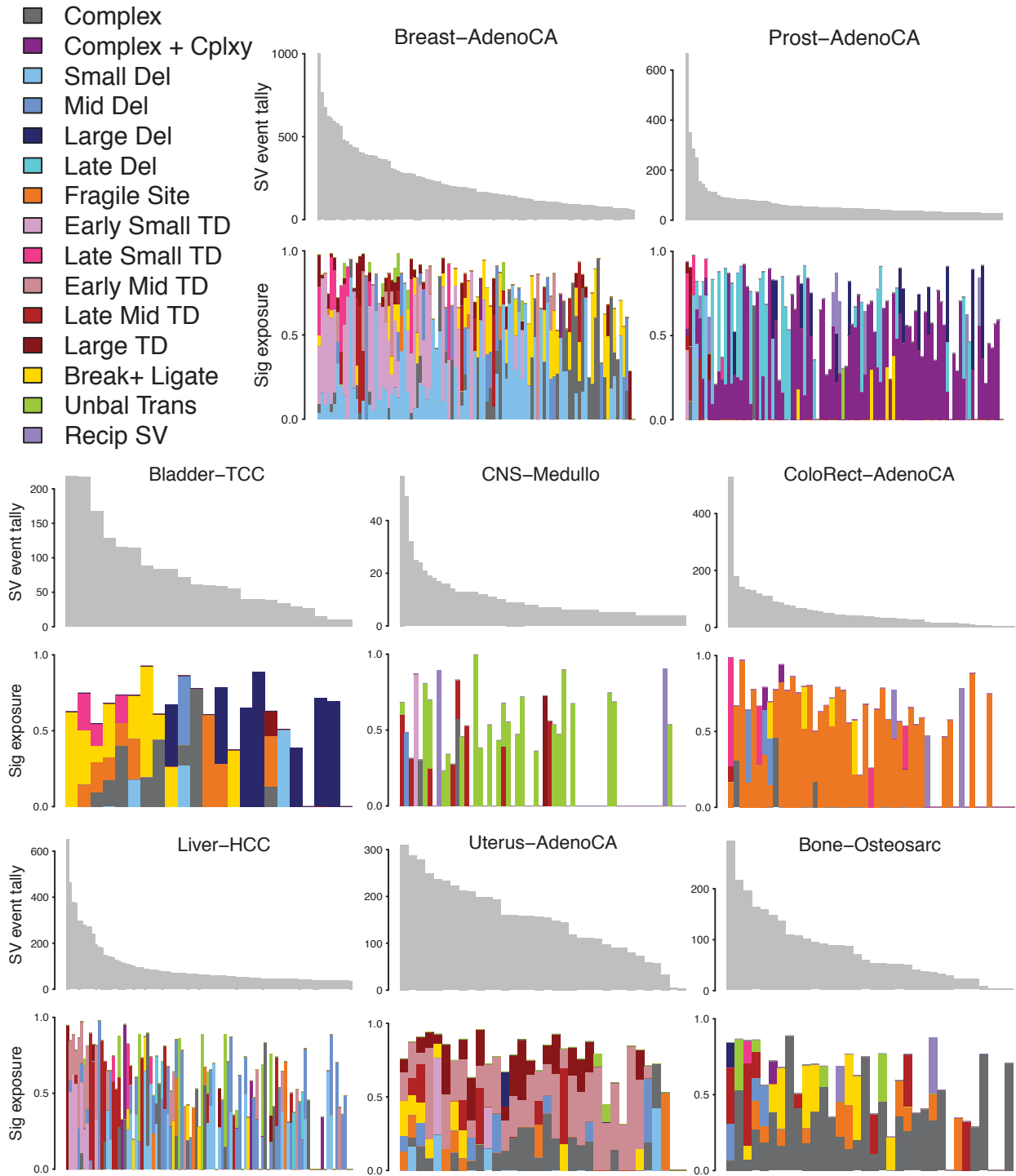


Figure 4.21: Average estimated sample exposures to HDP-extracted SV signatures (Figure 4.20) for eight of the PCAWG cancer types. Large cohorts are subset to a maximum of 100 samples for presentation. For each sample, mean estimated exposures are only plotted for significant signatures (95% credibility interval above zero), leaving a blank proportion with uncertain allocation to the same or different signatures.

