Supplementary Material for

‘Single-cell mutation calling and phylogenetic tree reconstruction with loss and recurrence’

Jack Kuipers, Jochen Singer and Niko Beerenwinkel

A Algorithmic details for loss and parallel mutations

A.1 Loss of the wild type allele

Moving beyond the infinite sites assumption, we now consider cases where alleles are lost after mutations occur. The simplest case is when, after a heterozygous mutation, the wild type allele is lost so that the mutation becomes hemizygous. This means that in the tree, we have a heterozygous branch with a hemizygous subtree. To account for this, we need to track the location of the original heterozygous mutation along with the location of the loss (Supplementary Figure S2a), leading to a quadratic (in m) number of possibilities. However, for each location of the original heterozygous mutation, we can compute all the partial sums

$$R_{\mathrm{wl}}(D_i \mid T, x) = \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x) \sum_{\substack{j=1 \\ j \prec x}}^{m} \frac{\tilde{P}_{\mathrm{hm}}(D_i \mid T, j)}{\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, j)} \tag{18}$$

with ‘wl’ standing for *wild type loss*. For the computation, we use the recursion

$$\begin{aligned} R_{\mathrm{wl}}(D_i \mid T,\, x) =& R_{\mathrm{wl}}(D_i \mid T,\, x_{\mathrm{l}})\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{r}}) + \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{l}})R_{\mathrm{wl}}(D_i \mid T,\, x_{\mathrm{r}}) \\ &+ \tilde{P}_{\mathrm{hm}}(D_i \mid T,\, x_{\mathrm{l}})\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{r}})I(\delta(x_{\mathrm{l}}) > 1) + \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{l}})\tilde{P}_{\mathrm{hm}}(D_i \mid T,\, x_{\mathrm{r}})I(\delta(x_{\mathrm{r}}) > 1) \end{aligned} \tag{19}$$

with a single tree traversal. The first line represents the case that the loss does not occur at one of the children of x making one subtree heterozygous and the loss occurring somewhere in the other subtree (Supplementary Figure S2b,c). The next two lines represent the cases that the loss occurs directly at either of the two children of x, so that a hemizygous and heterozygous subtree meet at x itself. In the recursion, we again do not allow a hemizygous mutation above a leaf, so the loss of the wild type allele must affect at least two cells, indicated in the equation by the indicator function I on the number of descendants $\delta(x)$. The recursion starts at 0 for each leaf:

$$R_{\mathrm{wl}}(D_i \mid T,\, j) = 0\,, \quad j = 1, \ldots, m \tag{20}$$

When we sum the different possibilities

$$S_{\mathrm{wl}}(D_i \mid T) = \frac{1}{m_{\mathrm{wl}}} \sum_{\tau_i} R_{\mathrm{wl}}(D_i \mid T,\, \tau_i)P_{\mathrm{wt}}(D_i) \tag{21}$$

we normalise by m_{wl}, the total number of permissible placements of the original mutation and the loss which can be counted with the same recursion as above by keeping track of the non-zero terms. In the depth-first tree traversal from the leaves upwards in Equation (19), for each node the contribution is the combination of the terms of its daughters in the tree, so we compute R_{wl} for all nodes in the tree in time $O(m)$. As each term itself is a sum over $O(m)$ terms from the definition in Equation (18), the tree recursion therefore allows us to sum the quadratic number of placements of the mutation and its loss in the tree is in $O(m)$.

A.2 Loss of the mutated allele

Alternatively, we may lose the mutated allele so that in the tree there is a heterozygous region with a wild type subtree. Again, we track the location of the original mutation

$$R_{\mathrm{ml}}(D_i \mid T,\, x) = \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x) \sum_{\substack{j=1 \\ j \prec x}}^{m} \frac{1}{\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, j)} \tag{22}$$

with ‘ml’ standing for *mutation loss*. This is computed using the recursion

$$\begin{aligned} R_{\mathrm{ml}}(D_i \mid T,\, x) =& R_{\mathrm{ml}}(D_i \mid T,\, x_{\mathrm{l}})\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{r}}) + \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{l}})R_{\mathrm{ml}}(D_i \mid T,\, x_{\mathrm{r}}) \\ &+ \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{r}})I(\delta(x_{\mathrm{l}}) > 1) + \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{l}})I(\delta(x_{\mathrm{r}}) > 1) \end{aligned} \tag{23}$$

where the wild type terms just give a factor of 1 because scores are relative to the wild type case. The definition above allows for the loss of the mutated allele directly after it appears in the tree: when the mutation occurs at node x it is lost at node x_{l} or x_{r}. A more parsimonious explanation, which leads to

the same observed data, is that the mutation was gained directly at x_r or x_l instead. We therefore exclude such scenarios and define

$$R'_\mathrm{ml}(D_i \mid T,\, x) = \tilde{P}_\mathrm{ht}(D_i \mid T,\, x) \sum_{\substack{j=1 \\ j \prec x \\ \mathrm{Pa}(j) \neq x}}^{m} \frac{1}{\tilde{P}_\mathrm{ht}(D_i \mid T,\, j)} \tag{24}$$

which can be computed as

$$R'_\mathrm{ml}(D_i \mid T,\, x) = R_\mathrm{ml}(D_i \mid T,\, x_\mathrm{l})\tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) + \tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l})R_\mathrm{ml}(D_i \mid T,\, x_\mathrm{r}) \tag{25}$$

which involves the unrestricted contributions R_ml of the daughter lineages. We sum the possibilities as

$$S_\mathrm{ml}(D_i \mid T) = \frac{1}{m_\mathrm{ml}} \sum_{\tau_i} R'_\mathrm{ml}(D_i \mid T,\, \tau_i) P_\mathrm{wt}(D_i) \tag{26}$$

where m_ml is the total number of permissible placements of the original mutation and the loss.

A.3 Parallel mutations

For parallel mutations, we track all cases where a mutation occurs in two distinct subtrees below any given node

$$R_\mathrm{pm}(D_i \mid T,\, x) = \sum_{\substack{j=1 \\ j \prec x \\ j \not\prec k}}^{m} \sum_{\substack{k=1 \\ k \prec x \\ k \not\prec j}}^{m} \tilde{P}_\mathrm{ht}(D_i \mid T,\, j)\tilde{P}_\mathrm{ht}(D_i \mid T,\, k) \tag{27}$$

To evaluate these sums, we first compute the partial sum that a heterozygous mutation occurs somewhere in the subtree below node j

$$R_\mathrm{ht}(D_i \mid T,\, x) = \sum_{\substack{j=1 \\ j \prec x}}^{m} \tilde{P}_\mathrm{ht}(D_i \mid T,\, j) \tag{28}$$

which is again computed using a tree traversal

$$R_\mathrm{ht}(D_i \mid T,\, x) = R_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l}) + R_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) + \tilde{P}_\mathrm{ht}(D_i \mid T,\, x) \tag{29}$$

since either the mutation is lower down one branch, or occurs at x itself.

