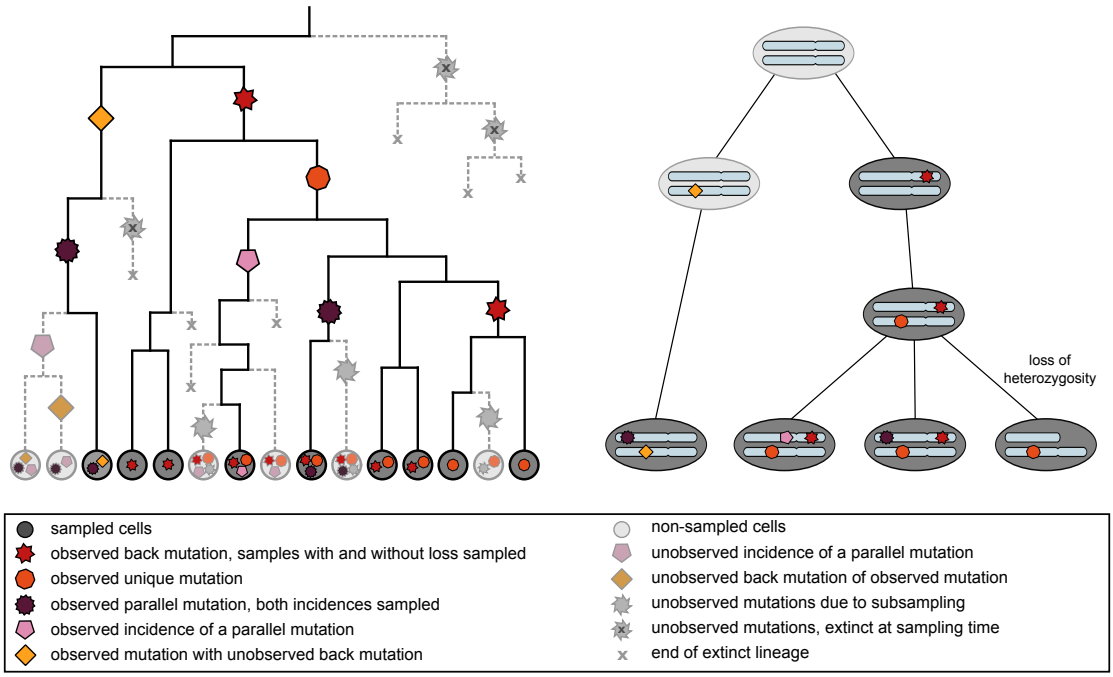
Supplemental material for "Single-cell sequencing data reveals widespread recurrence and loss of mutational hits in the life histories of tumors"

Jack Kuipers[1,2,⋆], Katharina Jahn[1,2,*], Benjamin J. Raphael[3], and Niko Beerenwinkel[1,2]

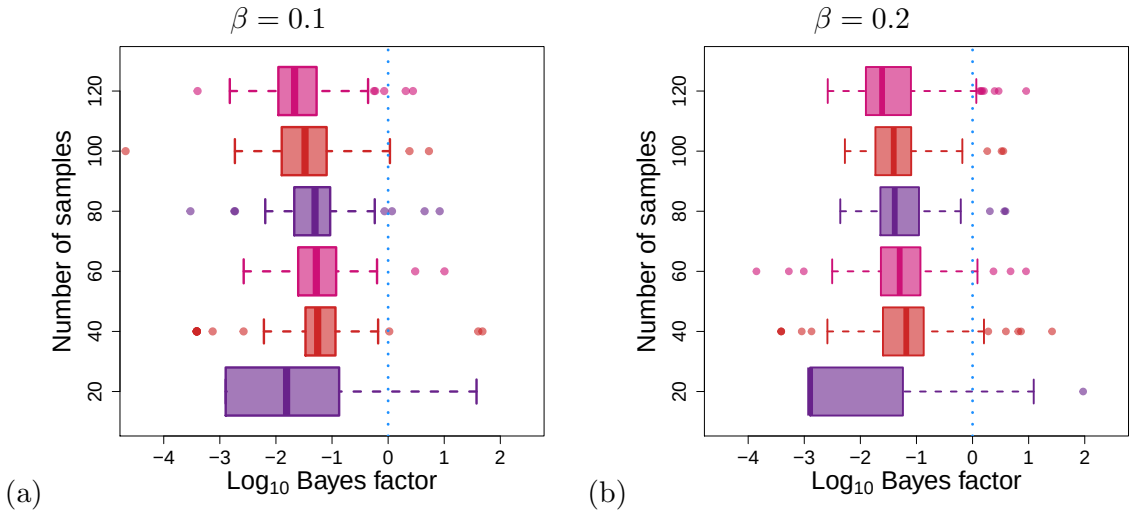
[1] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[2] SIB Swiss Institute of Bioinformatics, Basel, Switzerland
[3] Department of Computer Science, Princeton University, Princeton, NJ, USA

Supplemental Figures

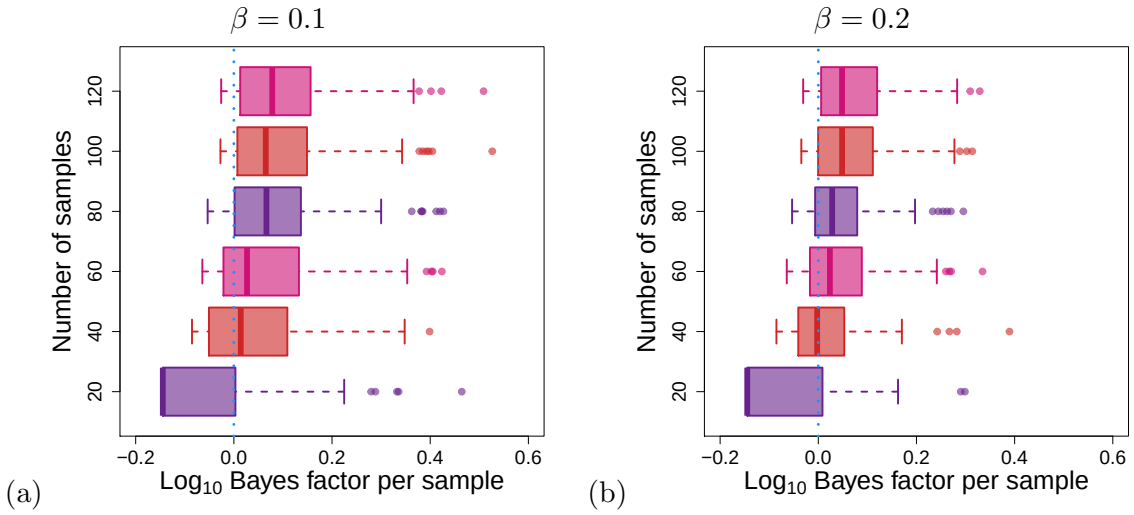


Supp. Figure 1. Left: A cell lineage tree illustrating somatic cell evolution. Due to subsampling and extinct lineages only part of the mutation history is reconstructable (black edges). Therefore some recurrent mutations may be mistaken as being unique (pink pentagon), or their back mutation may be overlooked (yellow diamond). Right: Reconstructable part of the mutation history showing that loss of heterozygosity is the cause of the lost mutation indicated by the red star: the chromosome segment containing the SNV is lost leaving only the normal allele on the matching chromosome behind.

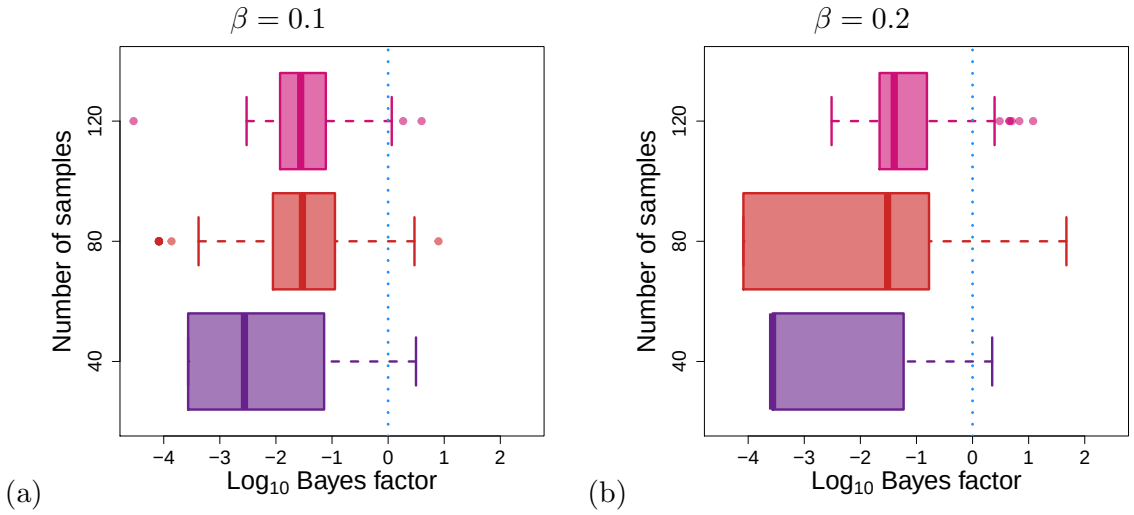
⋆ Equal contributors



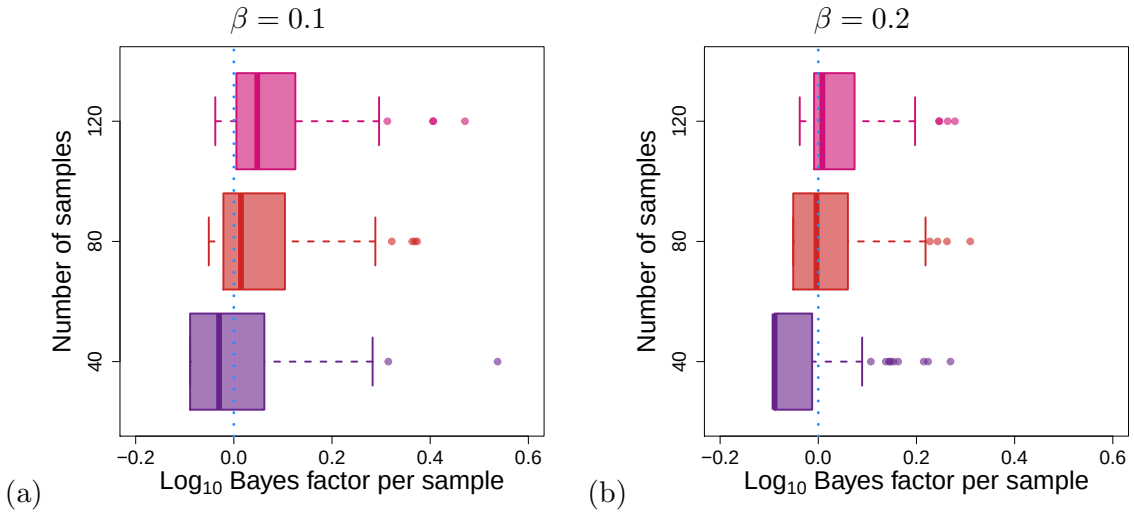
Supp. Figure 2. The range of Bayes factors estimated from simulated data with no recurrent mutations for trees with 20 mutations and up to 120 sampled cells. The false negative rate is 10% in (a) and 20% in (b).



