8. Appendices

8.1 Variant clustering with the n-dimensional hierarchical Dirichlet process

To cluster the VAFs across multiple samples from the same patient, an n-dimensional hierarchical Dirichlet process was employed. This algorithm was written and designed by Peter Campbell. Much like the Bayesian Dirichlet process used previously, this method requires the number of variant reads per sample as well as the sequencing depth (Nik-Zainal, Van Loo, et al., 2012). There are no prior definitions of how many mutations are in each cluster, and each mutation can be allocated to any cluster. An upper limit of 100 clusters was set; however the number of clusters generated here was far less than that. Diploid cells are assumed. Each cluster takes on a location within an n-dimensional VAF hypercube but the exact location of this is unknown until the completion of the process. Both the distribution of clone sizes and number of variants per clone are modelled as a Dirichlet process, in a hierarchical Bayesian model with the variant reads and total sequencing depth.

The total read depth for mutation *i* in sample *j* is defined as*:* $n_{i,j}, i = 1, \ldots, N, j = 1, \ldots, M$ where *N* is the number of somatic mutations across all *M* samples. The reference allele, $y_{i,j}$, is similarly defined using the number of reads supporting the reference sequence. The distribution of the reference allele approximates a binomial distribution of the mutation total read depth and the expected proportion of reads supporting the reference allele ($\pi_{i,j}$). As such, $y_{i,j} \sim \mathrm{Bin}(n_{i,j}, \pi_{i,j})$ where $\pi_{i,j}$ is a Dirichlet process:

$$\pi_i \sim DP(\alpha P_0) \in [0,1]^M$$

P is defined as:

$$P = \sum_{h=1}^{\infty} \omega_h \delta_{\pi_h}$$

In this case, $\pi_h \sim P_0$, where δ_π is the point mass at π and ω_h is the weight of the *h*th mutation cluster. As this is a stick-breaking representation of the Dirichlet process, ω_h is defined as:

$$\omega_h = V_h \Pi_{l<h}(1 - V_l)$$

The beta distribution is then used to estimate V_h as $V_h \sim Beta(1 - \alpha)$. As priors, $P_0 \sim U(0,1)^M$ and $\alpha \sim \Gamma(0.01, 0.01)$. The posterior distribution of the Dirichlet process is then modelled using the Gibbs sampler. This involves the sequential sampling of each parameter in the joint distribution and the extraction of a univariate conditional distribution, based on the previous sampling of all other parameters (Hines, 2015). In this case, each mutation is assigned a cluster and stick-breaking weights are adjusted based on the conditional conjugate beta posterior distributions. Draws from the posterior distribution of $(\pi_h|-)$ are then used to update the cluster positions in the *M*-dimensional VAF hypercube.

In large polyclonal samples, these clusters tend to be on the edges of the VAF hypercube and as such, a large region of low probability becomes apparent in the posterior distribution. This aspect of the posterior distribution can then go unsampled due to the low probability and the Gibbs sampling can then be limited. To counter this, a merge-split step is performed after each iteration of the Gibbs sampler using the Metropolis-Hastings proposal for conjugate distributions (Dahl, 2003). Briefly, this involves the random sampling of two mutations and if they are in different clusters, merging the two variants is considered based on the beta-binomial distribution of mutations already allocated. Should the two random variants be in the same clusters, a split step is then considered in a similar fashion (Dahl, 2003). Each merge-split option produces a Metropolis-Hastings ratio and the split or merge is accepted with this probability (Hastings, 1970). The posterior distribution for α can then be refined using this clustering.

The Gibbs sampler was run for 15,000 iterations and the first 10,000 were discarded. The R package label.switching (v1.7, https://CRAN.R-project.org/package=label.switching) (Papastamoulis, 2016) was integrated into the process in order to resolve the label switching problem associated with Markov Chain Monte Carlo outputs, through the use of an Equivalence Classes Representatives (ECR) algorithm (Papastamoulis & Iliopoulos, 2010).

8.2 Copy number analysis revealed no detectable gains or losses across samples

The mean ploidy was 1.97 with ASCAT and 2 with Battenberg (Nik-Zainal, Van Loo, et al., 2012; Raine et al., 2016; Van Loo et al., 2010). This fits with a normal sample containing no significant copy number changes. The ASCAT and Battenberg plot for sample PD37726d_lo0055 can be seen in Figure S1.

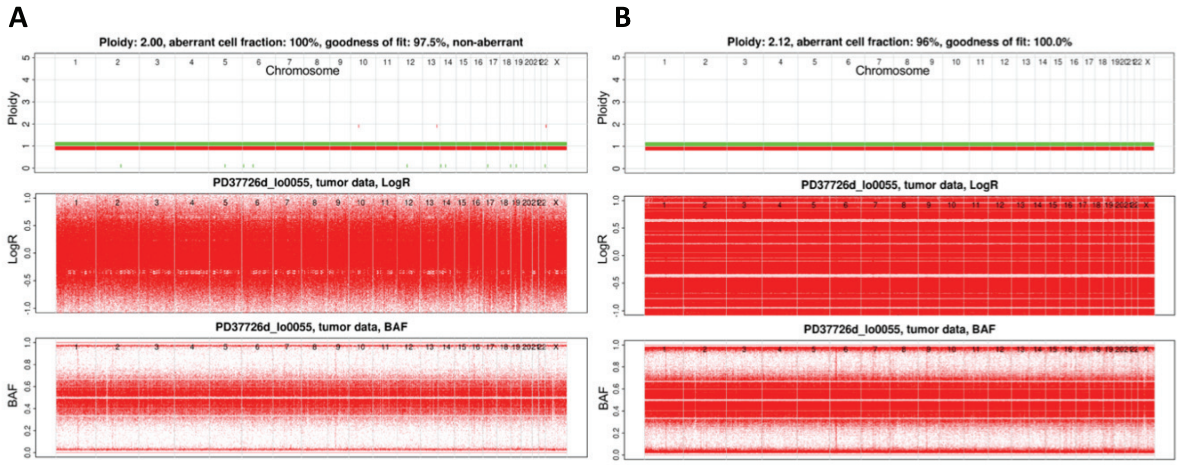


Figure S1 - The copy number analysis of sample PD37726d_lo0055

The x-axes in all six plots is the genomic position, split up by chromosome number. LogR is the log of the coverage at each variant site, compared to the reference sample. The BAF is the B-allele frequency with the B-allele being the mutant variant.

(A) The plots produced by ASCAT shows a clear ploidy of 2, with very few deviations. The top plot summarises this with the few individual red and green marks above and below the value y=1. The LogR and BAF plots show no significant changes in allele frequency or deviations in coverage.

(B) The plot produced by Battenberg supports the ASCAT result, with a ploidy of 2.12 and a goodness of it equal to 100%. The top plot summarises this with the few individual red and green marks above and below the value y=1. The LogR and BAF plots show no significant changes in allele frequency or deviations in coverage.

8.3 Additional support for the mutation calling filters

To validate these three filters, the variants that were filtered out in the unmatched data were analysed. These variants include those with a mean VAF≥0.4, a low over-dispersion parameter (ρ-value) or poor coverage at the variant site. This does not include any variants filtered out by manual inspection of the variants. Per sample, the range of mutations removed was 92,603 to 94,435 and the mean number of removed variants per sample was 94,204 (Figure S2).