The parallel mutation term is

$$R_\mathrm{pm}(D_i \mid T,\, x) = R_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l})R_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) \tag{30}$$

but this again includes terms with unnecessary complexity. For example, the case where the mutation occurs at both x_l and x_r can be more parsimoniously explained as the mutation occurring once at x instead. Another case is when the mutation occurs at one of the children of x, as well as one of the grandchildren on the other branch. This could be explained as a mutation at x, and then a loss at the non-mutated grandchild. We therefore exclude these possibilities and define

$$\begin{aligned} R'_\mathrm{pm}(D_i \mid T,\, x) =& R_\mathrm{pm}(D_i \mid T,\, x) - \tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l})\tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) \\ & - \tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l})\tilde{P}'_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) - \tilde{P}'_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l})\tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) \end{aligned} \tag{31}$$

with

$$\tilde{P}'_\mathrm{ht}(D_i \mid T,\, x) = \tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{l}) + \tilde{P}_\mathrm{ht}(D_i \mid T,\, x_\mathrm{r}) \tag{32}$$

We sum the possible placements of parallel mutations as

$$S_\mathrm{pm}(D_i \mid T) = \frac{1}{m_\mathrm{pm}} \sum_{\tau_i} R'_\mathrm{pm}(D_i \mid T,\, \tau_i) P_\mathrm{wt}(D_i) \tag{33}$$

where m_pm is the number of permissible placements of both mutations (corresponding to those with a non-zero score in R'). The time complexity is again $O(m)$.

A.4 Posterior probabilities for mutation loss and parallel mutations

In order to compute the posterior probability of the loss of the mutated allele, from the recursions above, all we know are the relative probabilities that the mutation occurred at node x, but we have not recorded where the loss occurred. To remedy this, we also compute $Q_{\mathrm{ml}}(D_i \mid T,\, x)$, the probability that the mutational loss itself occurred at x, and we sum all possibilities that the original mutation occurred at an ancestor of x. To perform the tree traversal from the root down, we define x_{p} as the parent of x and x_{s} as its sibling, so that we can use the recursion

$$\begin{aligned} Q_{\mathrm{ml}}(D_i \mid T,\, x) =& Q_{\mathrm{ml}}(D_i \mid T,\, x_{\mathrm{p}})\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{s}}) \\ &+ \frac{\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{p}})}{\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x)} \end{aligned} \tag{34}$$

where the first term represents shifting the loss from a parent to the child and setting the other sibling to be heterozygous, while the second term is the additional case where the mutation occurs at the parent and is directly lost at one child. These need to be computed for the recursion to function, but excluded from the final set of possibilities as

$$Q'_{\mathrm{ml}}(D_i \mid T,\, x) = Q_{\mathrm{ml}}(D_i \mid T,\, x) - \frac{\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{p}})}{\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x)} \tag{35}$$

The other case is when a cell is below a parallel mutation. To compute this, we track the sum of possibly placements of a parallel mutation on alternative subtrees on the path to the root

$$Q_{\mathrm{ht}}(D_i \mid T,\, x) = Q_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{p}}) + R_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{s}}) \tag{36}$$

We must also remove the restricted cases when the sibling of the parent, or a child of the sibling are mutated

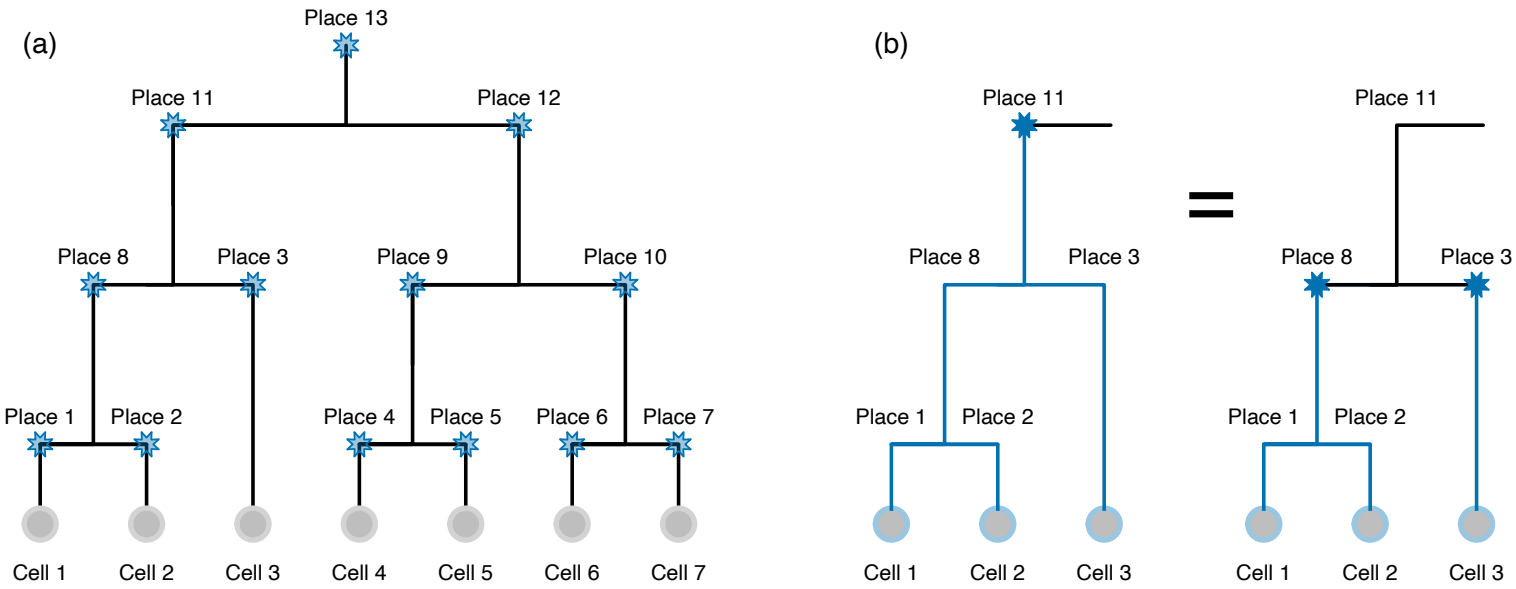
$$\begin{aligned} Q'_{\mathrm{ht}}(D_i \mid T,\, x) =& Q_{\mathrm{ht}}(D_i \mid T,\, x) - \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{p_s}}) \\ &- \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{s_l}}) - \tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x_{\mathrm{s_r}}) \end{aligned} \tag{37}$$