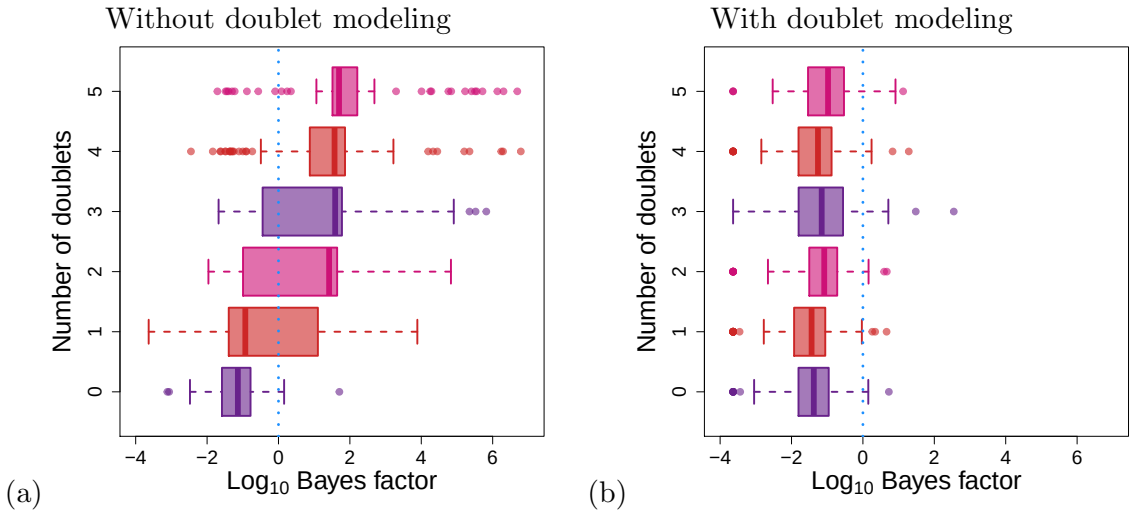
Supp. Figure 3. The range of Bayes factors estimated from simulated data with a single recurrent mutations for trees with 20 mutations and up to 120 sampled cells. To compare the Bayes factors for different numbers of samples, their log values are divided by the number of samples. The false negative rate is 10% in (a) and 20% in (b).



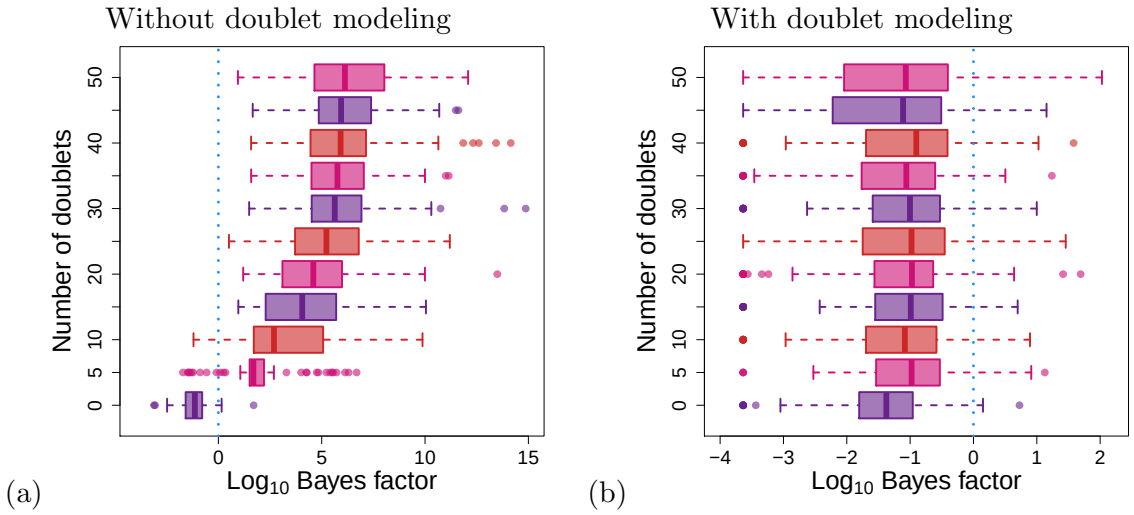
Supp. Figure 4. The range of Bayes factors estimated from simulated data with no recurrent mutations for trees with 40 mutations and up to 120 sampled cells. The false negative rate is 10% in (a) and 20% in (b).



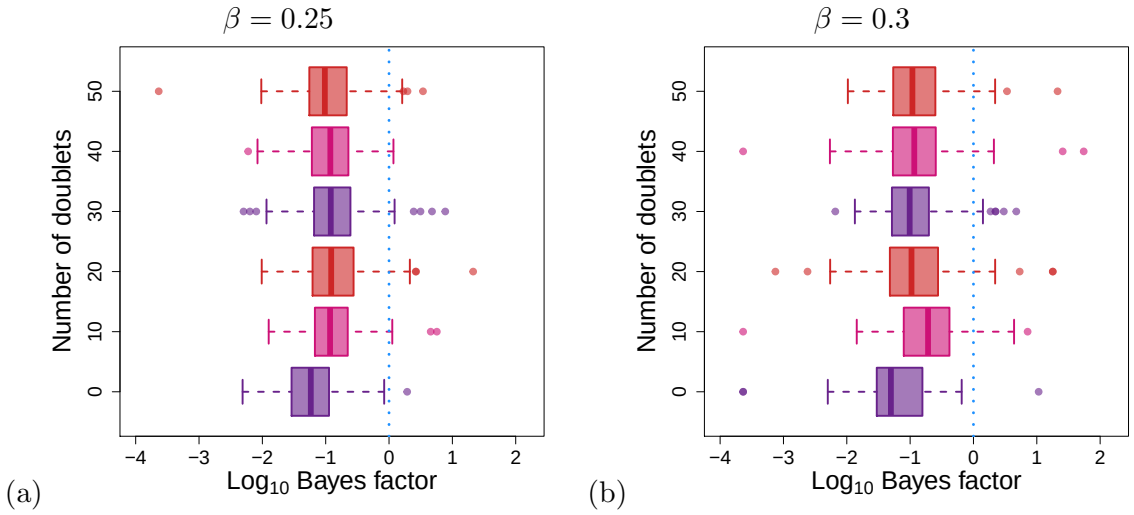
Supp. Figure 5. The range of Bayes factors estimated from simulated data with a single recurrent mutation for trees with 40 mutations and up to 120 sampled cells. To compare the Bayes factors for different numbers of samples, their log values are divided by the number of samples. The false negative rate is 10% in (a) and 20% in (b).



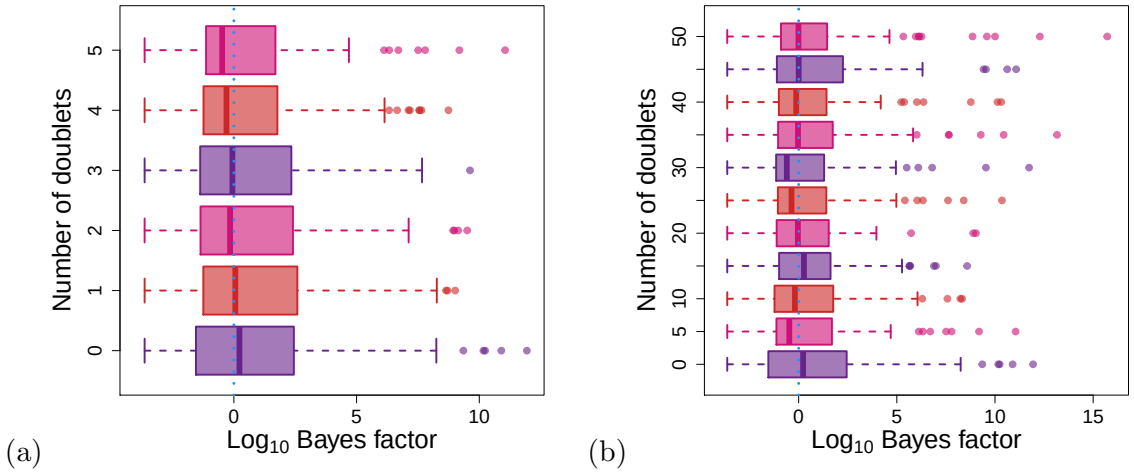
Supp. Figure 6. The range of Bayes factors estimated from simulated data with 20 mutations and no recurrent mutation for an increasing number of doublets out of 50 sampled cells. Doublet modeling is not employed in (a) but its use in (b) removes the spurious signals from sequencing two cells at once.



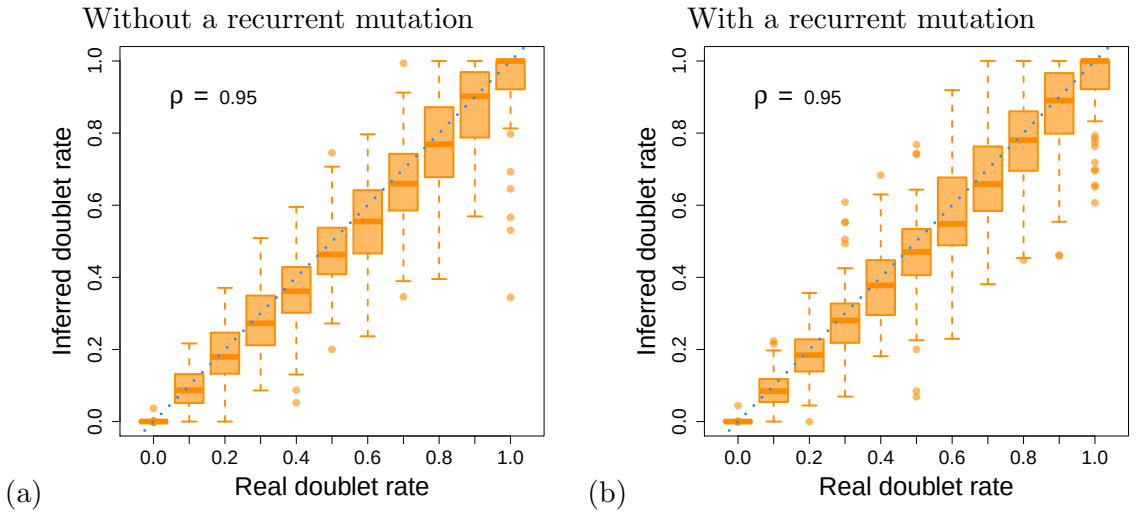
Supp. Figure 7. The range of Bayes factors estimated from simulated data with 20 mutations and no recurrent mutation for an increasing number of doublets out of 50 sampled cells. The BFs in (a) are calculated without modeling the presence of doublets, while they are accounted for in (b).



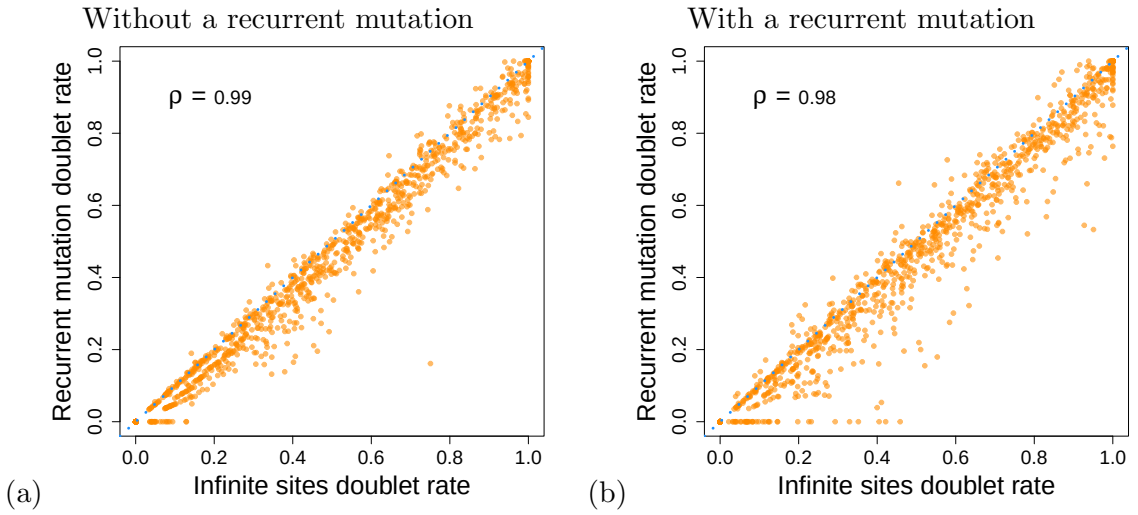
Supp. Figure 8. The range of Bayes factors estimated from simulated data with 20 mutations and no recurrent mutation for an increasing number of doublets out of 50 sampled cells when modelling the presence of doublets. The false negative rate is 25% in (a) and 30% in (b).