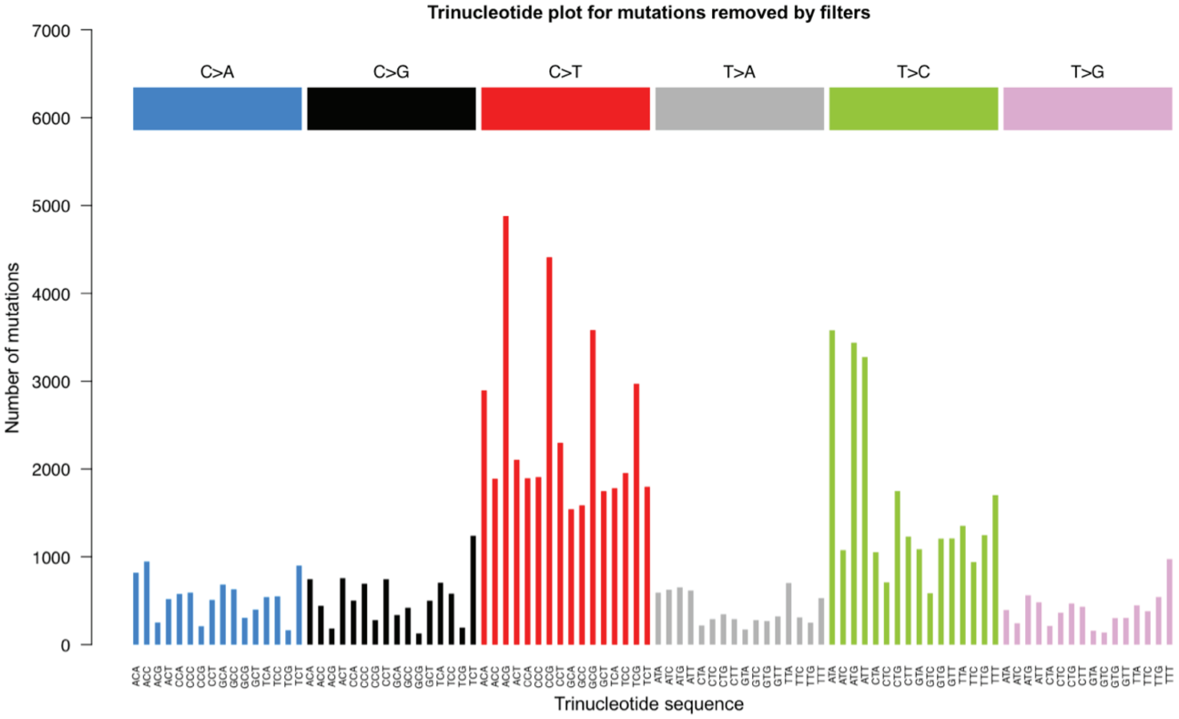


Figure S2 – The entire set of variants filtered out from the unmatched data

The 96-trinucleotide bar plot shows the unique variants removed from the unmatched data. There is a clear dominance of C>T mutations, followed by T>C mutations.

Many of the mutations included in this filtered out cohort would be expected to be germline SNVs. These variants take on high VAFs and are present throughout all samples. Artefacts however would be expected to have a lower VAF and be less common amongst the samples. The efficient removal of artefacts, through the use of the filters described, appears to be supported when subsetting these mutations to those with a VAF less than 0.1 (Figure S3).

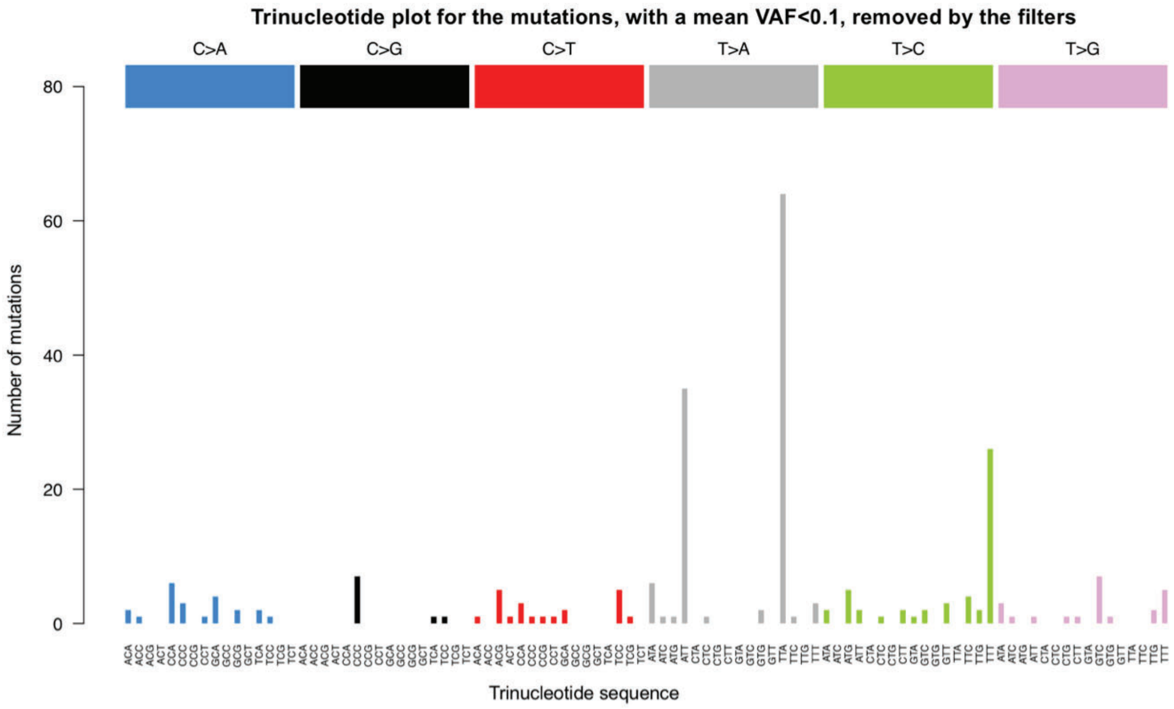


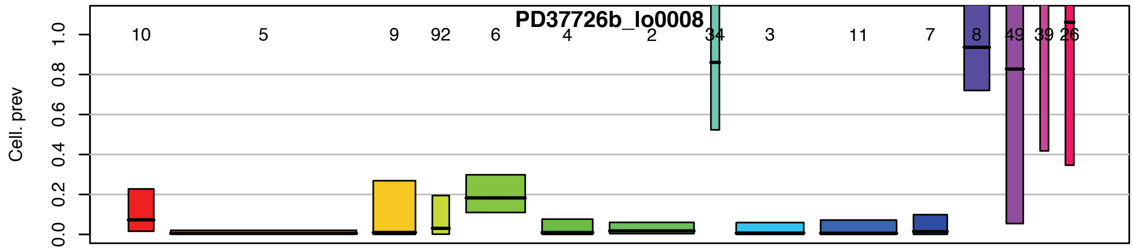
Figure S3 – The low VAF variants that were filtered out

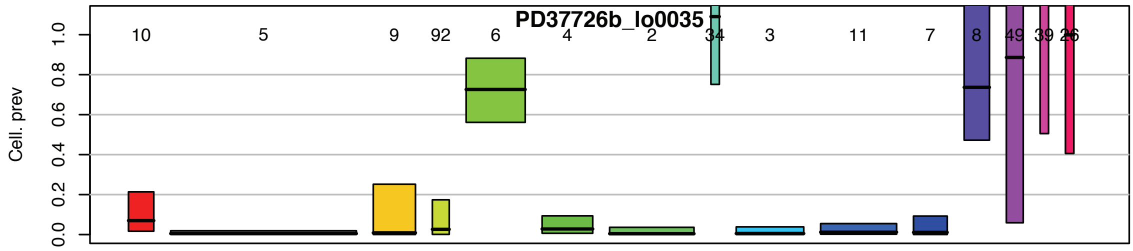
From the 94,548 variants filtered out of the unmatched data, 497 had a mean VAF<0.1. The majority of these are T>A mutations and resemble known sequencing artefacts.