Placing the other mutation at x then gives the contributions

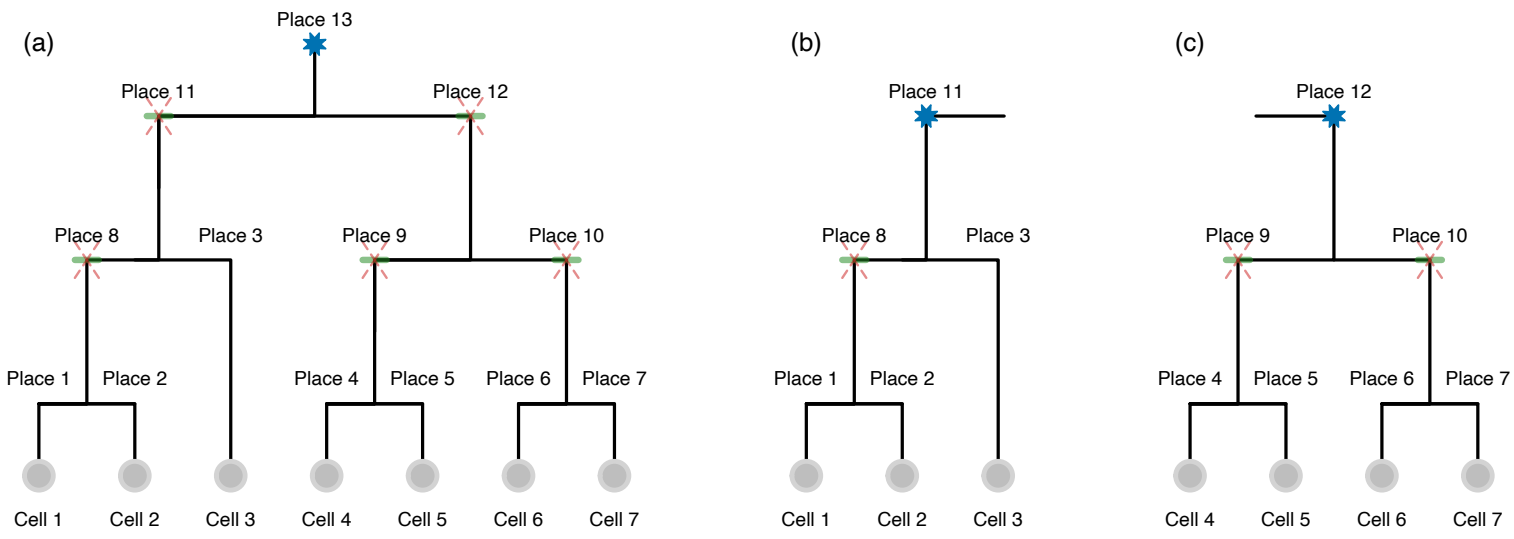
$$Q'_{\mathrm{pm}}(D_i \mid T,\, x) = Q'_{\mathrm{ht}}(D_i \mid T,\, x)\tilde{P}_{\mathrm{ht}}(D_i \mid T,\, x) \tag{38}$$

from which we again normalise and propagate down the tree to obtain the conditional probabilities of a mutation being present in each cell.

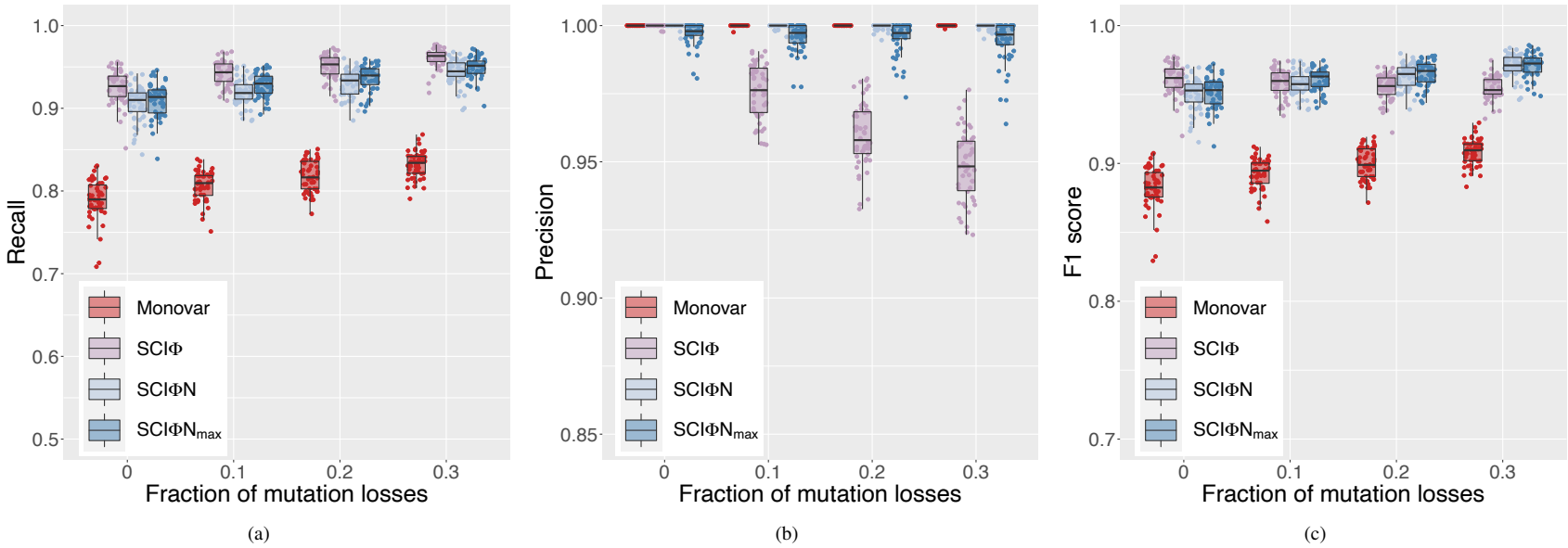
B Supplementary Figures



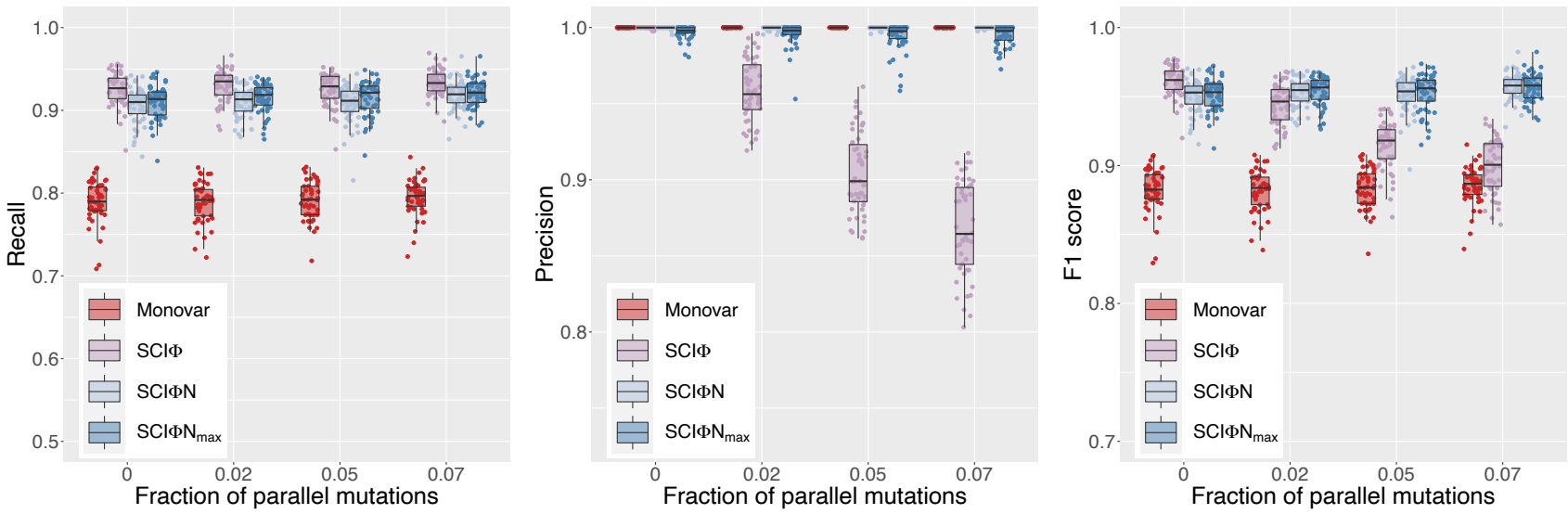
Supplementary Figure S1. Mutations in the cell lineage tree. (a) When a mutation occurs in tumour evolution it affects all descendant cells so the possible placements of the mutation in the tree are at all inner branches. (b) A mutation occurring in the tree (for example at the place labelled 11) has equivalent genotypes at the cells as two mutations occurring along the child branches (3 and 8 in this example). The likelihood contribution of a mutation at a particular branch is therefore computed from the contributions of its children using a tree traversal.