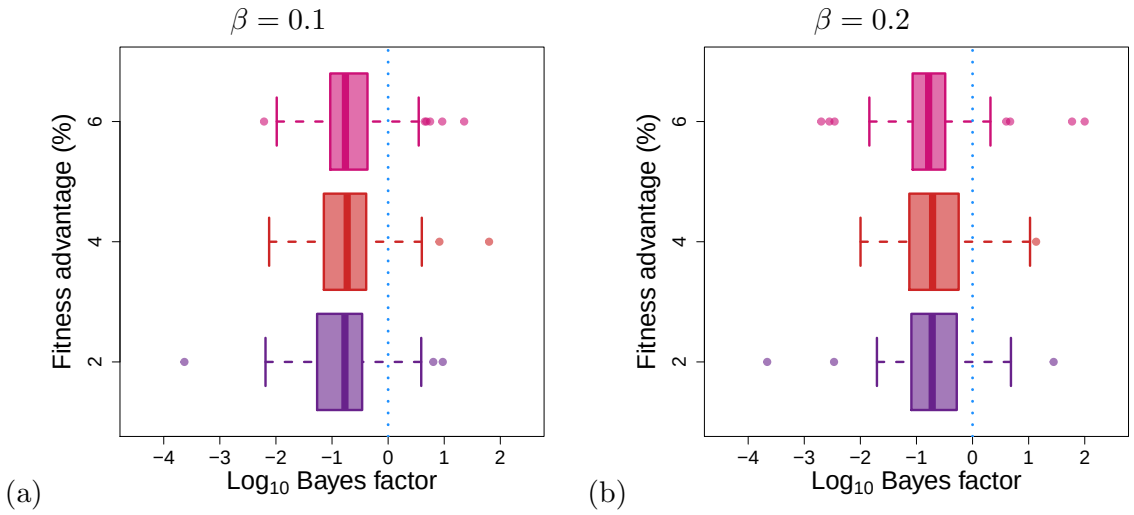
Supp. Figure 9. The range of Bayes factors estimated from simulated data with 20 mutations and a single recurrent mutation for an increasing number of doublets out of 50 sampled cells. The doublet rate increases to 10% in (a) and 100% in (b).



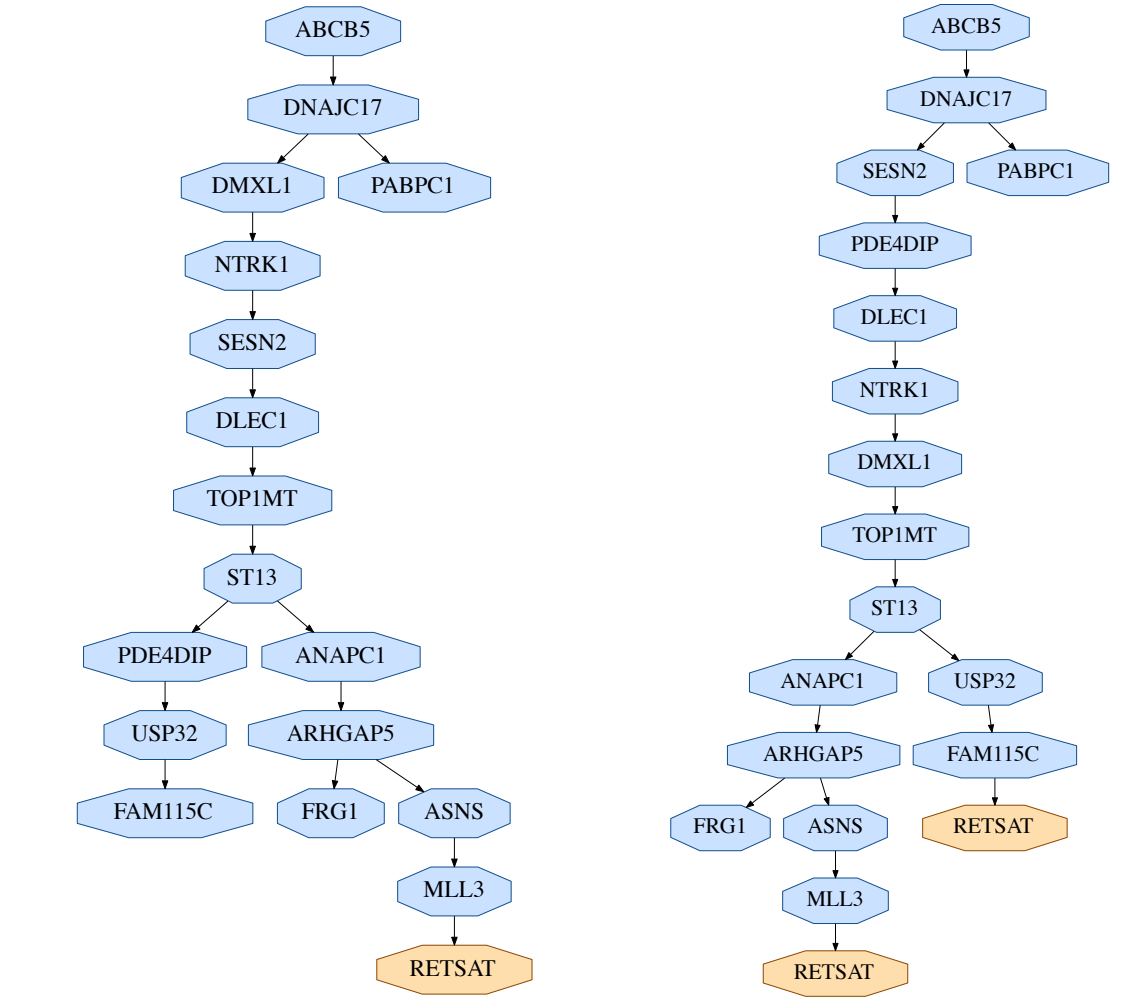
Supp. Figure 10. The doublet rates estimated from simulated data with 20 mutations and 50 sampled cells as the fraction of doublets increases. The simulations do not include any recurrent mutations in (a) but permit a single recurrence in (b).



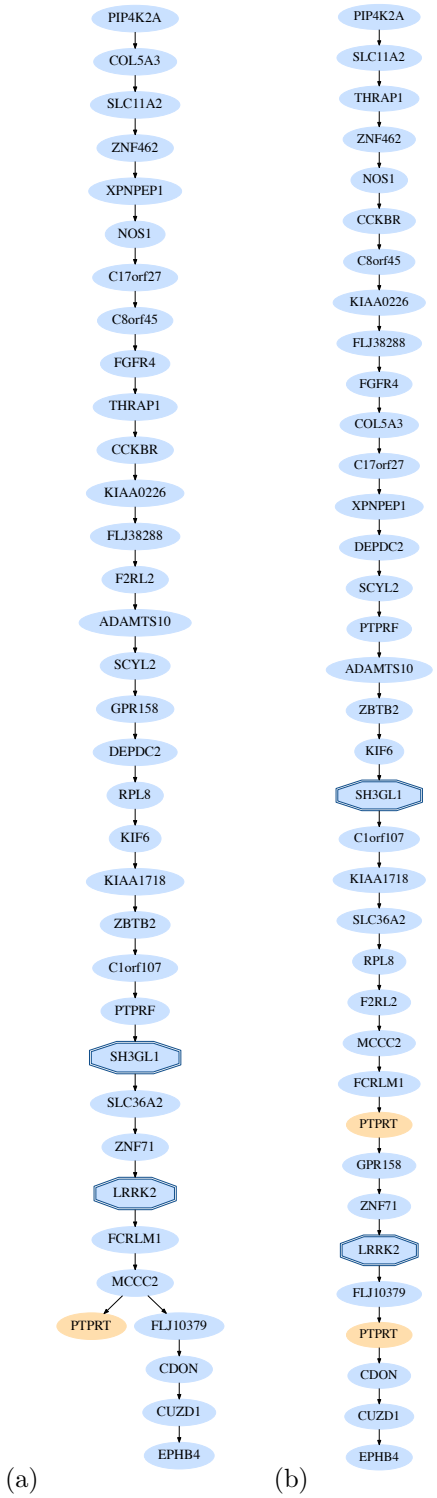
Supp. Figure 11. The doublet rates estimated from simulated data with 20 and 50 sampled cells for a range of doublet rates. For each simulated tree, the doublet rate is learnt under both the infinite sites hypothesis and allowing for a recurrent mutation. The simulations do not include any recurrent mutations in (a) and possess a single recurrence in (b).



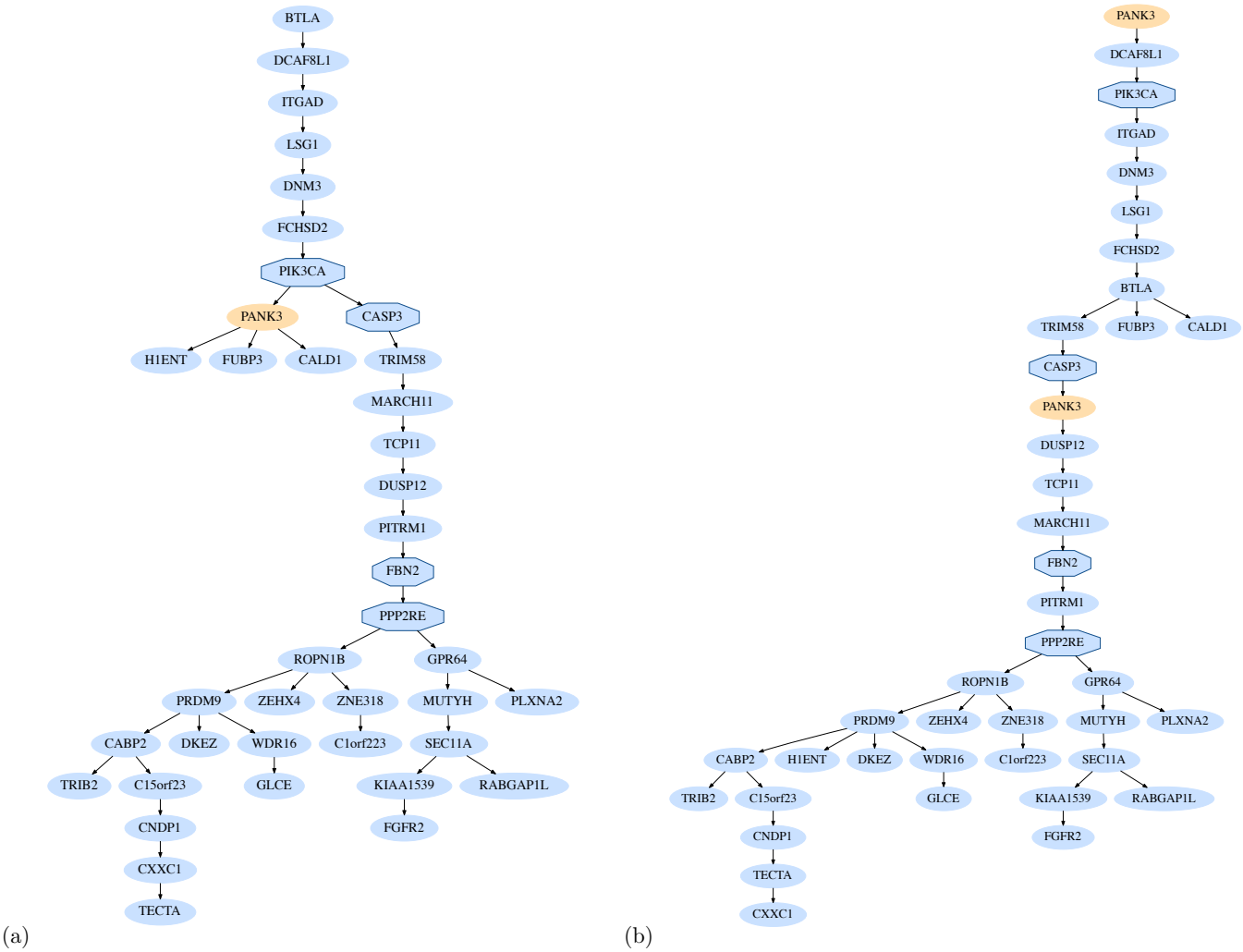
Supp. Figure 12. The range of Bayes factors estimated from data simulated under the spatial model of Waclaw et al. (2015) which utilises the infinite sites assumption. For different values of the driver fitness advantage, 50 cells (of which 5 are doublets) were sampled. The set of non-private driver mutations was used for reconstructing phylogenies and computing BFs. The false negative rate is 10% in (a) and 20% in (b).