(Nik-Zainal et al., 2016; Hillman et al., 2018), this analysis of 76 SV categories—including many novel structures—tallied across 2050 samples in a pan-cancer cohort is the most detailed and comprehensive summary of rearrangement signatures yet produced. However, the scope for further improvement is vast. My HDP method currently requires genome alterations to be tallied in discrete, unordered categories, as do the NMF-based methods reviewed in Section 4.1. When classifying SV events, a large swathe of the complex rearrangement landscape remains intractable to simple categorisation, and is largely precluded from signature analysis. Even if BPJ in complex events like chromothripsis or BFB were to be classified, it remains unclear how these large-scale phenomena should be tallied for meaningful comparison against a simple deletion count, for example. For the SV events that do have classifiable structures, their other pertinent features of size and replication timing are best measured as quantitative variables. My current categorisation of size and timing is a crude substitute for the real value, causing edge effect bias and violating the assumed independence of separate alteration classes. Ideally, future signature analysis methodologies will handle quantitative event features (perhaps using a similar approach to Shiraishi et al. (2015)), and SV signatures may extend to additional features such as microhomology and chromatin state.

4.7 Discussion

In this chapter, I introduced the hierarchical Dirichlet process as a novel strategy for mutational signature decomposition, and derived a set of fifteen somatic rearrangement signatures with unprecedented detail and scale.

The HDP model was first developed by Teh et al. (2006) for topic modelling in corpora, but it is also well-suited to mixed-membership cluster problems in biology, such as the mutational signature decompositions explored in this thesis. The flexible tree of hierarchical DP nodes provides a natural framework for grouping samples by any number of pertinent factors, such as cancer type, germline genotype, mutagen exposure, or patient of origin (if multiple metastases or subclones are available from the same individual). This consideration of sample relatedness empowers the clustering procedure to borrow information across disparate groups, while upholding the prior expectation of differences between groups. In contrast, most other methods perform siloed signature extraction in separate cancer types, with a post hoc consolidation to match results across groups. As the MCMC posterior sampling method naturally gener-

ates credibility intervals for every signature and sample exposure estimate, HDP quantifies significant differences between samples and groups with a justifiable comparison often lacking in other methods. Two further advantages of the HDP approach stem from its nonparametric Bayesian assumption of infinitely many generative processes. First, this enables the number of underlying signatures to be automatically determined from the complexity of the data itself so that—unlike many other methods—HDP does not need to separately assess all plausible signature numbers to find the optimal fit. Second, HDP easily conditions on a prior set of known signatures while simultaneously finding novel clusters in a new dataset. This property is particularly important for small and/or heterogeneous cancer cohorts which might be underpowered for completely *de novo* signature extraction, but which nevertheless contain some number of previously undescribed signatures (particularly artefacts).

One of the major downsides to my current `hdp` R package implementation is that runtime and memory both scale at a roughly linear rate with the number of observed mutations. For the dataset of about 5 million SNVs analysed in Section 4.4, every 1000 MCMC iterations required approximately 3 CPU hours. Although this speed was sufficient to complete analysis in under a week (human time) with parallel computing, the computational expense is prohibitive for larger datasets such as the entire PCAWG SNV catalogue of almost 44 million mutations. For a collection of 10–100 million items, MCMC inference could still be made in separate silos of relevant cancer type groups (as is done for NMF and other methods), *or* an alternative variational inference method (reviewed by Blei et al. (2017)) could approximate the optimal solution in far less time. Several variational inference methods have been proposed for the HDP model, with two available Python packages to support multinomial data (Wang et al., 2011; Hughes et al., 2015).

Another limitation of my current HDP package is the multinomial distribution definition for mutational signatures. By modelling genome alterations as discrete, unordered categories, any quantitative features are forced into crude bins, relationships between similar alteration types are ignored, and the parameter space multiplies with each subdivision for additional features. The requirement for a modest number of separate mutation categories (perhaps less than one thousand) is the major reason most SNV analyses are restricted to 96 classes defined by trinucleotide context, despite the relevance of other signature features such as pentanucleotide context, replication timing, chromatin state, and transcriptional and replication strand bias (Shiraishi et al., 2015; Haradhvala

et al., 2016; Morganella et al., 2016). With the simplifying assumption of independence between features, Shiraishi et al. (2015) proposed a novel approach where each signature is modelled as a collection of distributions over different mutational features. Although currently implemented for categorical features only, the same principle should extend to quantitative variables, and offers an appealing solution for SV alterations defined by many disparate properties such as form, size, microhomology, complexity, and genome topography. In future work, I propose that the nonparametric Bayesian HDP framework—with its many advantages for modelling sample relatedness, conditioning on prior knowledge, quantifying uncertainty, and learning the signature number—could be extended to signature clustering based on sets of independent multinomial and Gaussian distributions, and thus combine the best aspects of both strategies.

Other directions for future improvement include: adoption of formal convergence diagnostics to assess MCMC chains; refinement of my post-processing signature extraction procedure (Section 4.2.2); and incorporation of other topic modelling developments to explicitly account for correlation between topics/signatures (Blei and Lafferty, 2007; Kim and Sudderth, 2011) and/or impose sparsity constraints (Wang and Blei, 2009; Williamson et al., 2010).