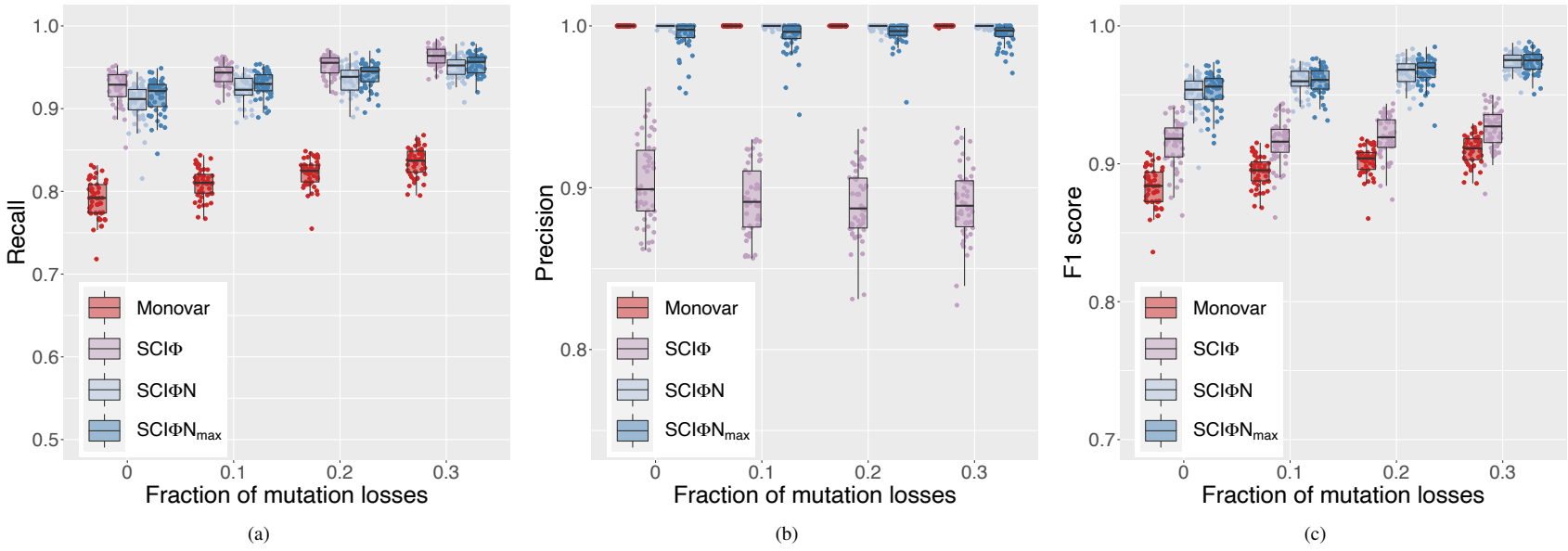
Supplementary Figure S2. Loss in the cell lineage tree. (a) For a given placement of the original mutation (here at the place labelled 13), we wish to count all highlighted possible placements of the loss of the wild type allele (which are above at least two cells). Along with placing the loss directly at a child branch (11 and 12 in this example) the possible placements lower down in the tree are covered by the contributions of the two child subtrees (b) and (c) which have already been computed in the tree traversal.



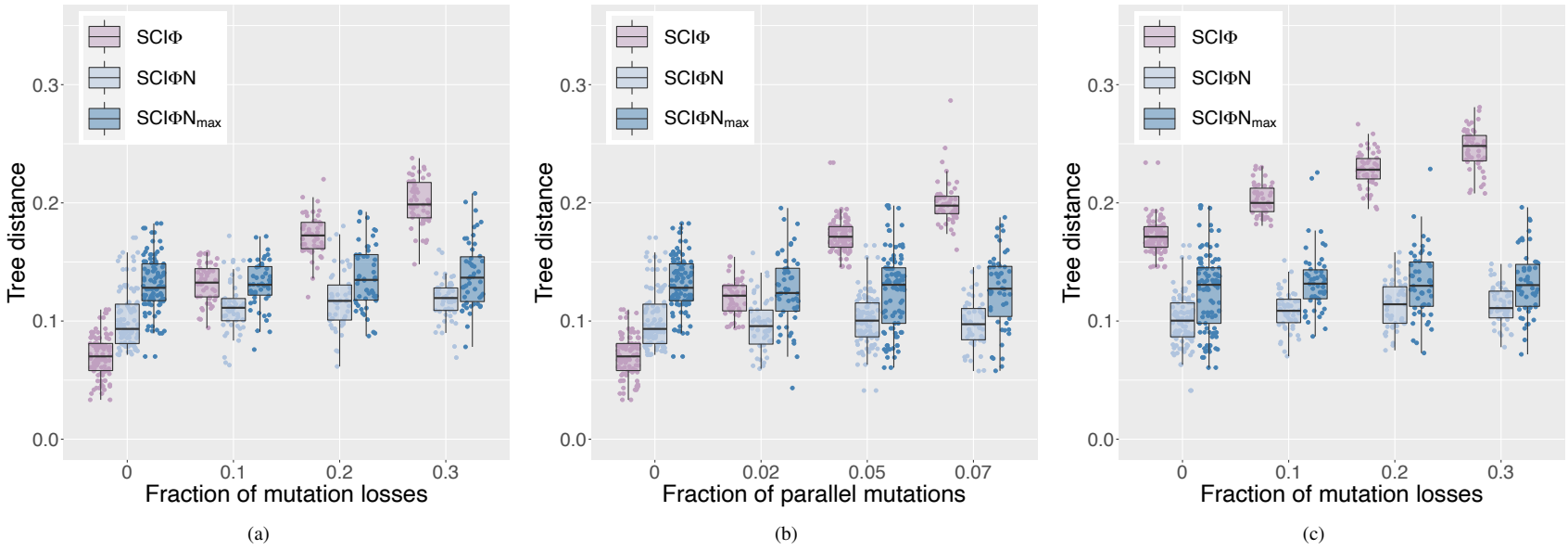
Supplementary Figure S3. Effect of loss on single cell mutation calling.. The fraction of losses is increased (with no parallel mutations) to demonstrate their effect on the recall (a), precision (b) and the F1 score (c).