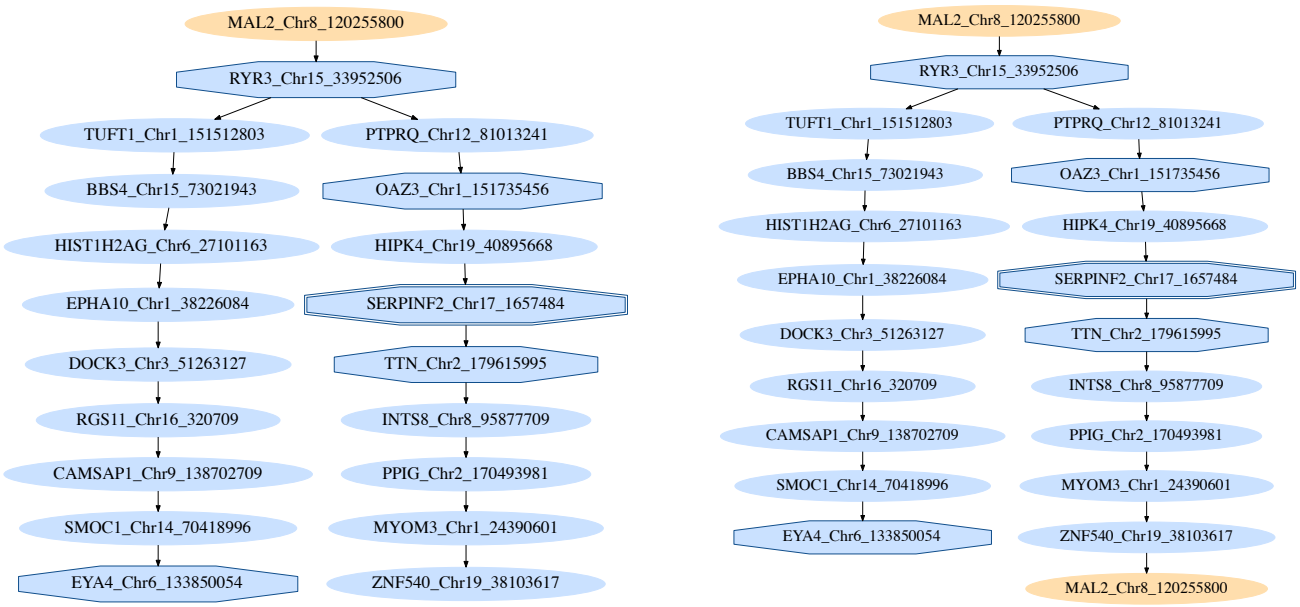
Supp. Figure 13. The best scoring trees learnt for the data on 18 selected mutations from the whole exome sequencing of 58 cells from a *JAK2*-negative myeloproliferative neoplasm (Hou et al., 2012). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation which occurs in the gene *RETSAT*. The presence or absence of the point mutation in *PDE4DIP* is unknown for over 60% of the single cells, resulting in high uncertainty and variability in its placement.



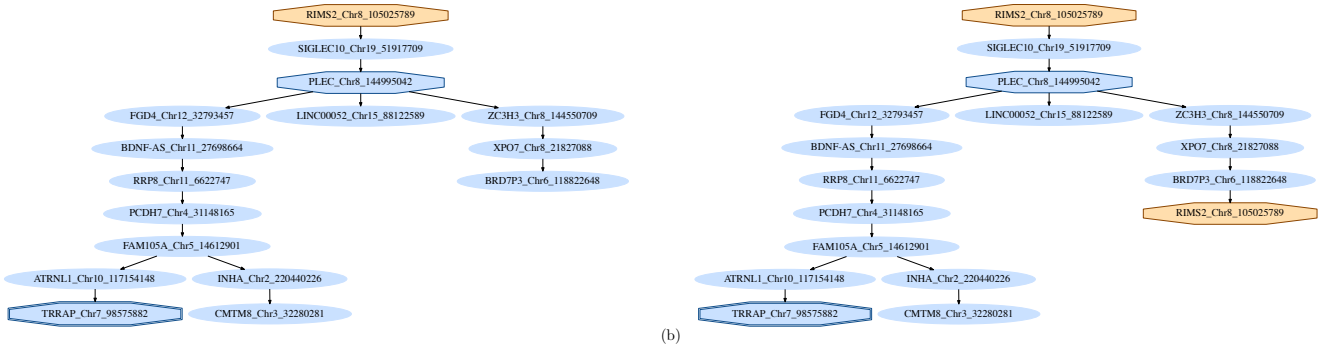
Supp. Figure 14. The best scoring trees learnt for the whole exome sequencing of 17 cells from a clear cell renal cell carcinoma (Xu et al., 2012). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation which occurs in the gene *PTPRT*. Potential driver mutations identified by Xu et al. (2012) are marked with octagonal nodes.



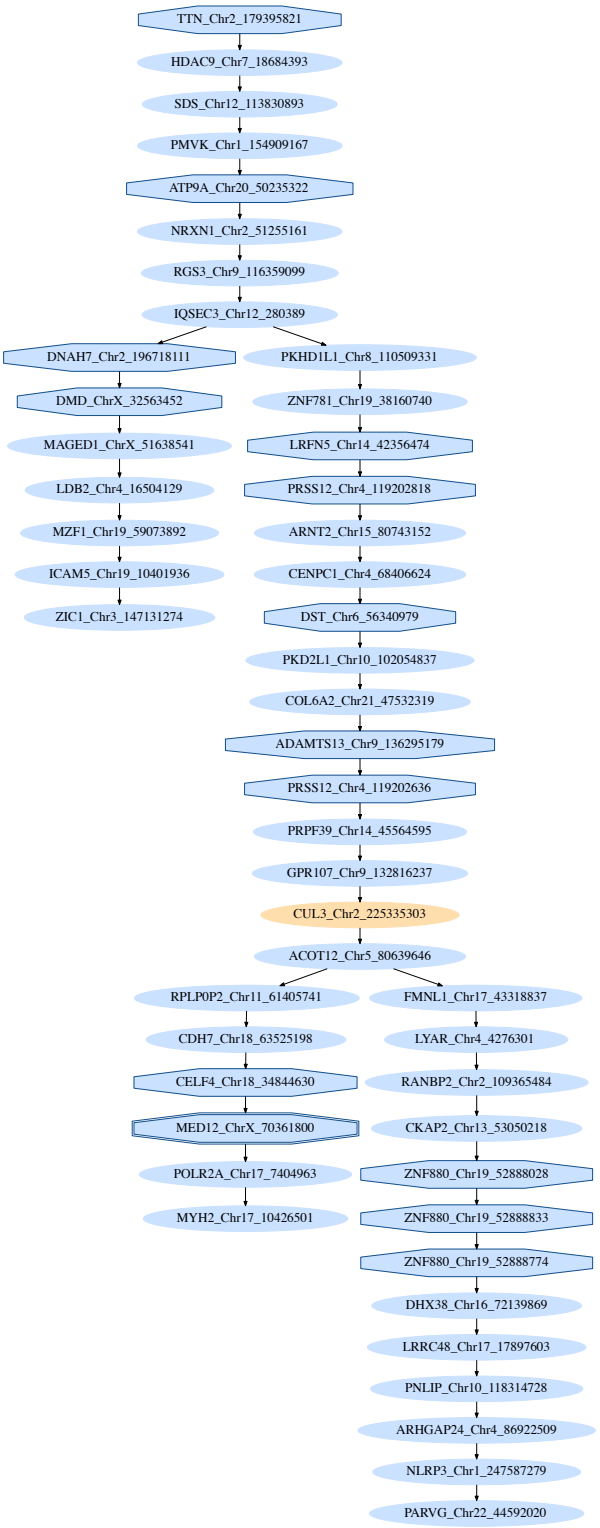
Supp. Figure 15. The best scoring trees learnt for the whole exome sequencing of 47 cells from a estrogen-receptor positive (ER+) breast cancer tumor (Wang et al., 2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation which occurs in the gene *PANK3*. Mutations identified by Wang et al. (2014) as affecting cancer genes are marked with octagonal nodes.



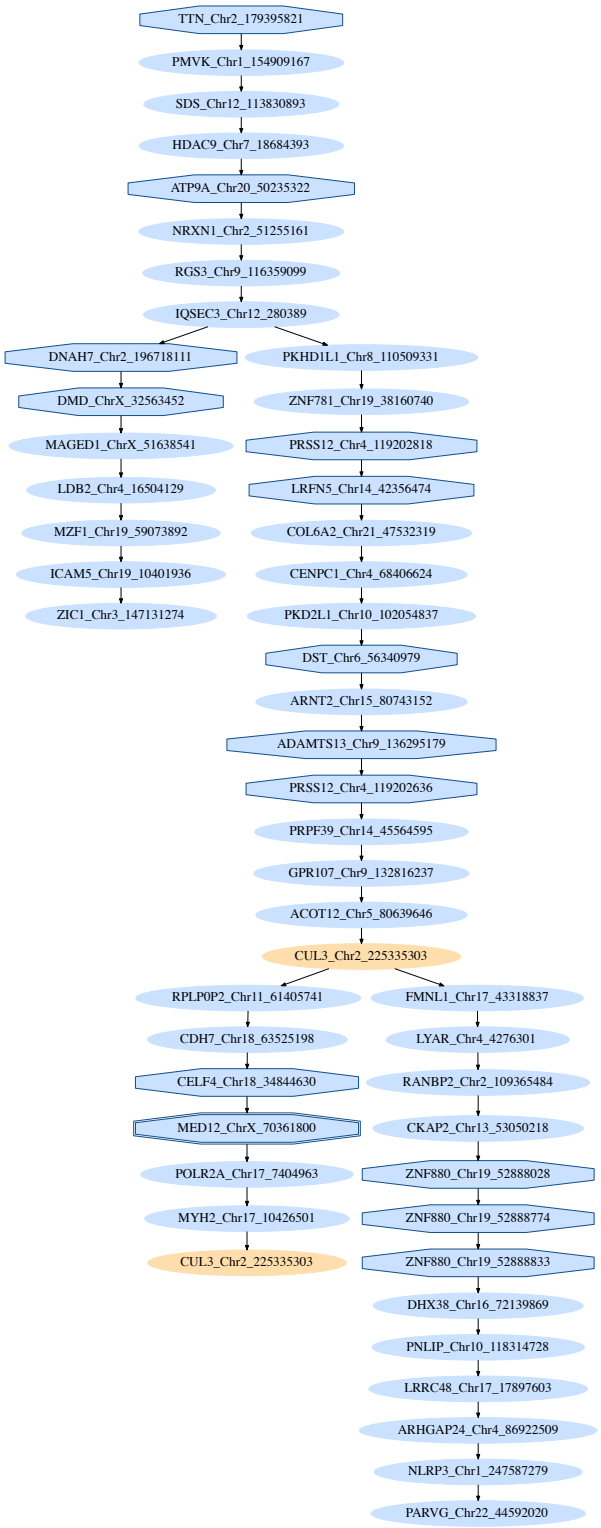
Supp. Figure 16. The best scoring trees learnt for the mutation data for patient 1 from the leukemia dataset of Gawad et al. (2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 8 at position 120255800 (hg19) in the gene *MAL2*. Common ALL mutations highlighted by Gawad et al. (2014) are indicated as octagonal nodes with the plausible driver depicted with a double border.