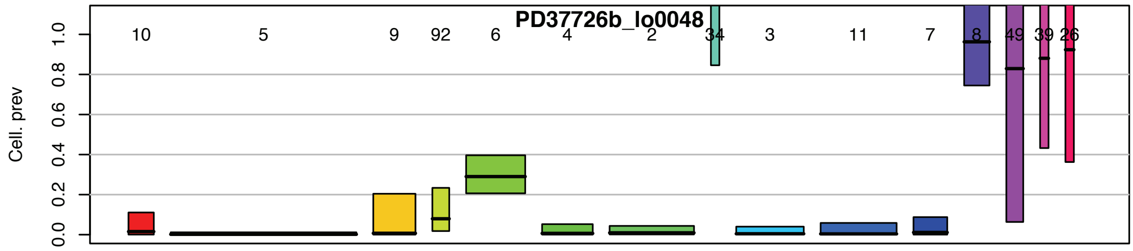
The 497 variants with a VAF less than 0.1 are made up mostly of T>A mutations, a transversion likely generated by the fragmentase enzyme mix (New England Biolabs), during whole-genome library preparation. Comparing Figure S3 to Figures 11 and 12, there is a clear removal of these T>A mutations through the use of the beta-binomial distribution and a depth filter. It appears therefore, that these two filters have proved useful in reducing sequencing artefacts.

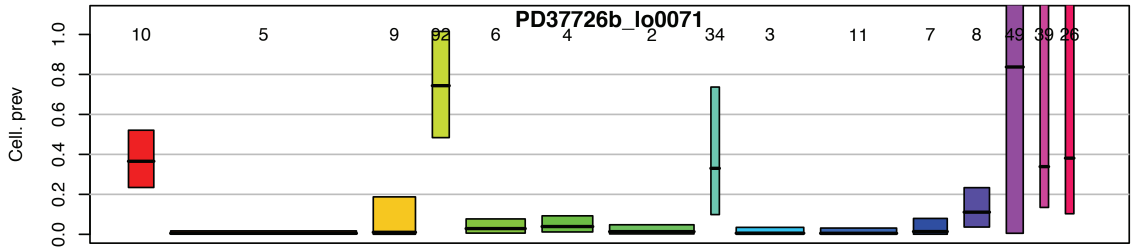
8.4 The pigeonhole principle was applied to each of the 42 samples to reconstruct the phylogenetic tree

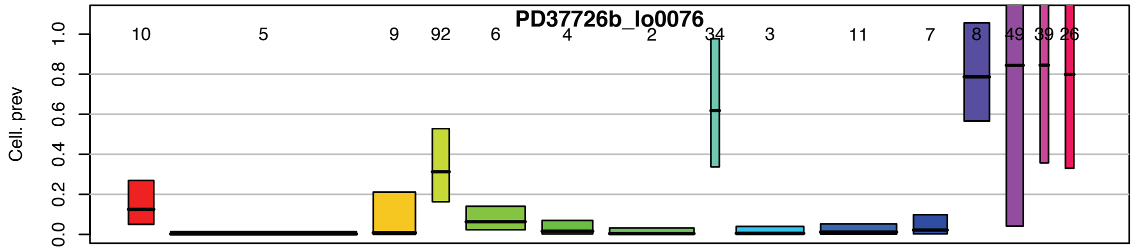
The boxes that were generated from the n-HDP clustering algorithm for all 42 samples are shown in Figure S4. These figures were generated in collaboration with Federico Abascal.