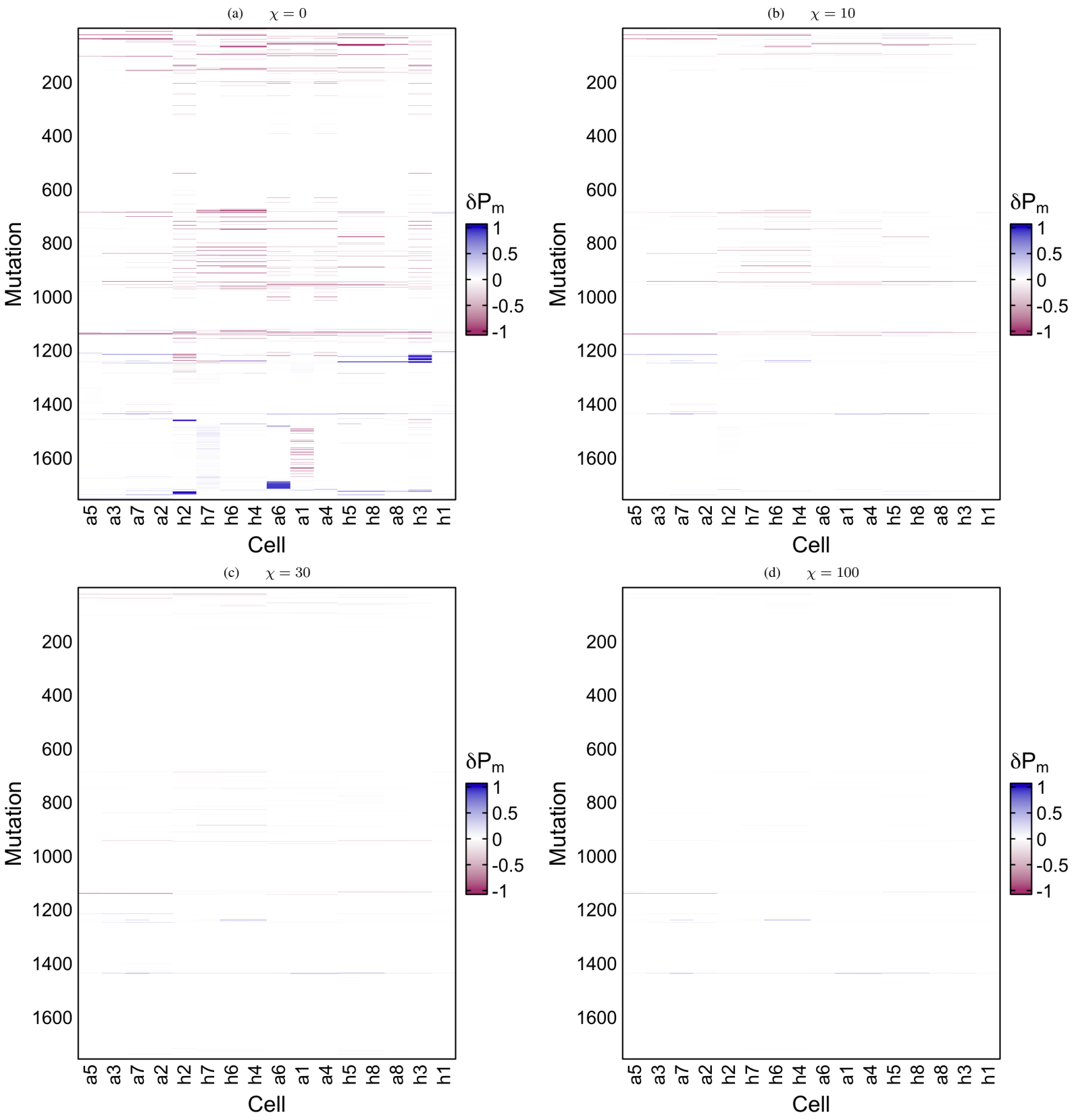
Supplementary Figure S4. Effect of parallel mutations on single cell mutation calling. The fraction of parallel mutations is increased (with no losses) to demonstrate their effect on the recall (a), precision (b) and the F1 score (c).



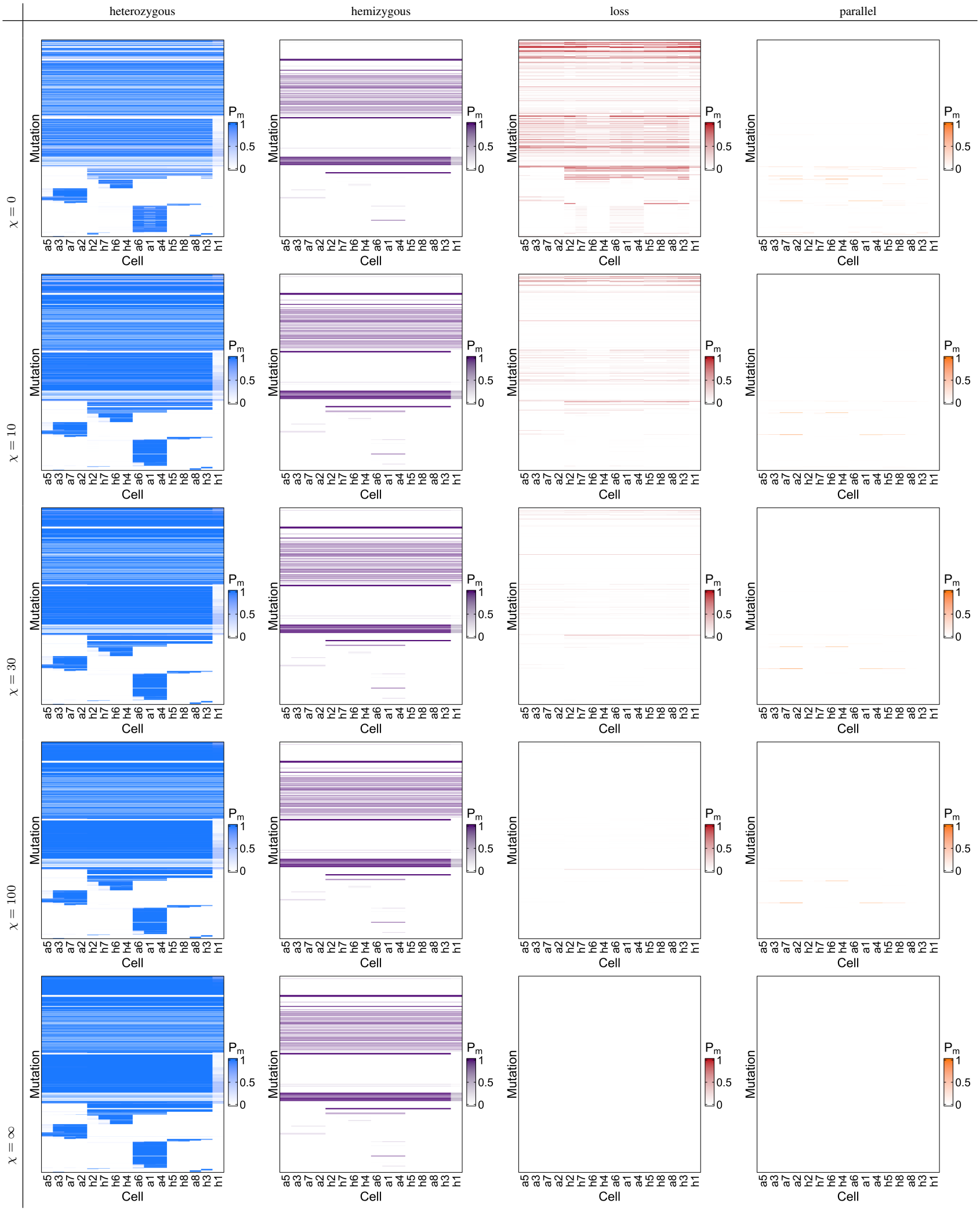
Supplementary Figure S5. Effect of loss and parallel mutations on single cell mutation calling. The fraction of losses is increased (with 5% parallel mutations) to demonstrate their effect on the recall (a), precision (b) and the F1 score (c).