Supp. Figure 17. The best scoring trees learnt for the mutation data for patient 2 from the leukemia dataset of Gawad et al. (2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 8 at position 105025789 (hg19) in the gene *RIMS2*. Common ALL mutations highlighted by Gawad et al. (2014) are indicated as octagonal nodes with the plausible driver depicted with a double border.

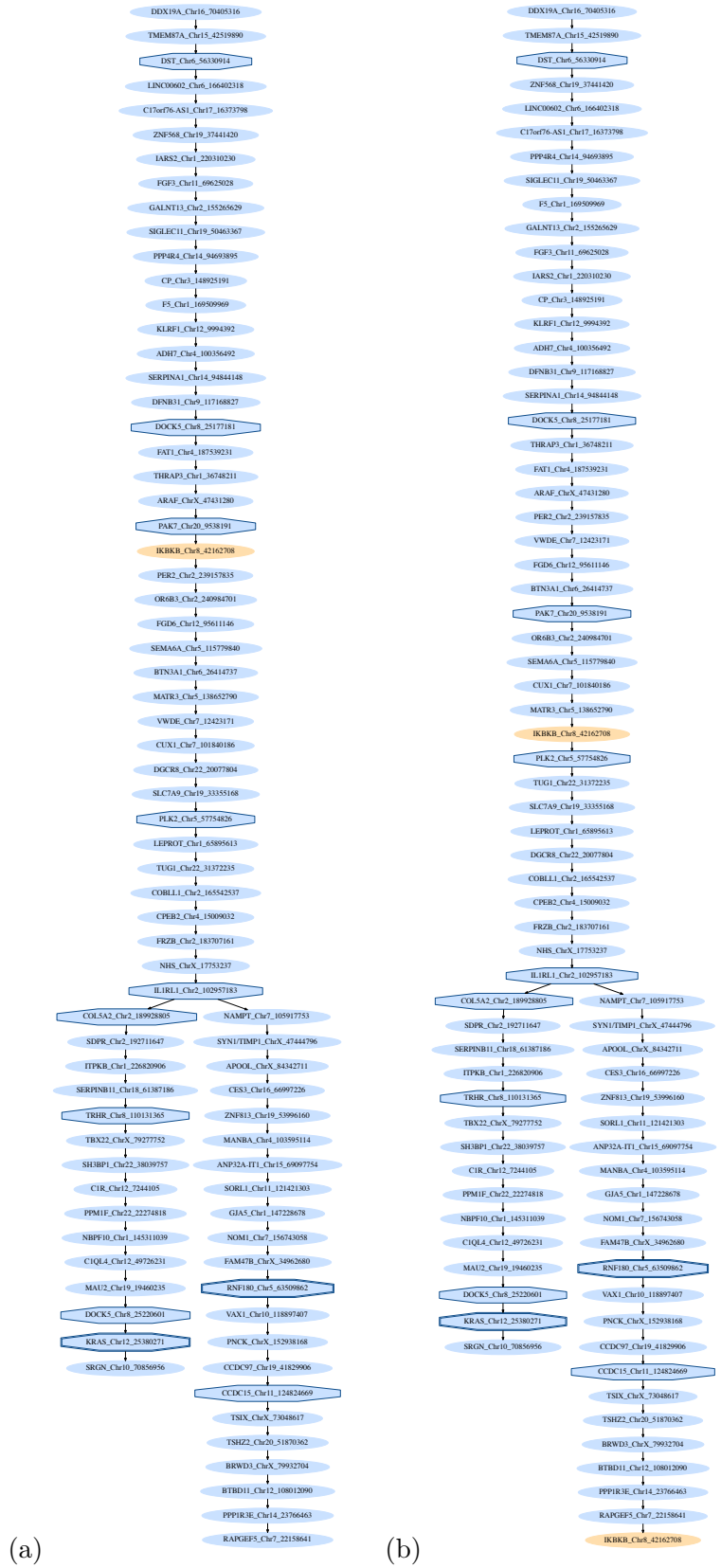


(a)

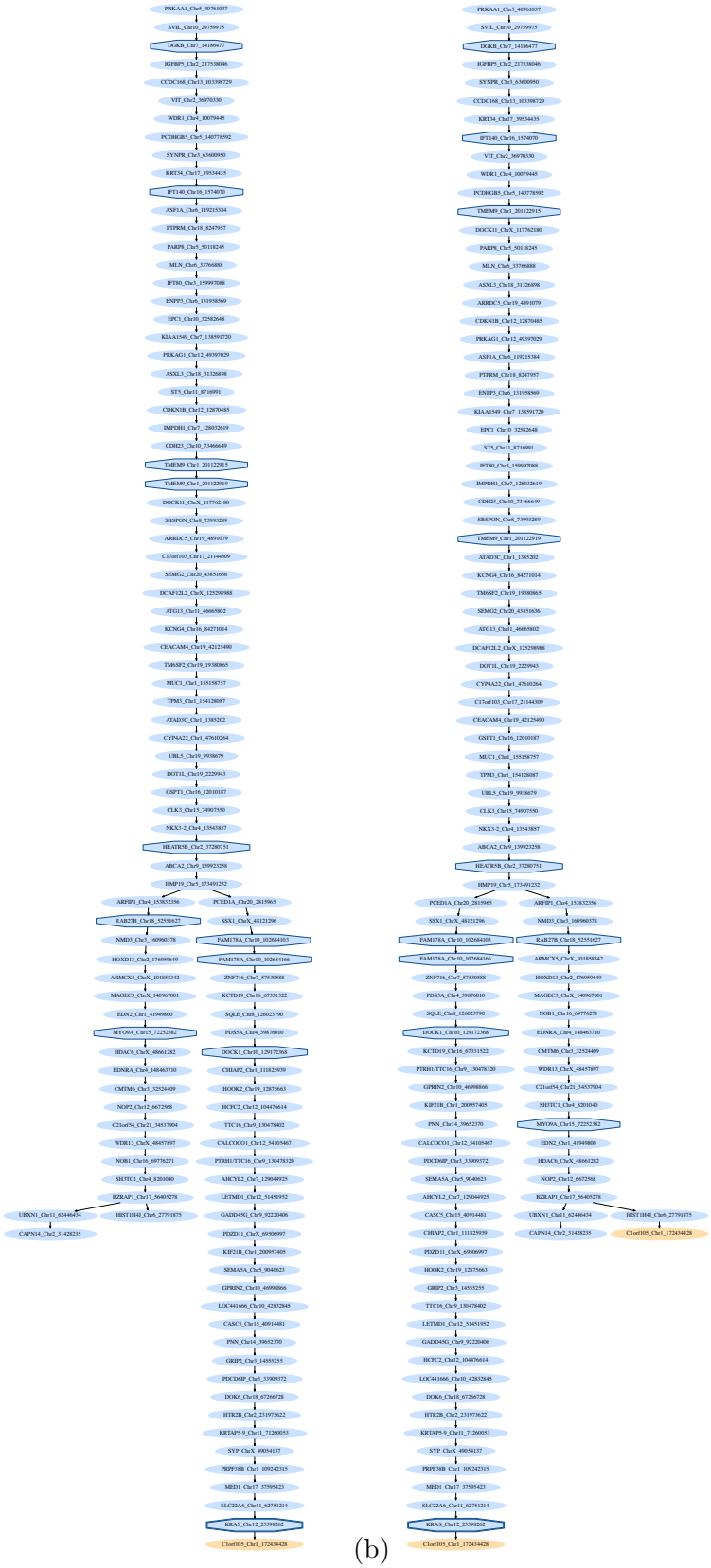


(b)

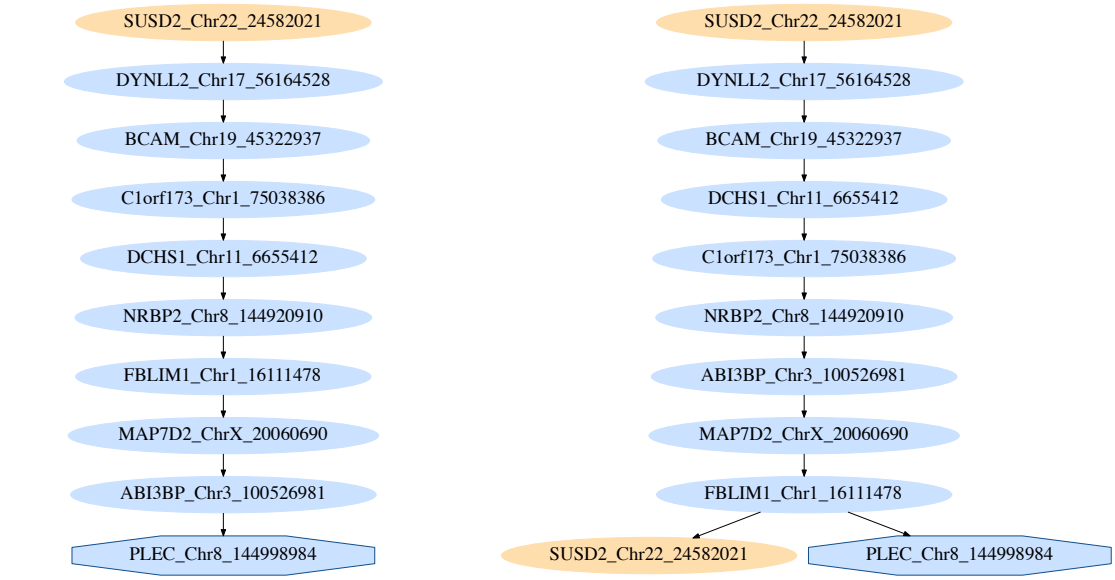
Supp. Figure 18. The best scoring trees learnt for the mutation data for patient 3 from the leukemia dataset of Gawad et al. (2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 2 at position 225335303 (hg19) in the gene *CUL3*. Common ALL mutations highlighted by Gawad et al. (2014) are indicated as octagonal nodes with the plausible driver depicted with a double border.



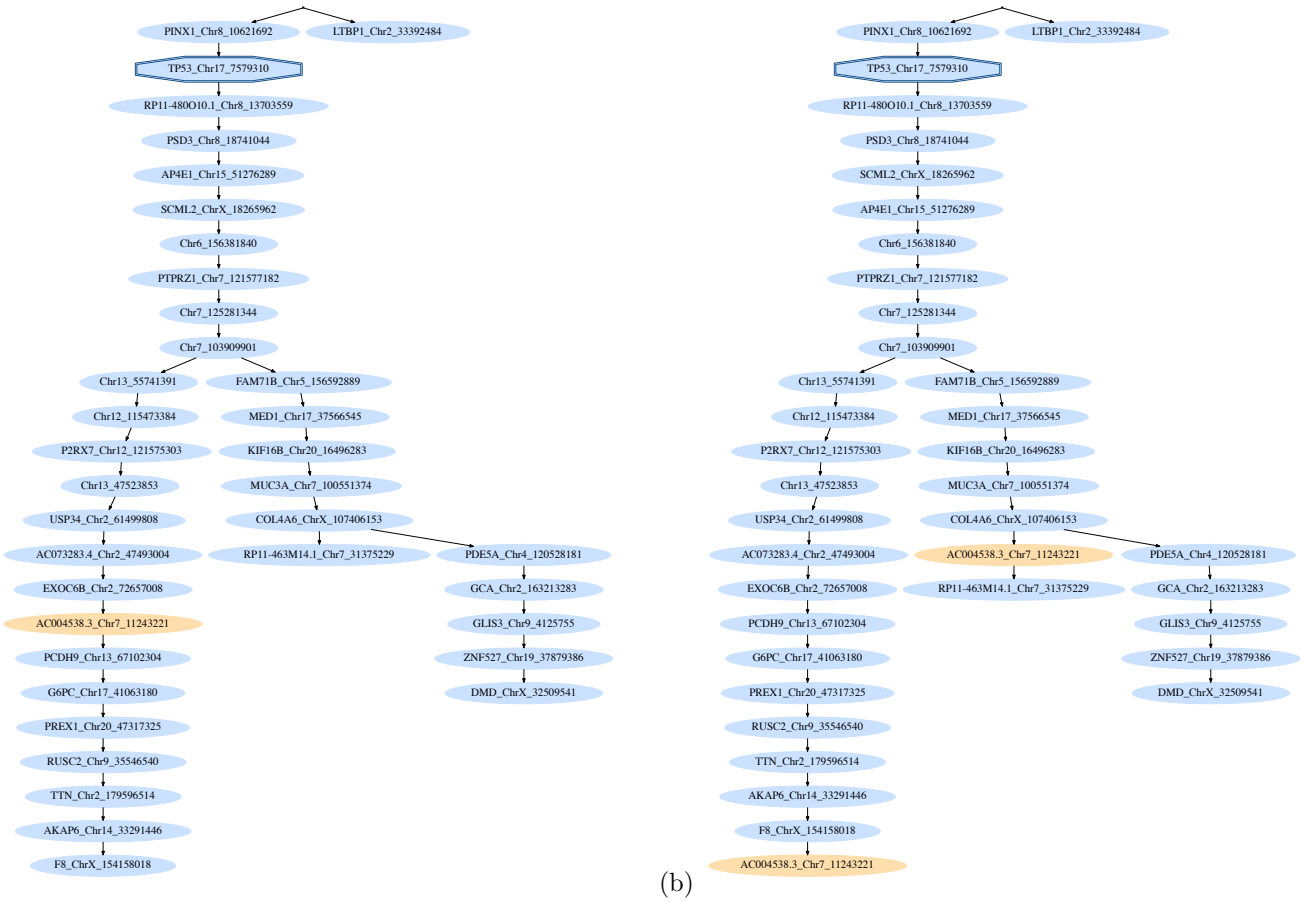
Supp. Figure 19. The best scoring trees learnt for the mutation data for patient 4 from the leukemia dataset of Gawad et al. (2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 1 at position 42162708 (hg19) in the gene *IKBKB*. Common ALL mutations highlighted by Gawad et al. (2014) are indicated as octagonal nodes with the plausible drivers depicted with a double border.