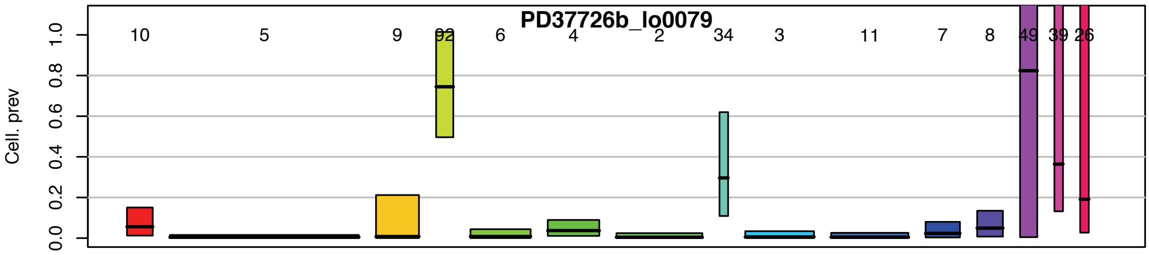

PD37726b_lo0008
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0

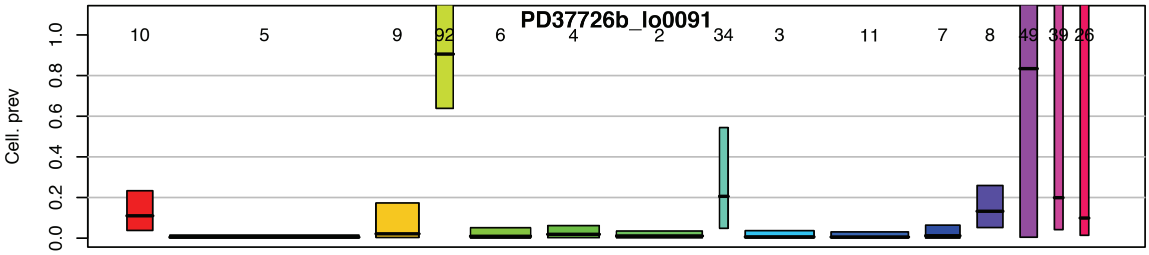

PD37726b_lo0035
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0

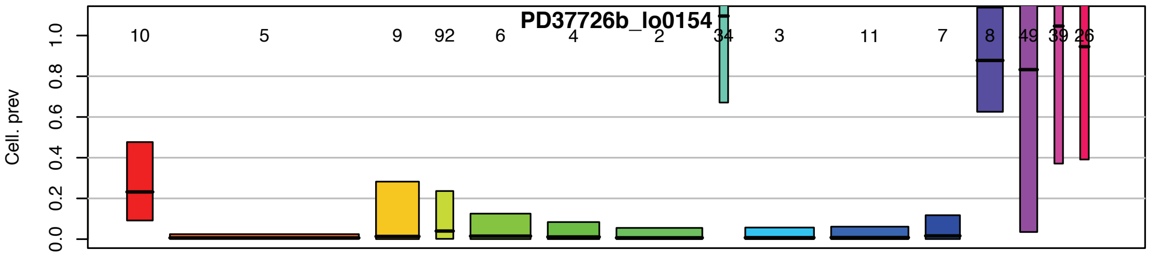

PD37726b_lo0048
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0

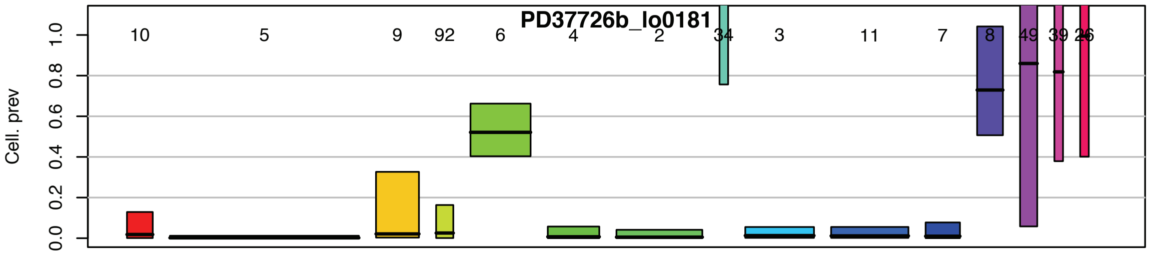

PD37726b_lo0071
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0

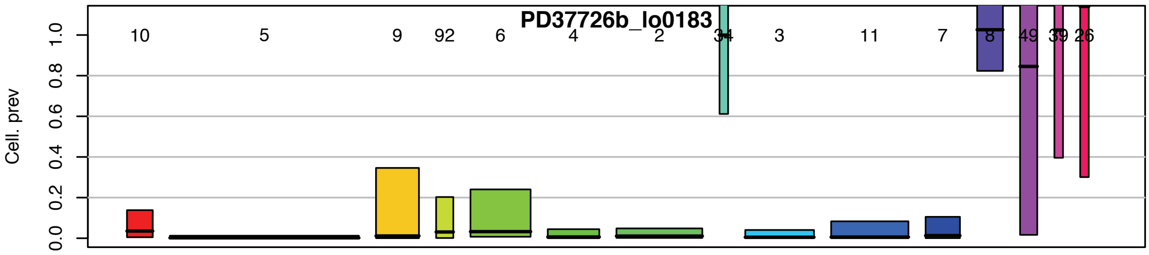

PD37726b_lo0076
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0

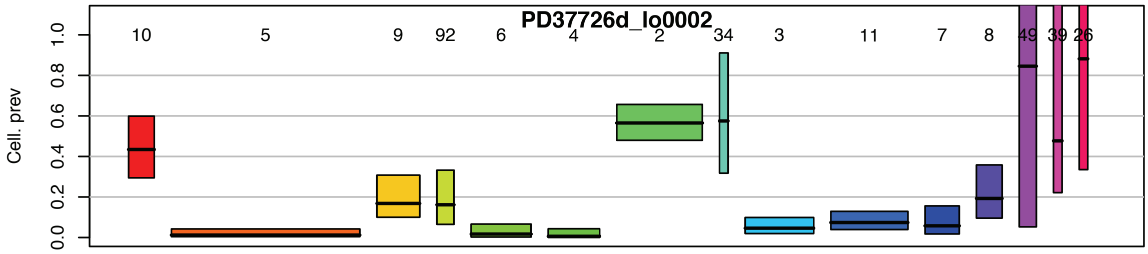

PD37726b_lo0079
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

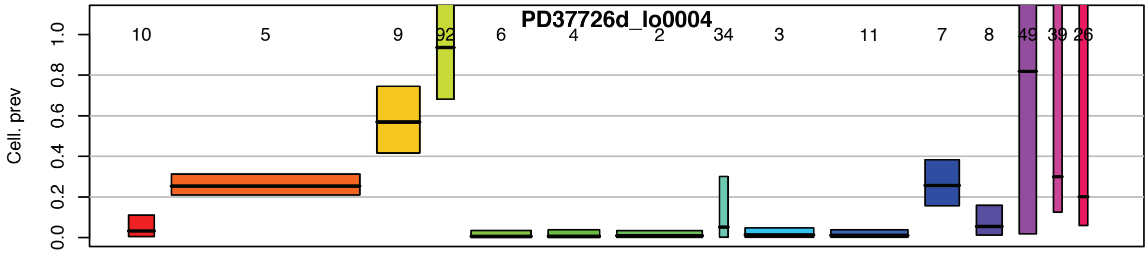

PD37726b_lo0091
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

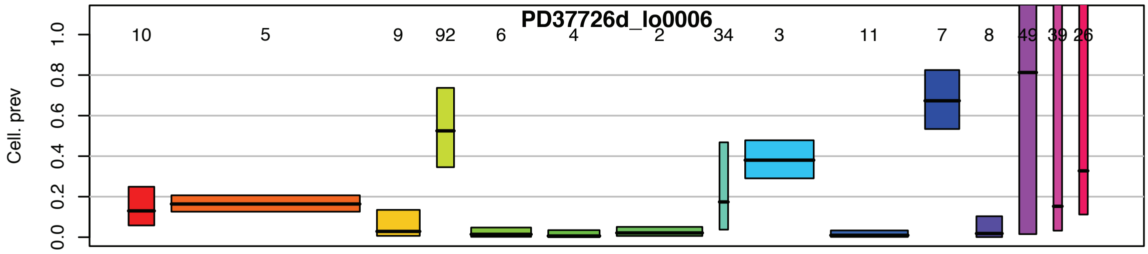

PD37726b_lo0154
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

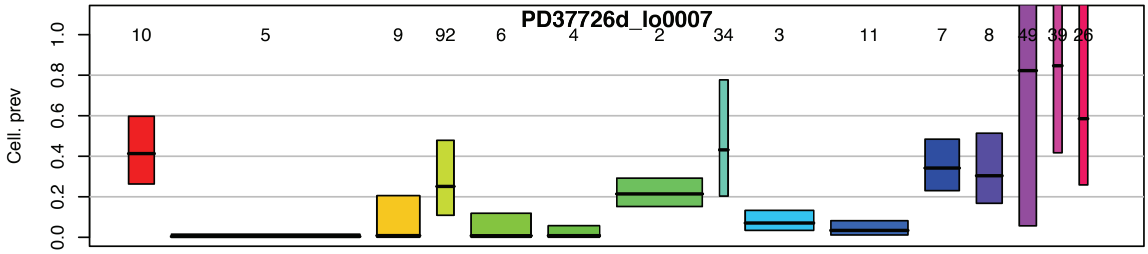

PD37726b_lo0181
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