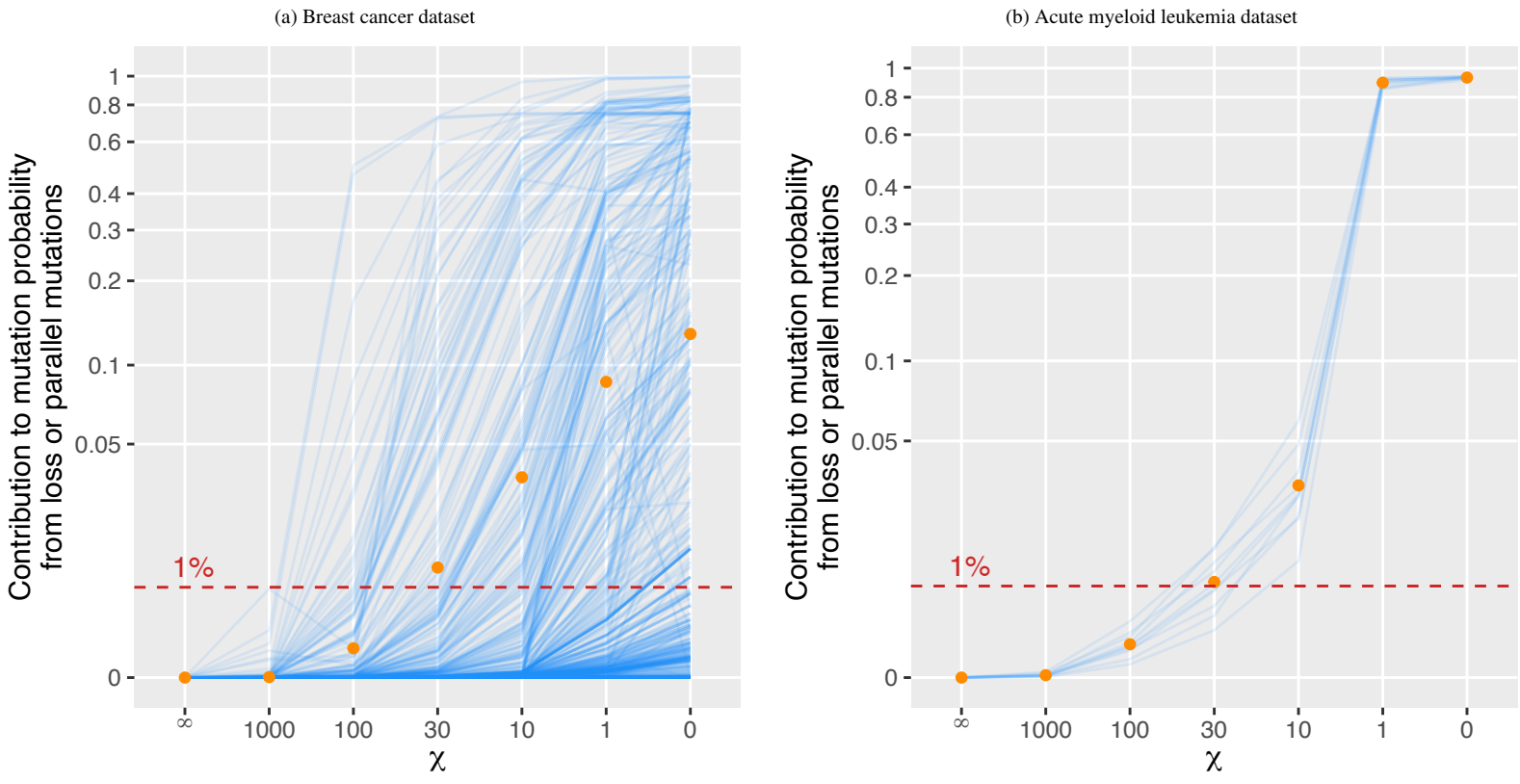
Supplementary Figure S6. Effect of loss and parallel mutations on tree reconstruction. The distance between the inferred and generating tree as the fraction of losses is increased with no parallel mutations (a), as the fraction of parallel mutations is increased with no loss (b) and as the fraction of losses in increased with 5% parallel mutations (c). The simulation under the infinite-sites assumption is the leftmost column of panels (a) and (b).



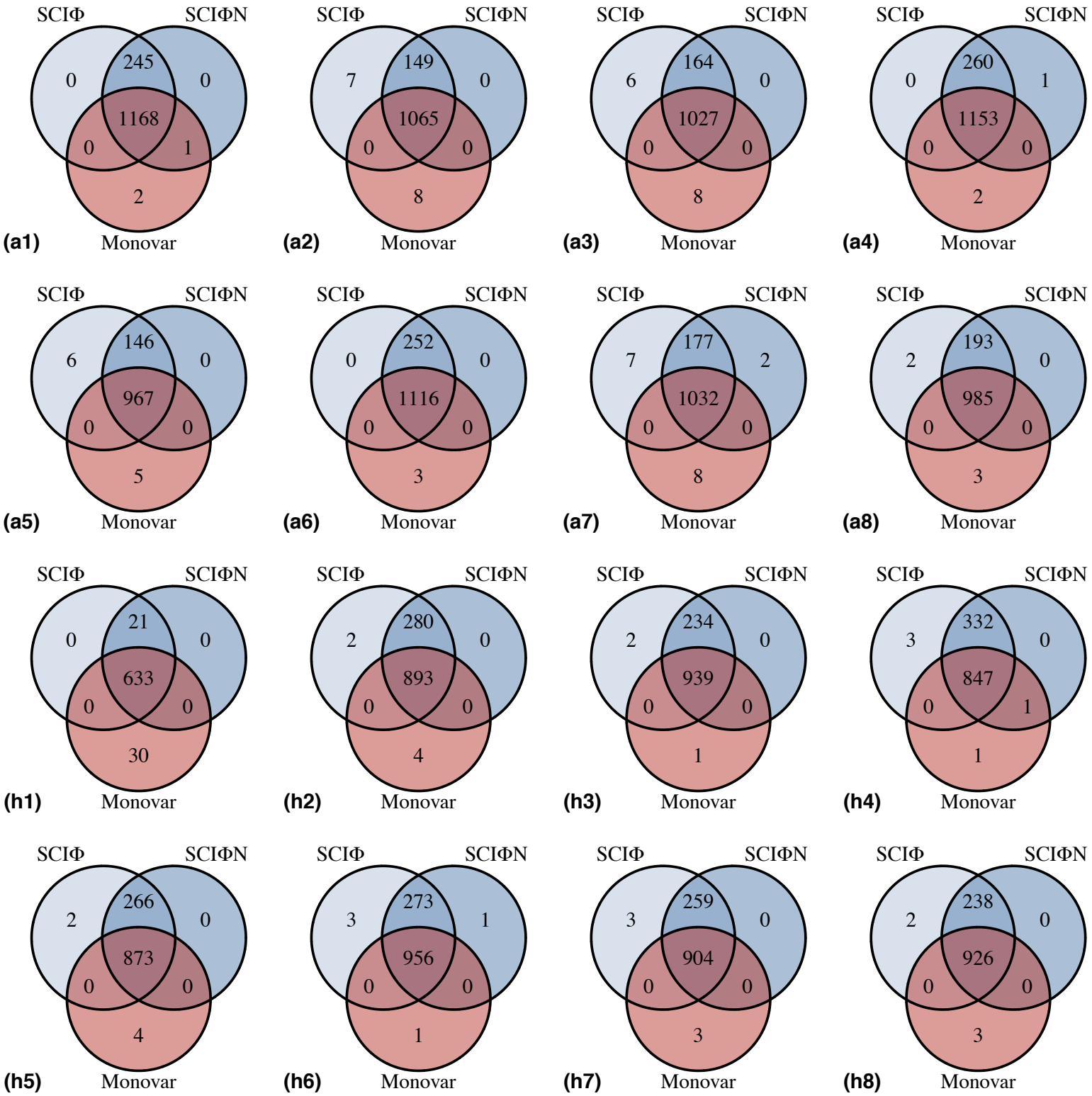
Supplementary Figure S7. Mutation calling differences on 16 breast cancer cells. The difference in probability of mutation presence, δP_m, relative to the infinite sites limit ($\chi = \infty$) for the calls displayed in Figure 3.