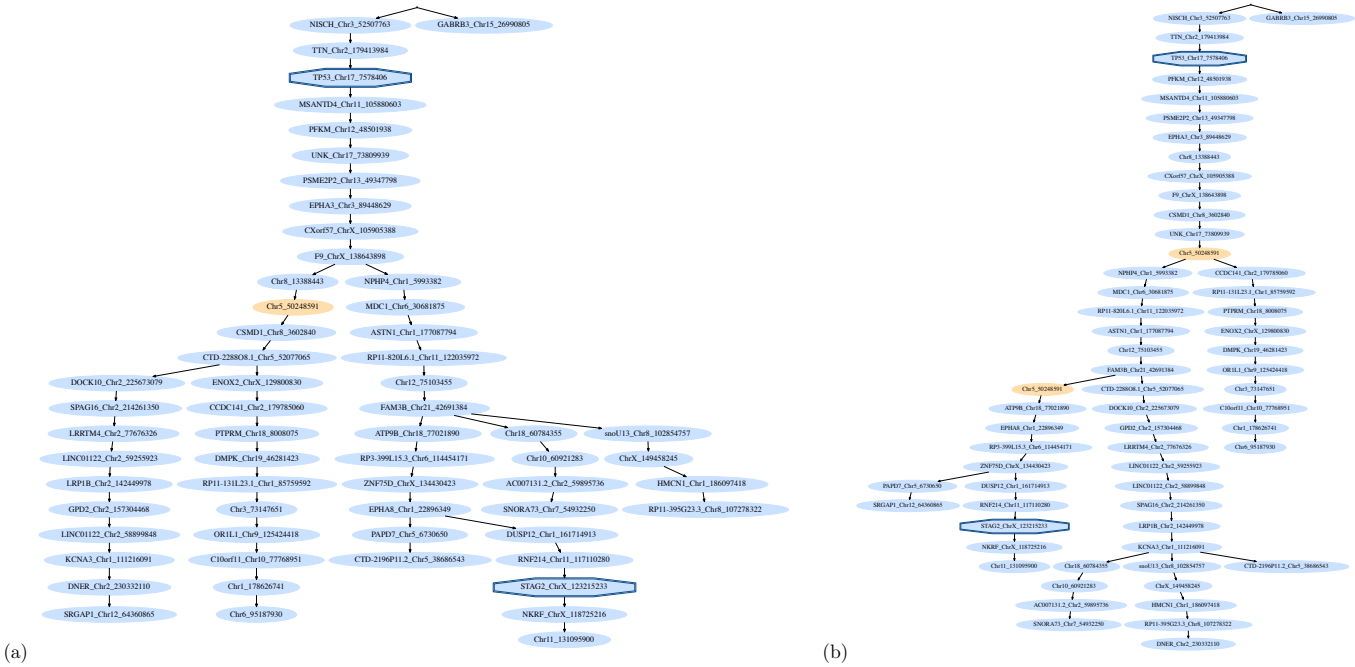
Supp. Figure 20. The best scoring trees learnt for the mutation data for patient 5 from the leukemia dataset of Gawad et al. (2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 8 at position 172434428 (hg19) in the gene *C1orf105*. Common ALL mutations highlighted by Gawad et al. (2014) are indicated as octagonal nodes with the plausible driver depicted with a double border.



Supp. Figure 21. The best scoring trees learnt for the mutation data for patient 6 from the leukemia dataset of Gawad et al. (2014). The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 22 at position 24582021 (hg19) in the gene *SUSD2*. The common ALL mutation highlighted by Gawad et al. (2014) is indicated as an octagonal node.

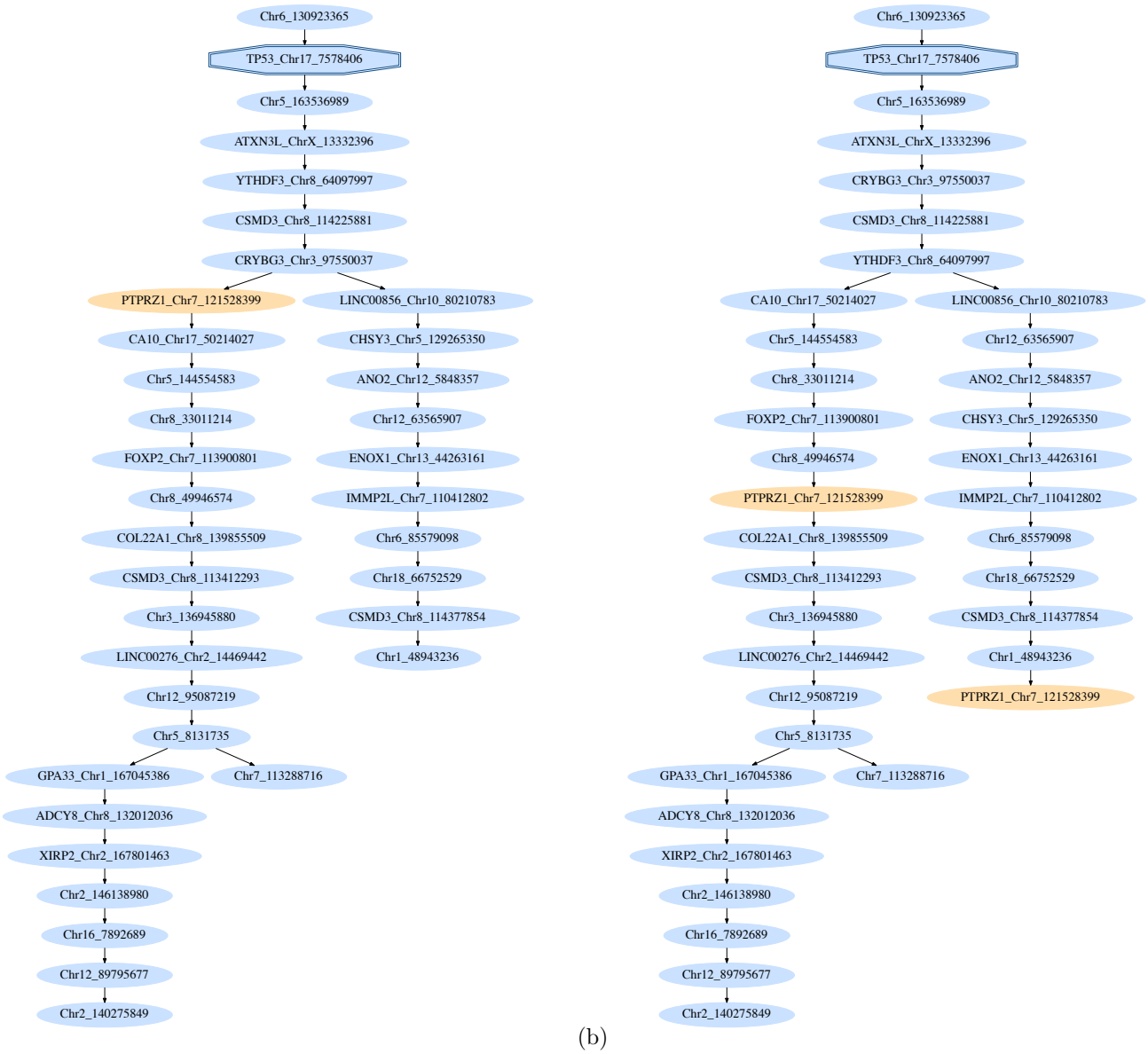


Supp. Figure 22. The best scoring trees learnt for the mutation data for patient 2 from the ovarian cancer dataset of McPherson et al. (2016) when mutations present in the normal cells are excluded from the analysis. The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 7 at position 11243221 (hg19) in the gene *AC004538.3*. The driver mutation in *TP53* highlighted by McPherson et al. (2016) is indicated as an octagonal node.

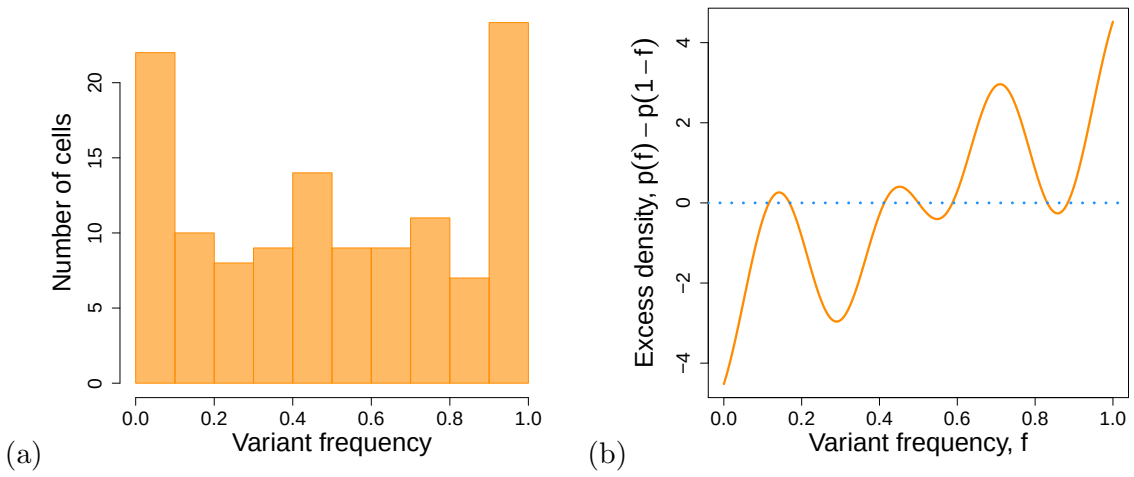


(a) (b)

Supp. Figure 23. The best scoring trees learnt for the mutation data for patient 3 from the ovarian cancer dataset of McPherson et al. (2016) when mutations present in the normal cells are excluded from the analysis. The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 5 at position 50248591 (hg19). The driver mutations highlighted by McPherson et al. (2016) are indicated as octagonal nodes.