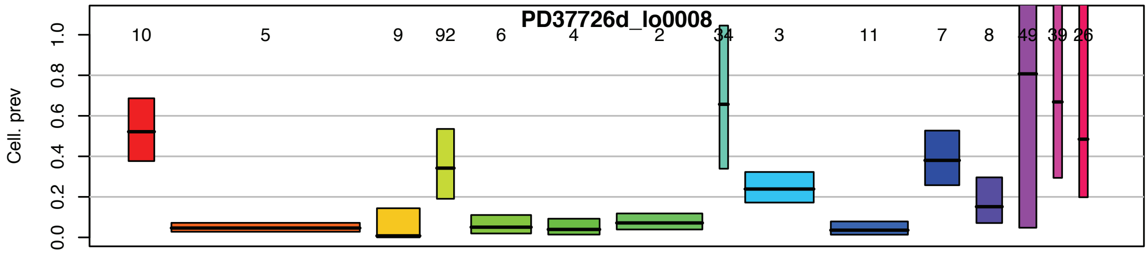

PD37726b_lo0183
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

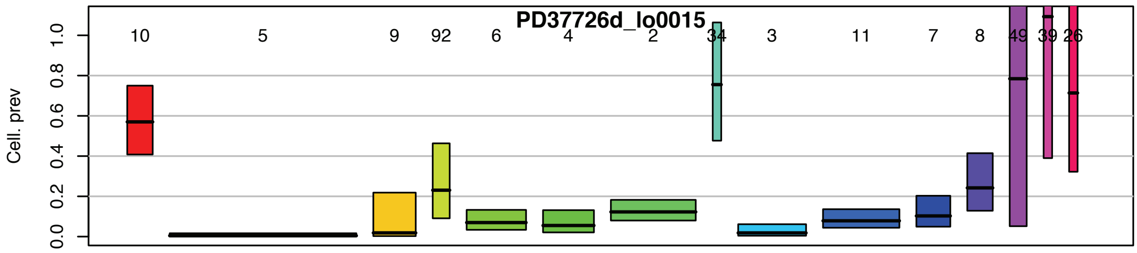

PD37726d_lo0002
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

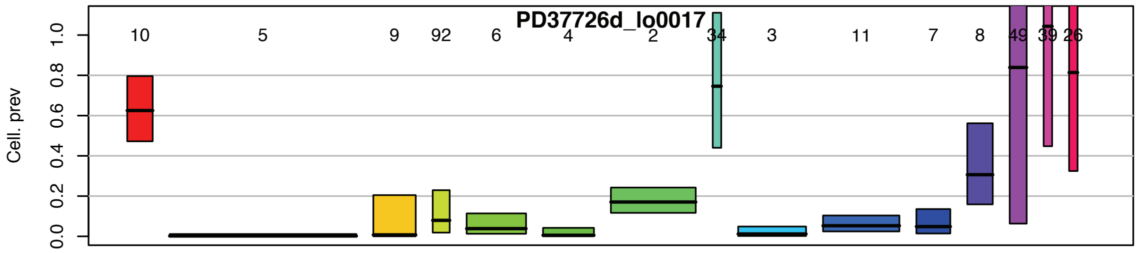

PD37726d_lo0004
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

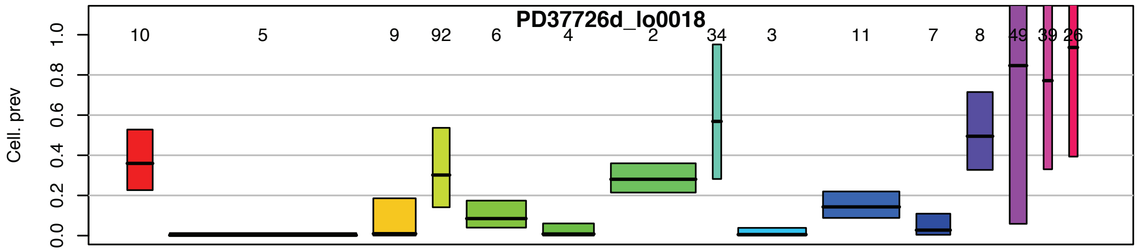

PD37726d_lo0006
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

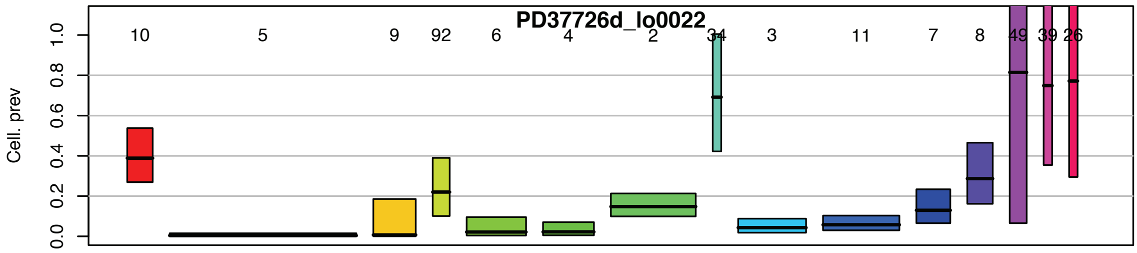

PD37726d_lo0007
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

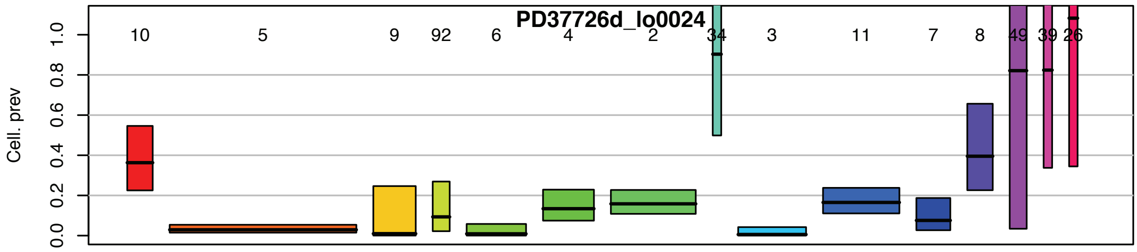

PD37726d_lo0008
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

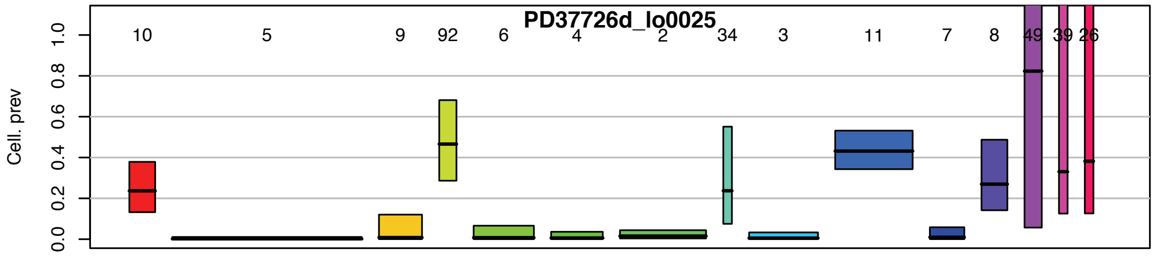

PD37726d_lo0015
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

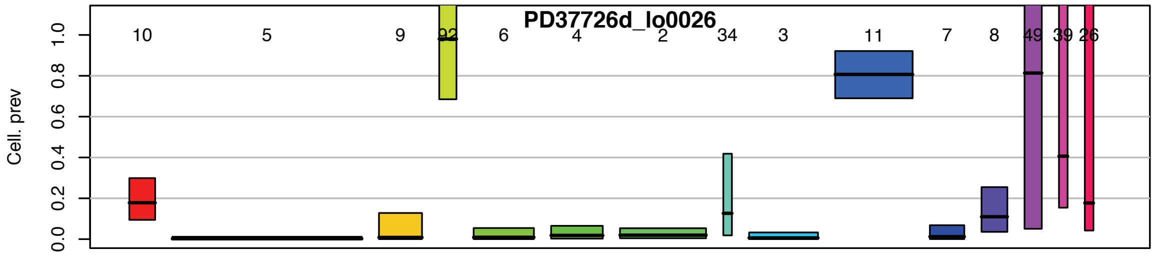

PD37726d_lo0017
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

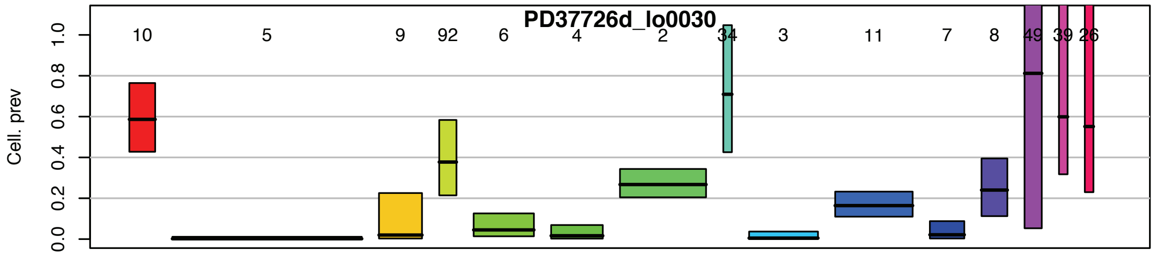

PD37726d_lo0018
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

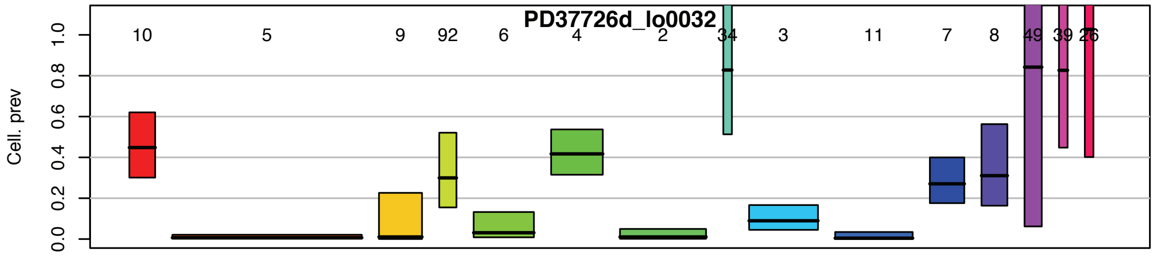