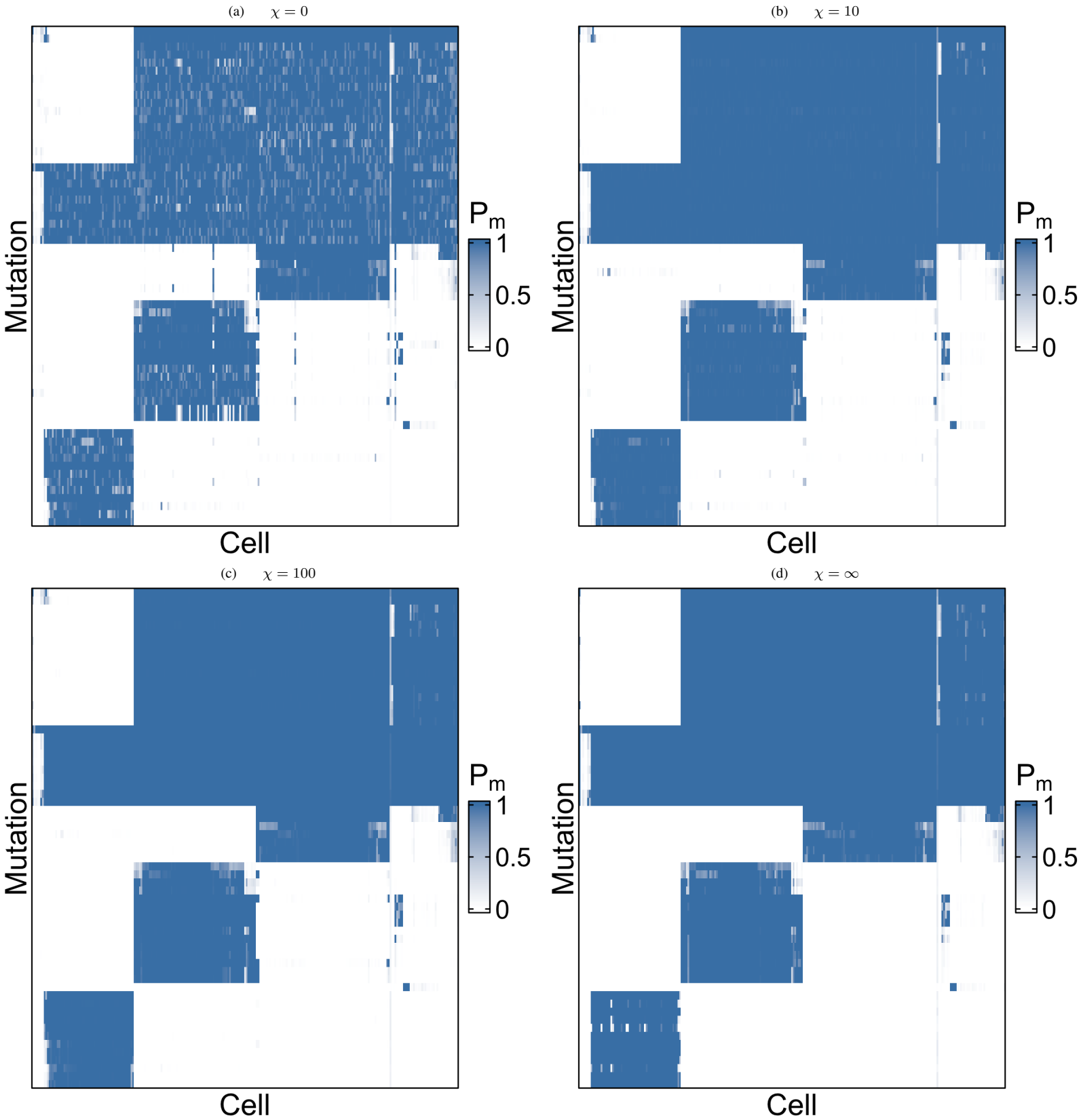
Supplementary Figure S8. Mutation subtype calling on 16 breast cancer cells. The probability of mutation presence, P_m displayed in Figure 3, broken down into components coming from heterozygous mutations, hemizygous mutations, loss of the wild type allele or loss of the mutation elsewhere in the tree, and from parallel mutations.