Supp. Figure 24. The best scoring trees learnt for the mutation data for patient 9 from the ovarian cancer dataset of McPherson et al. (2016) when mutations present in the normal cells are excluded from the analysis. The tree in (a) respects the infinite sites hypothesis while (b) allows for the highest scoring recurrent mutation on Chromosome 7 at position 121528399 (hg19) in the gene *PTPRZ1*. The driver mutation in *TP53* highlighted by McPherson et al. (2016) is indicated as an octagonal node.



Supp. Figure 25. (a) Histogram of the proportion of reads supporting the parallel mutation on Chromosome 8 at position 172434428 (hg19) in the gene *C1orf105* in patient 5 from the leukemia dataset of Gawad et al. (2014) across all the sequenced single cells. Cells with only supporting reads, or no supporting reads have been excluded. (b) The excess of low frequency variants is plotted by taking the kernel density $p(f)$ of the variant frequencies and subtracting a symmetrized version: $p(f) - p(1-f)$. Since this is negative for low frequencies, this excludes the possibility of excess false positive mutation calls at this genomic position.

Supplemental Tables

Dataset	Hou et al. (2012)	Xu et al. (2012)	Wang et al. (2014)
cancer type	myeloproliferative neoplasm	renal cell carcinoma	ER^+ breast cancer
no. of mutations	18	35	40
no. of cells	58	17	47
Recurrent mutation			
type	parallel	lost	lost
gene	*RETSAT*	*PTPRT*	*PANK3*
Estimated parameters			
false positive rate α	6.04×10^{-5}	2.67×10^{-5}	1.24×10^{-6}
false negative rate β	0.401	0.192	0.171
doublet rate δ ($\mathcal{M}_{\mathrm{I}}$)	1.000	1.000	1.000
doublet rate δ ($\mathcal{M}_{\mathrm{F}}$)	1.000	1.000	1.000
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{I}}$)	0.346	0.070	0.557
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{F}}$)	0.307	0.000	0.280
Inferred trees			
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{I}}$)	-132.62	-112.25	-189.29
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{F}}$)	-127.65	-109.63	-182.32
$\log_{10}$ Bayes factor $\tilde{B}_{\mathrm{FI}}$	1.13	-0.57	3.31
Bayes factor $\tilde{B}_{\mathrm{FI}}$	13	0.27	2000

Supp. Table 1. The characteristics of the three exome sequencing datasets along with their inferred parameters and BFs derived from comparing the highest scoring trees in each class.

Patient	**1**	**2**	**3**	**4**	**5**	**6**
no. of mutations	20	16	49	78	105	10
no. of cells	111	115	150	143	96	146
panel size	98	82	196	155	248	85
Recurrent mutation						
type	lost	lost	lost	lost	parallel	lost
gene	*MAL2*	*RIMS2*	*CUL3*	*IKBKB*	*C1orf105*	*SUSD2*
chromosome	Chr8	Chr8	Chr2	Chr8	Chr1	Chr22
genomic position	120255800	105025789	225335303	42162708	172434428	24582021
substitution	C→G	T→G	G→C	C→G	C→A	C→T
Estimated parameters						
false positive rate α	4.3×10^{-4}	1.0×10^{-3}	5.1×10^{-4}	1.9×10^{-3}	1.2×10^{-3}	$< 10^{-6}$
false negative rate β	0.173	0.160	0.267	0.249	0.291	0.163
doublet rate δ ($\mathcal{M}_{\mathrm{I}}$)	1.000	0.578	1.000	0.971	0.944	1.000
doublet rate δ ($\mathcal{M}_{\mathrm{F}}$)	1.000	0.588	1.000	0.965	0.943	1.000
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{I}}$)	0.239	0.075	0.132	0.109	0.235	0.000
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{F}}$)	0.239	0.077	0.132	0.108	0.235	0.027
Inferred trees						
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{I}}$)	-282.91	-181.88	-1040.84	-1906.52	-1877.99	-195.17
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{F}}$)	-272.07	-174.10	-1025.47	-1894.66	-1857.92	-173.67
$\log_{10}$ Bayes factor $\tilde{B}_{\mathrm{FI}}$	5.93	2.47	10.62	7.25	15.68	13.99
Bayes factor $\tilde{B}_{\mathrm{FI}}$	8.6×10^5	300	4.1×10^{10}	1.8×10^7	4.8×10^{15}	9.7×10^{13}

Supp. Table 2. The characteristics of the panel sequencing datasets of six leukemia patient samples of Gawad et al. (2014) along with their inferred parameters and BFs derived from comparing the highest scoring trees in each class. The genomic positions are according to the hg19 assembly.

Patient	**2**	**3**	**9**
no. of mutations	37	60	37
no. of cells	588	672	420
Recurrent mutation			
type	parallel	lost	parallel
gene	*AC004538.3*	–	*PTPRZ1*
chromosome	Chr7	Chr5	Chr7
genomic position	11243221	50248591	121528399
substitution	G→T	T→C	G→C
Estimated parameters			
false positive rate α	1.5×10^{-2}	9.8×10^{-3}	3.0×10^{-2}
false signal rate β	0.465	0.438	0.378
doublet rate δ ($\mathcal{M}_{\mathrm{I}}$)	0.370	0.743	0.499
doublet rate δ ($\mathcal{M}_{\mathrm{F}}$)	0.367	0.392	0.302
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{I}}$)	0.198	0.567	0.089
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{F}}$)	0.202	0.168	0.047
Inferred trees			
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{I}}$)	-3022.22	-5564.23	-3011.09
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{F}}$)	-2995.70	-5522.35	-2991.31
$\log_{10}$ Bayes factor $\tilde{B}_{\mathrm{FI}}$	16.65	33.47	11.81
Bayes factor $\tilde{B}_{\mathrm{FI}}$	4.4×10^{16}	2.9×10^{33}	6.5×10^{11}

Supp. Table 3. The characteristics, inferred parameters and BFs for three ovarian cancers from the study of McPherson et al. (2016) when mutations present in normal cells are excluded from the analysis. Genomic positions are according to the hg19 assembly.

Patient	**2**	**3**	**9**
no. of mutations	43	84	43
no. of cells	588	672	420
Recurrent mutation			
type	parallel	lost	lost
gene	*AC004538.3*	–	*PTPRZ1*
chromosome	Chr7	Chr5	Chr7
genomic position	11243221	50248591	121528399
substitution	G→T	T→C	G→C
Estimated parameters			
false positive rate α	2.0×10^{-2}	1.6×10^{-2}	2.0×10^{-2}
false signal rate β	0.436	0.442	0.384
doublet rate δ ($\mathcal{M}_{\mathrm{I}}$)	0.501	1.000	0.923
doublet rate δ ($\mathcal{M}_{\mathrm{F}}$)	0.496	1.000	1.000
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{I}}$)	0.252	0.458	0.229
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{F}}$)	0.256	0.452	0.254
Inferred trees			
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{I}}$)	-3551.47	-9114.51	-3225.17
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{F}}$)	-3524.84	-9076.18	-3205.58
$\log_{10}$ Bayes factor $\tilde{B}_{\mathrm{FI}}$	17.53	30.96	12.12
Bayes factor $\tilde{B}_{\mathrm{FI}}$	3.4×10^{17}	9.0×10^{30}	1.3×10^{12}

Supp. Table 4. The characteristics of the panel sequencing of three ovarian cancers from the study of McPherson et al. (2016) along with their inferred parameters and BFs derived from comparing the highest scoring trees in each class. The genomic positions are according to the hg19 assembly.

Patient	**1**	**2**	**3**	**4**
no. of mutations	20	16	49	78
no. of cells	111	115	150	143
panel size	98	82	196	155
Recurrent mutation				
type	parallel	parallel	parallel	parallel
gene	*CAMSAP1*	*LINC00052*	*FMNL1*	*SYN1/TIMP1*
chromosome	Chr9	Chr15	Chr17	ChrX
genomic position	138702709	88122589	43318837	47444796
substitution	C→T	G→T	C→T	G→A
Estimated parameters				
false positive rate α	4.3×10^{-4}	1.0×10^{-3}	5.1×10^{-4}	1.9×10^{-3}
false negative rate β	0.173	0.160	0.267	0.249
doublet rate δ ($\mathcal{M}_{\mathrm{I}}$)	1.000	0.578	1.000	0.971
doublet rate δ ($\mathcal{M}_{\mathrm{F}}$)	1.000	0.474	1.000	0.969
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{I}}$)	0.239	0.075	0.132	0.109
rel. doublet rate $\tilde{\delta}$ ($\mathcal{M}_{\mathrm{F}}$)	0.239	0.056	0.117	0.110
Inferred trees				
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{I}}$)	-282.91	-181.88	-1040.84	-1906.52
$\log_{10}$ tree score ($\mathcal{M}_{\mathrm{F}}$)	-276.15	-175.42	-1029.71	-1900.74
$\log_{10}$ Bayes factor $\tilde{B}_{\mathrm{FI}}$	1.85	1.16	6.38	1.17
Bayes factor $\tilde{B}_{\mathrm{FI}}$	71	14	2.4×10^{6}	14

Supp. Table 5. We consider the leukemia patient samples of Gawad et al. (2014) with lost mutations, patients 1–4 and 6. For each of these we focus on possible additional parallel mutations. In particular we compute the BFs for the highest scoring parallel mutation by comparing the highest scoring trees in each class. For patients 1–4 we find some evidence of a parallel mutations, and fairly strong evidence for patient 3, suggesting that parallel mutations may occur along with lost mutations and that there may be multiple violations of the infinite sites hypothesis. The genomic positions are according to the hg19 assembly.

tissue type	lifetime stem cell divisions in tissue	# lifetime point mutations in tissue		probability of having recurrent mutations	
		genome (3.23 Gb)	exome (32.3 Mb)	genome	exome
basal cell carcinoma	3.55×10^{12}	8.8×10^{12}	8.8×10^{10}	1	1
leukemia	1.30×10^{11}	3.2×10^{11}	3.2×10^{9}	1	1
osteosarcoma (pelvis)	3.15×10^{6}	7.8×10^{6}	7.8×10^{4}	≈ 1	≈ 1

Supp. Table 6. Probability of having recurrent mutations in different tissue types for the complete genome and the exome. Figures are based on a mutation rate of 0.77×10^{-9} per site per somatic cell division as estimated by (Lynch, 2010), and the estimates of (Tomasetti and Vogelstein, 2015) for the cumulative number of stem cell divisions per lifetime in different human tissues. For fast dividing cells we have more point mutations than positions in the genome. However even for slow dividing cells such as the osteoblastic cells of the pelvis, the probability that at least two mutations hit the same site is still practically 1.

cancer type	dataset		# of selected mutations	# of detected mutations	# of cells	prior recurrence probability
myeloproliferative neoplasm	Hou et al. (2012)		18	712	58	2.0×10^{-4}
kidney	Xu et al. (2012)		35	35	17	2.7×10^{-5}
breast	Wang et al. (2014)		40	40	47	2.5×10^{-5}
		patient				
leukemia	Gawad et al. (2014)	1	20	20	111	6.3×10^{-6}
		2	16	16	115	4.0×10^{-6}
		3	49	49	150	3.8×10^{-5}
		4	78	78	143	9.5×10^{-5}
		5	105	105	96	1.7×10^{-4}
		6	10	10	146	1.6×10^{-6}
		patient				
ovarian	McPherson et al. (2016)	2	37	7009	588	4.1×10^{-5}
		3	60	5576	672	5.2×10^{-5}
		9	37	3577	420	2.1×10^{-5}

Supp. Table 7. Lower bound estimates for the probability of having a recurrent mutation among the selected set of mutations for the three exome sequencing datasets (Hou et al., 2012; Xu et al., 2012; Wang et al., 2014) and for the panel sequencing of single cells from six leukemia patient samples (Gawad et al., 2014) and three ovarian cancers (McPherson et al., 2016).

Supplemental Material

In this supplemental material we first detail the simulation studies showing how our statistical test performs in Supp. Section S.1. Then we examine the inference of parameters and trees for three single-cell exome datasets (Supp. Section S.2), six personalized panels for leukemia and three targeted panels for ovarian cancer (Supp. Section S.3). In particular we exclude the effect of sequencing bias on the Bayes factor for the parallel mutation uncovered in the data of patient 5 of Gawad et al. (2014) in Supp. Section S.5, and the possibility that the parallel mutations are due to allelic dropout in Supp. Section S.6.