PD37726d_lo0022
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

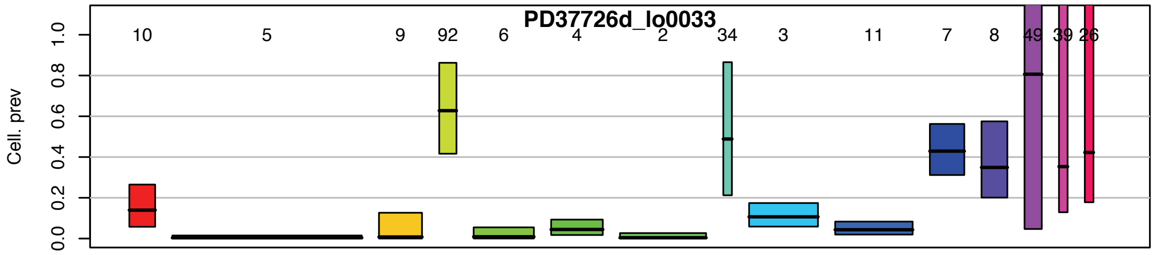

PD37726d_lo0024
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

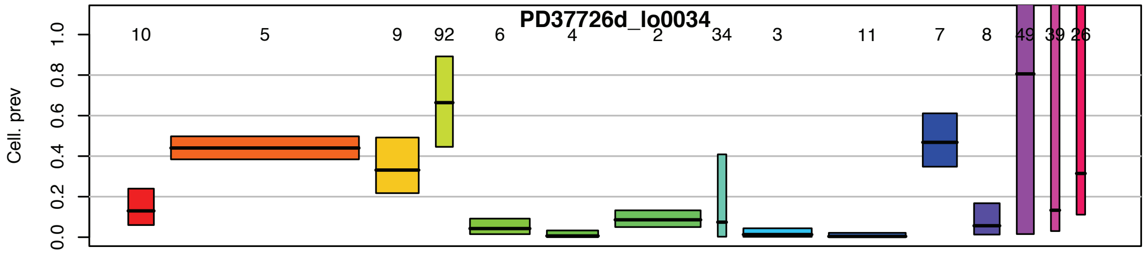

PD37726d_lo0025
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

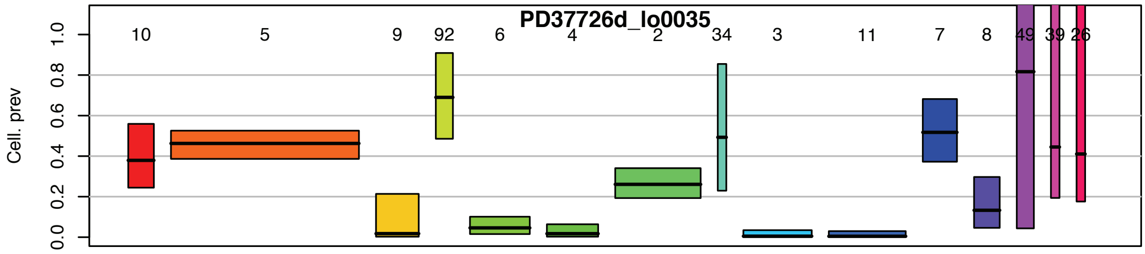

PD37726d_lo0026
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

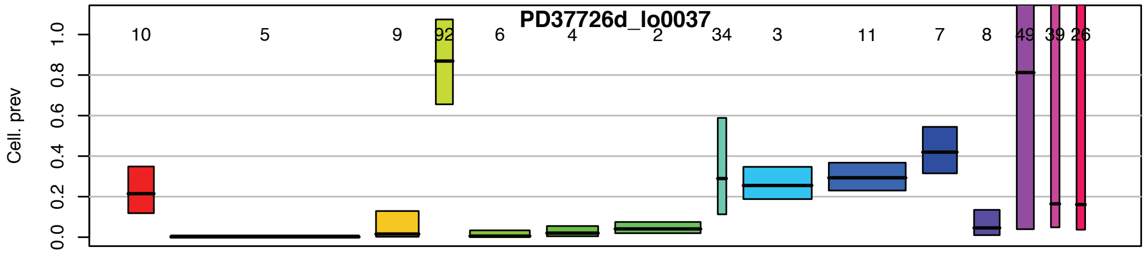

PD37726d_lo0030
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

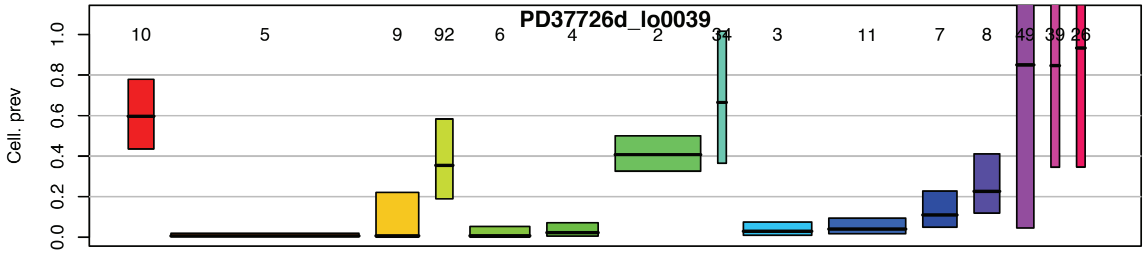

PD37726d_lo0032
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

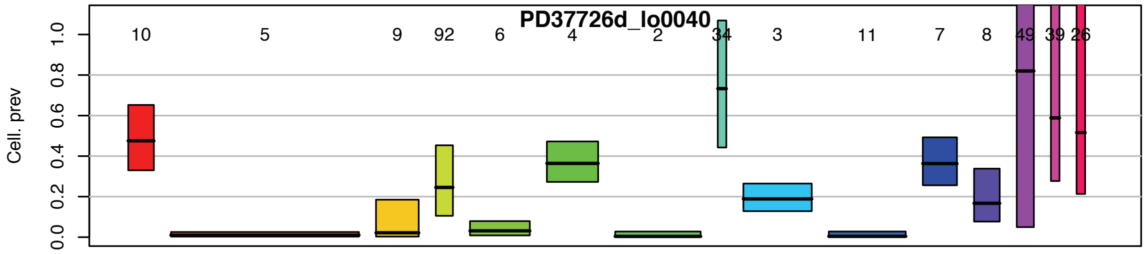

PD37726d_lo0033
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

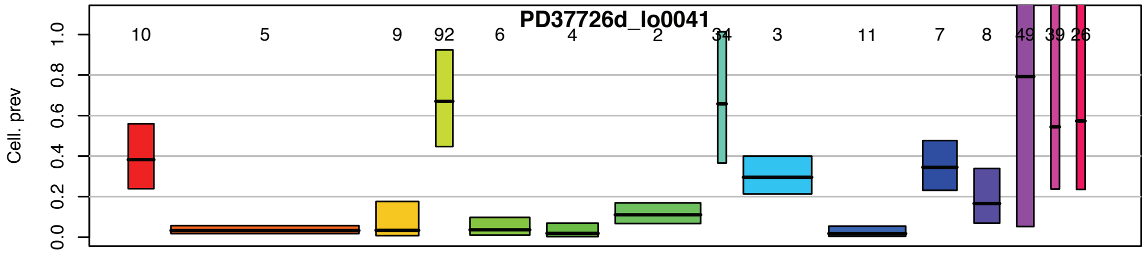

PD37726d_lo0034
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

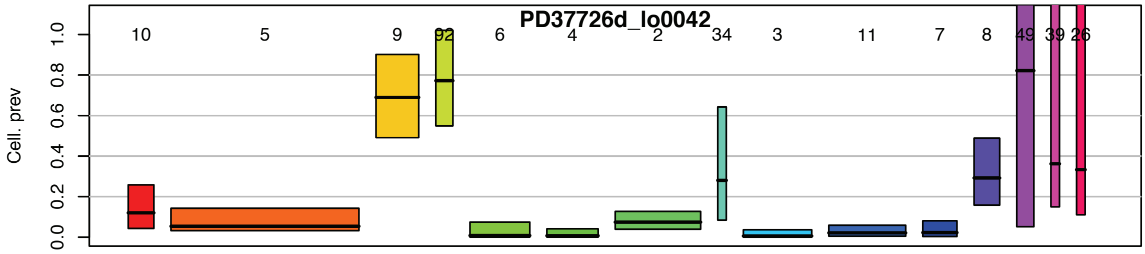

PD37726d_lo0035
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

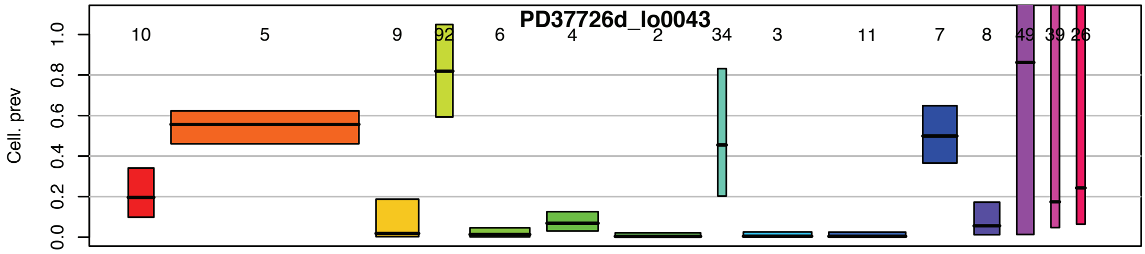

PD37726d_lo0037