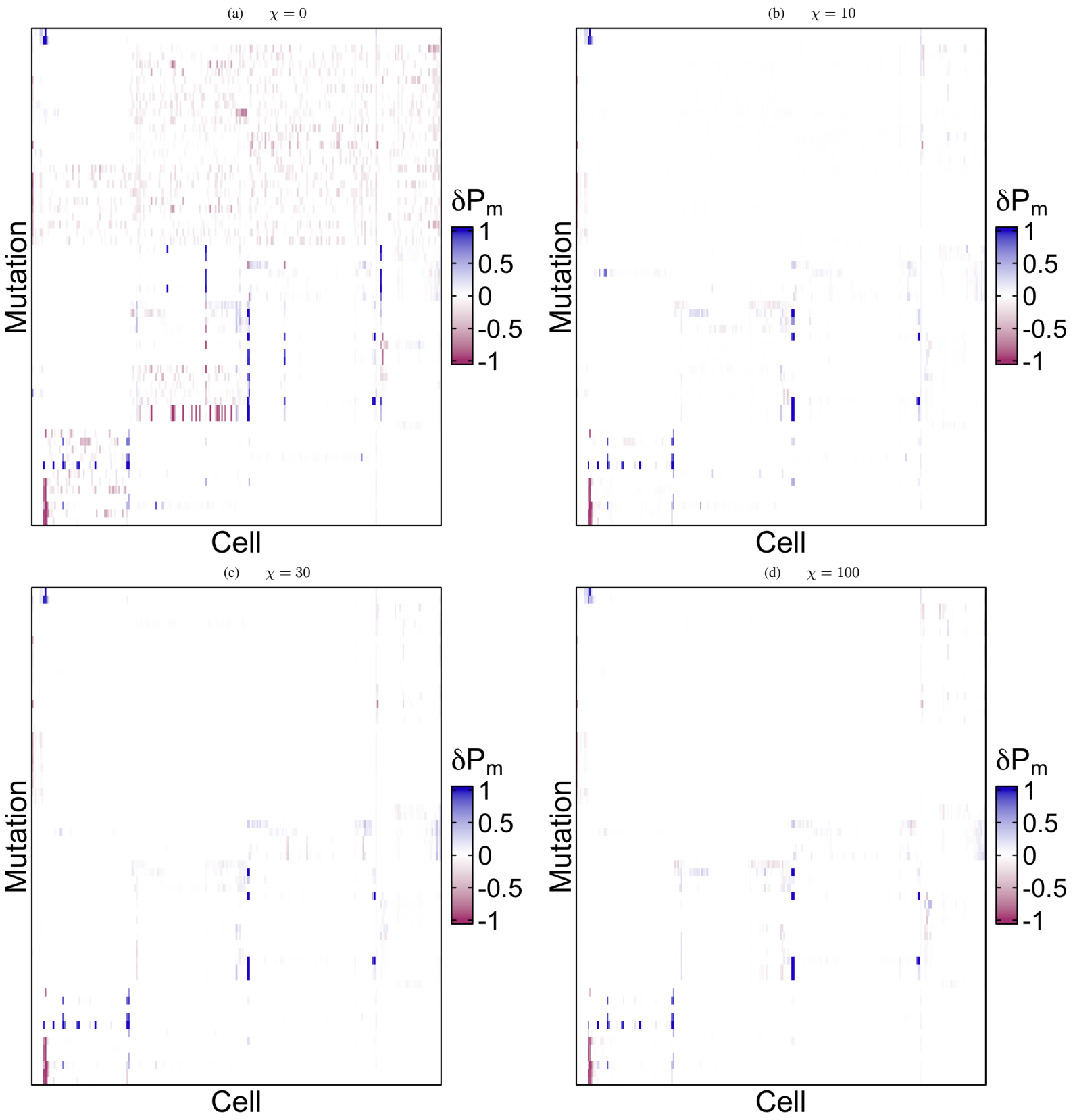
Supplementary Figure S9. Interpreting the penalisation. For each of the mutations present with a posterior probability above 95% in at least 95% of cells under the infinite sites model $\chi = \infty$, we extract its average contribution from loss or parallel mutations across cells as we relax the penalisation (blue solid lines). The average (orange dots) increases with lower χ and we may select a value to keep this below a threshold, for example 1% (red dashed line). (a) For the data from the 16 breast cancer cells, (b) for the data from the 255 acute myeloid leukemia cells.



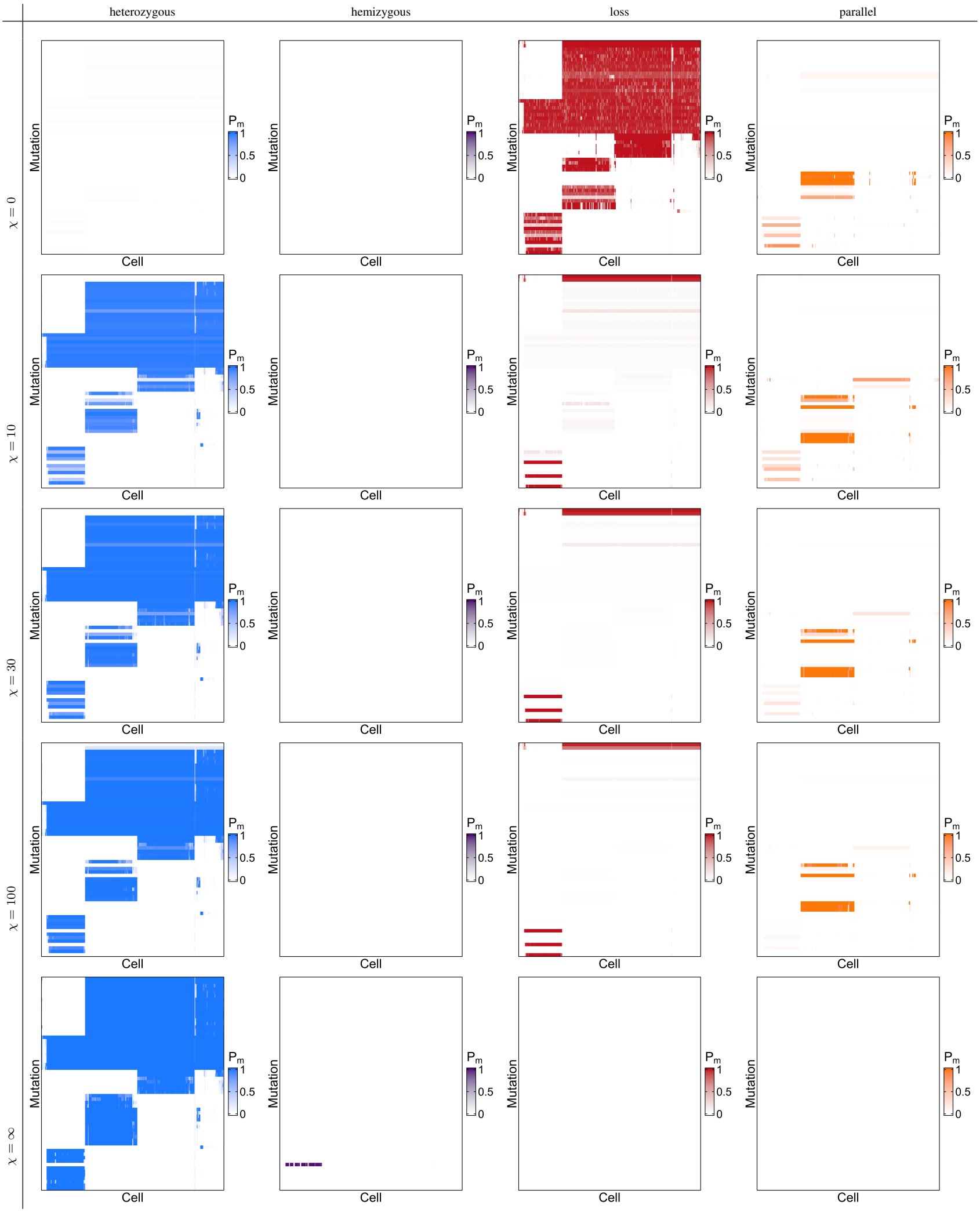
Supplementary Figure S10. Concordance of mutation calls on 16 breast cancer cells. For each of the 16 single cells (labelled in brackets) we compare the called mutations from Monovar, SCIΦ and SCIΦN with penalisation $\chi = 100$. Only loci identified by both Monovar and SCIΦN are included. For each cell, Monovar does not call mutations for loci with no coverage and these are therefore excluded from the comparison, even though SCIΦN can call mutations for such loci by sharing information from other cells through their phylogenetic relationships.



Supplementary Figure S11. Mutation calling on 255 acute myeloid leukemia cells. (a – d) the probability P_m of mutation presence in the cells as the penalisation χ is varied. The ordering of rows and columns is fixed to match panel (c).



Supplementary Figure S12. Mutation calling differences on 255 acute myeloid leukemia cells. (a – d) the difference in probability δP_m of mutation presence relative to the infinite sites limit ($\chi = \infty$) calls of Supplementary Figure S11.



Supplementary Figure S13. Mutation subtype calling on 255 acute myeloid leukemia cells. The breakdowns of the probability of mutational presence from Supplementary Figure S11, into components of heterozygous mutations, hemizygous mutations, lost mutations (where the wild type allele is lost or the mutation is lost elsewhere in the phylogeny) and parallel mutations.