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

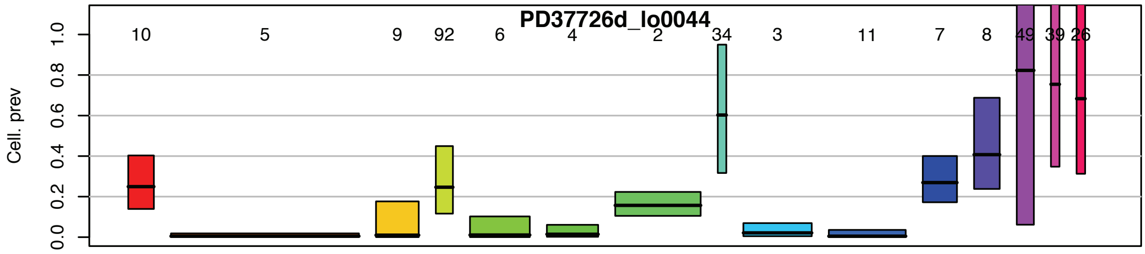

PD37726d_lo0039
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

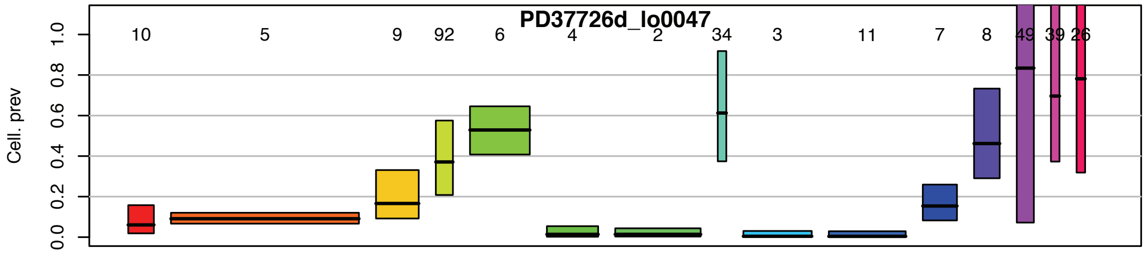

PD37726d_lo0040
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

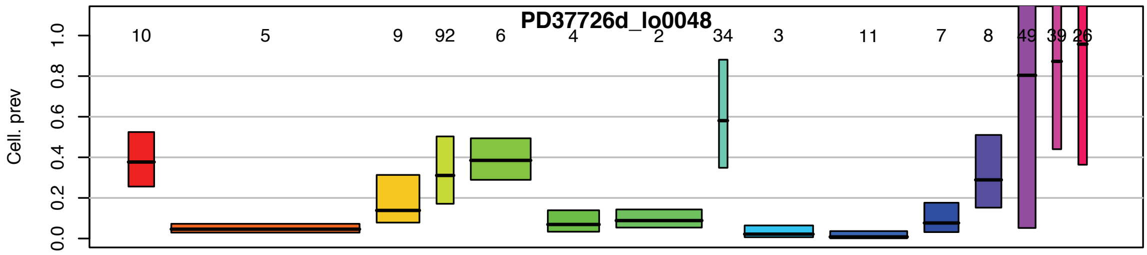

PD37726d_lo0041
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

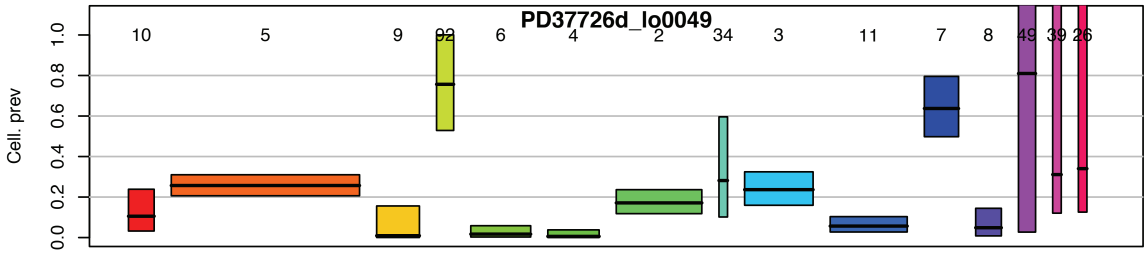

PD37726d_lo0042
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

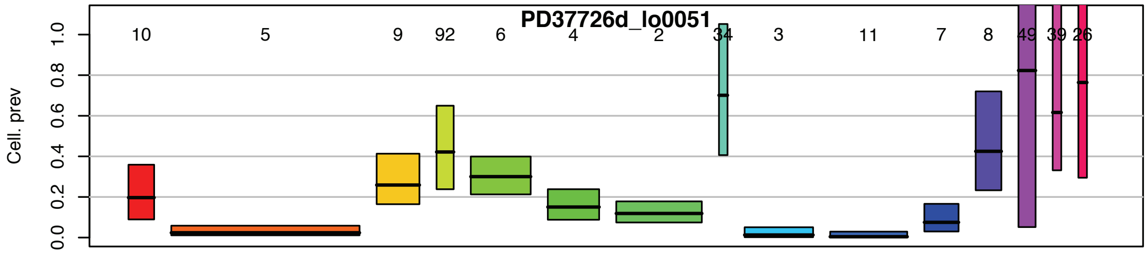

PD37726d_lo0043
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

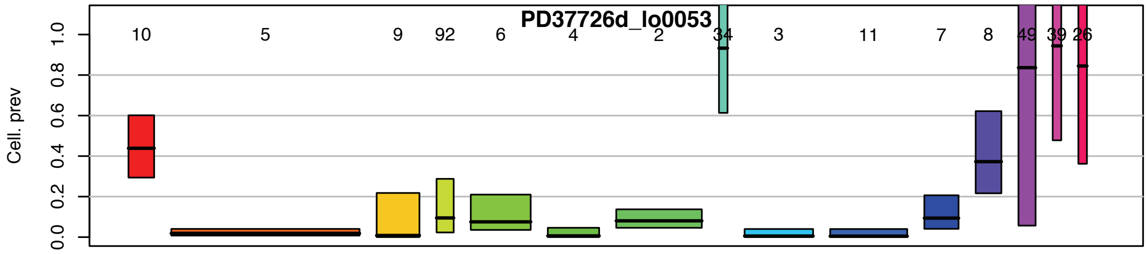

PD37726d_lo0044
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

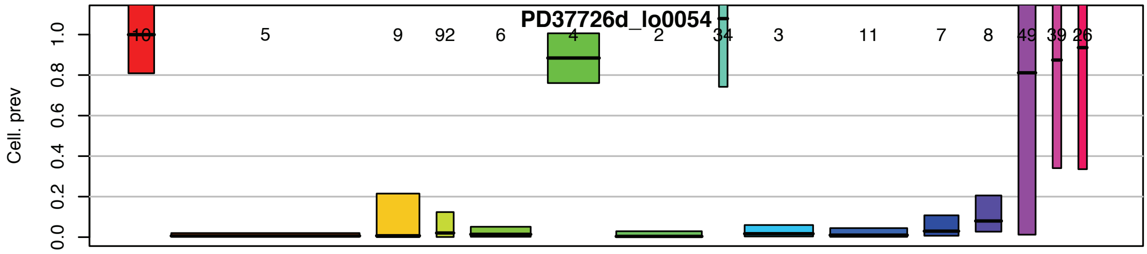

PD37726d_lo0047
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0048
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0049
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0051
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0053
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26


PD37726d_lo0054
Cell. prev
0.0
0.2
0.4
0.6
0.8
1.0
10
5
9
92
6
4
2
34
3
11
7
8
49
39
26

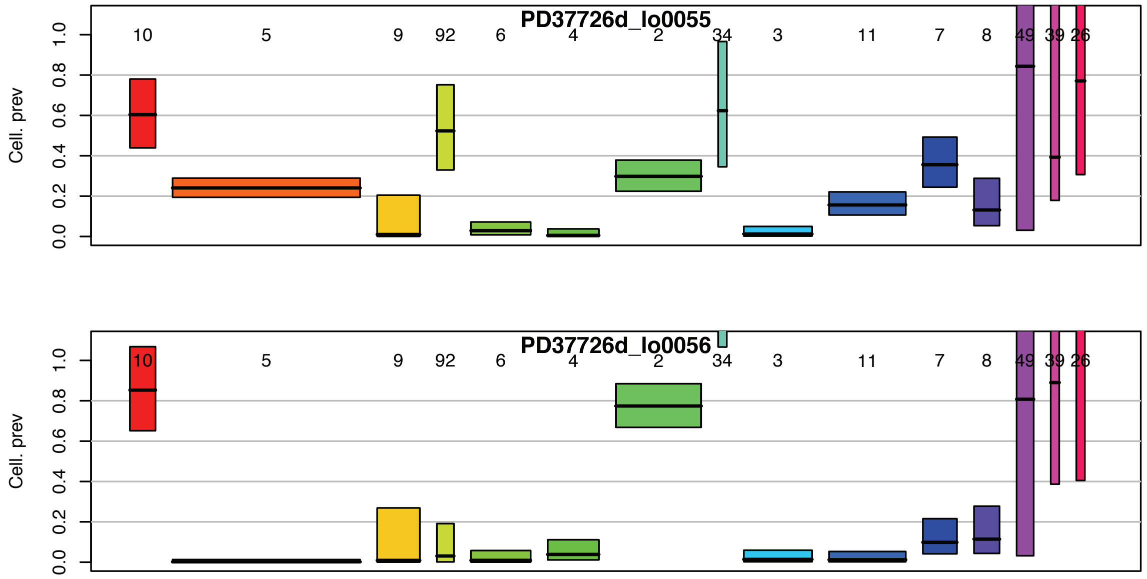


Figure S4 – All 42 boxes used in the phylogenetic tree reconstruction

These boxes represent all 32 islets and ten bladder urothelium samples that underwent clustering with n-HDP. Per sample, clusters (x-axis) and the cell fraction per cluster (y-axis) are shown. The cell fraction is equal to double the VAF. The sample name is at the top with the prefix “PD37726b” representing a bladder urothelium sample and “PD37726d” representing an islet sample. Box width is proportional to the number of mutations while the length is the 95% credible interval. Cluster 49 was discarded as it appeared to be present in all cells.

Each of these boxplots was used to reconstruct the phylogenetic tree, placing the 32 islets and the ten bladder urothelial samples onto the tree. This revealed a shared MRCA and a missing split accounting for three bladder samples (Figures 30 and 31).