On the methodological side we consider the prior probability that a mutation occurs twice at the same position in the genome when they occur uniformly at random. First that any mutation occurs twice across the whole genome or exome in Supp. Section S.7 and second that a recurrence occurs within a selected set of mutations in Supp. Section S.8. Finally we look at the set of trees with a recurrent mutation to characterize and count the relevant trees for testing the infinite sites assumption in Supp. Section S.9.

S.1 Validation on simulated data

To verify our framework for testing the infinite sites hypothesis, we ran a series of detailed simulations. First we checked the type I errors by simulating random mutation trees under the infinite sites assumption with no mutations recurring. For $n = 20$ mutations we generated 100 trees uniformly with up to 120 sampled cells attached for each simulation. We then created a noisy data matrix by adding false positives with a rate of $\alpha = 10^{-5}$ and we considered two false negative rates of $\beta = 10\%$ and $\beta = 20\%$. To create a discrepancy between the simulation and the parameters used for the inference, we randomized the false negative rate with a 10% misspecification (as in Jahn et al., 2016). Finally, 1% of the data was removed and labelled as missing.

The maximally scoring trees were learned and sampled for both the model class where no recurrent mutations are permitted and for the set of all models with one of the 20 mutations duplicated. The BFs are largely unaffected by the different false negative rates or increasing numbers of samples, although there is a slow decrease in the BFs as more cells are sequenced (Supp. Figure 2). More importantly, the overwhelming majority of the BFs are smaller than one. In fact only 5% of the simulations with the higher false negative rate have BFs larger than 1 (and a slightly smaller percentage with $\beta = 10\%$) so that comparing to the null hypothesis significance testing framework, a BF cutoff of 1 can be thought of as translating to a significance level of 5%. Higher cutoffs would then correspond to lower significance levels.

To test type II errors, random mutation trees were simulated with the settings as above but including a recurrent mutation. For low numbers of sampled cells, only a few simulations allowed the recurrence to be detected. However, this rapidly improves as more samples are added before improving more modestly (Supp. Figure 3). There is also a marginal improvement in the detection rate with a lower false negative rate. With 60 samples, corresponding to several samples per mutation, around 60% of the infinite sites breaking recurrences are detected as having a BF above 1 for the simulation with the higher error rate (and a slightly higher rate with $\beta = 10\%$).

From the simulation with no recurrence (Supp. Figure 2), the highest BFs with 60 samples were around 10, while for the simulations with a recurrent mutation 55% were above this mark (Supp. Figure 3). Reducing type I errors to a low level, our statistical test would still be able to detect the majority of infinite sites violations for a reasonable number of cells sequenced per mutation of interest. This ratio of cells to mutations is also in line with the results of Jahn et al. (2016). While a higher sensitivity would be welcome, these simulations show that our statistical test is rather conservative with a high specificity giving us confidence that BFs signaling the violation of the infinite sites hypothesis in real data can be trusted.

Repeating both experiments with $n = 40$ mutations we see similar rate of type I errors (4% when $\beta = 10\%$ and 8% when $\beta = 20\%$, Supp. Figure 4). When a recurrent mutation is present the BFs are slightly more tightly packed giving a slightly higher sensitivity when considering several cells per mutation. For 120 sampled cells the BFs were greater than 1 in 79% of the simulations for $\beta = 10\%$ and 68% for $\beta = 20\%$ (Supp. Figure 5).

A possible contamination of single cell experiments, which can mimic violations of the infinite sites hypothesis, comes from doublet sequencing where two cells are accidentally captured together and sequenced as one. Their effect on the statistical test can be studied in a simulation where pairs of single cells are randomly selected and have their mutations joined. Again 100 trees with no recurrent mutations and with up to 100 attached samples were generated and then pairs of cells combined until 50 doublets were left. In preparing the data matrices, the values of $\alpha = 10^{-5}$, $\beta = 20\%$ with the same misspecification as above and 1% missing data were used. For a fixed total size of 50 cells (including any doublets) the effect of an increasing number of doublets on the BFs is rather marked (Supp. Figure 6a). On one hand, doublet samples do violate the infinite sites assumption and this violation in the data is picked up by the BFs increasing with the doublet rate. On the other hand, this type of violation is a technical problem with single cell experiments and not biologically relevant. However, by modeling the presence of this type of contamination (as discussed in the Methods), we can correct for its signal in the data and filter it completely from the BFs (Supp. Figure 6b). Continuing the plot up to a doublet rate of 100% we observe no real deterioration in the BFs (Supp. Figure 7b), showing that the doublet correction provides a test with a similar specificity to when only single cells are sampled.

Repeating the simulation with higher false negative rates of $\beta = 25\%$ and $\beta = 30\%$ we see similar behaviour (Supp. Figure 8) with a slight narrowing of the range of BFs compared to $\beta = 20\%$ above (Supp. Figure 7b). This correspondingly reduces the type I error rate of observing a BF larger than 1 when there are doublet samples and no recurrent mutations to 4–6% compared to approximately 10% with $\beta = 20\%$.

On the computational side, correcting for doublets is necessarily more intensive (by a factor proportional to the number of mutations considered). To test the sensitivity we repeat the experiment with $\beta = 20\%$ but now also including a recurrent mutation. The BFs are highly insensitive to the presence of doublets (Supp. Figure 9) confirming that modeling them removes their unwanted effects. However, the number of simulations with a BF above 1 is around 48% compared to 56% when there are no doublets, suggesting a slight loss in sensitivity and a higher type II error rate for our test. Despite their great scope for contaminating single cell data these simulations show that our statistical test is robust against such contamination and effective at removing the effect of doublets. The test also remains conservative suggesting that it would provide high confidence for large BFs observed in real experimental data.

For the simulations where we model the presence of doublets, the doublet rate is learnt both under the infinite sites model and the model with a single recurrent mutation. The median estimates inferred from the data are very much in line with the real values used in its generation (Supp. Figure 10). There is some spread in the estimates – the pair of cells merged into a doublet may only differ in a small number of mutations from different lineages (or none) and these mutations might be lost as false negatives. Other signals may also be compensated for by adjusting the tree structure. However the overall correlation between the true and inferred values is high at 95%. More relevant for our model comparison is how the doublet rate changes for the same dataset when considering the infinite sites model or its violation. The resulting plot (Supp. Figure 11) shows that the inferred doublet rates are highly consistent under both model classes. Since some doublet effects may be similar to a recurrent mutation, we observe marginally higher doublet rates under the infinite sites hypothesis, but still correlations of 98% or 99%. This concordance shows that even if the inferred doublet rate differs from the true value, a highly similar value will be used on both sides of the model comparison providing confidence in the resulting BFs.

To examine the effect of selection and the fitness advantage of driver mutations on the tree structure and our test of violations of the ISA, we simulated tumors using the spatial model of Waclaw et al. (2015). 100 simulations were run until the tumors consisted of 10^7 cells for driver advantages of 2, 4 and 6%. The mutation rate was set to the default 0.02 and each mutation has a probability of 10^{-3} of being a driver. From each simulated tumor, 55 cells were sampled uniformly and 10 combined as doublets to leave 50 sampled cells including 5 doublets. Errors were added to the mutation profiles at with a false positive rate of $\alpha = 10^{-5}$, with false negative rates of $\beta = 10\%$ and $\beta = 20\%$ (misspecified as above) and with 1% missing data. Only mutations which occurred in more than one cell were retained and only driver mutations were considered for the tree reconstruction. Since the simulation setting of Waclaw et al. (2015) uses the ISA, this simulated dataset allows us to test the type I error rate as a function of driver advantage. Compared to the

type I error rates of around 10% for the doublet simulation above with uniformly sampled trees and attachments, we find similar, but slightly higher rates of around 13%. The distributions of BFs obtained under this null model (Supp. Figure 12) also have a smaller spread and depend little on the fitness advantage.

S.2 Parameter and tree inference in single-cell exome sequencing data

To test the infinite site hypothesis in real data we ran our statistical test on three single-cell exome sequencing datasets (Hou et al., 2012; Xu et al., 2012; Wang et al., 2014). To avoid any confounding from doublet samples they were modeled and included in the comparison. Treating the cells as a mixture of singlets and doublets, we infer the mixture component assigned to doublets δ and the proportion $\tilde{\delta}$ exhibiting mutations from different lineages in the phylogenetic history. The false negative rates were inferred under the infinite sites model to favor that model wherever possible and then the best tree is learnt under each model class to provide an estimate of the Bayes factor. All the estimates are summarized in Supp. Table 1.

The first tumor data is from a *JAK2*-negative myeloproliferative neoplasm (essential thrombocythemia) (Hou et al., 2012) from which 58 tumor cells were sequenced. For testing the infinite sites hypothesis we focused on a selection of 18 mutations chosen by Hou et al. (2012) as cancer-related. The false positive rate of the single-cell sequencing was estimated to be $\alpha = 6.04 \times 10^{-5}$ while we inferred the false negative rate under the infinite sites hypothesis as $\beta = 0.401$ (with the prior as in Jahn et al., 2016). We use the same value when allowing a recurrent mutation to ensure that we favor the infinite sites assumption wherever possible. The dataset also has 45% missing data and distinguishes between heterozygous and homozygous mutations (modeled following Kim and Simon, 2014; Jahn et al., 2016).

The doublet rate δ is estimated as 1, although this is driven by the attachment of sampled cells lower down in the mutation tree. The more pertinent estimate is $\tilde{\delta}$ of doublets which contain mutations from different lineages (and which cannot be explained by singlet cells). This is around a third which is reasonably elevated. When we infer the best scoring trees for both model classes we obtain a preliminary BF estimate of 13. A more accurate BF estimation is obtained via a full MCMC sampling (Online Methods) providing a value of 30 with a 95% confidence interval of $[27, 34]$. Despite the high error, missing data, and doublet rates in this dataset, this BF represents reasonably strong evidence of a recurrent mutation. Looking at the trees for both model classes (Supp. Figure 13) we observe that the mutation is a parallel one in two different lineages and also that it occurs late in the mutation tree. Apart from the recurrence, the trees are highly similar under both the infinite sites hypothesis and its violation.

In general the exact ordering of mutations within linear stretches of the tree is not strongly identified by the data and can depend on the random occurrences of allelic dropout and missing data. More certain, and more relevant for understanding the tumor's composition, are the branching points and partitioning of mutations into separate lineages. Even with the high error and missing data rates of the Hou et al. (2012) sequencing, and the uncertainty this brings, only the point mutation in the gene *PDE4DIP* changes its lineage when we allow the recurrent parallel mutation in *RETSAT*, aside from the parallel mutation itself. Data for the mutation in *PDE4DIP* is however missing for over 60% of the single cells, making its placement less precise.

The second data set is from a clear cell renal cell carcinoma (Xu et al., 2012). A total of 17 tumor cells were sequenced and 35 mutation sites were detected which are informative for the phylogenetic reconstruction. For these mutations, heterozygous and homozygous mutations were recorded. The data has 22% missing entries and the false positive rate was estimated by Xu et al. (2012) as $\alpha = 2.67 \times 10^{-5}$. The false negative rate was again learnt under the infinite sites assumption allowing for doublets, resulting in a value of $\beta = 0.192$. The trees exhibit a highly linear structure (Supp. Figure 14). The infinite sites hypothesis leads to a branching while a recurrent mutation would suggest a lost mutation. The doublet rate δ of 1 again comes from samples being preferentially attached lower down the tree, while the highly linear tree means that they could nearly all be explained as singlets instead so that $\tilde{\delta}$ is very low or zero. The BF for this data set is somewhat below 1 so there is no evidence for a violation of the infinite sites hypothesis.

The third dataset consists of 47 oestrogen-receptor positive (ER^+) breast cancer tumor cells (Wang et al., 2014) for which 40 mutations occurred in two cells or more. The error and missing data rates

are improved upon the earlier datasets with the false positive rate estimated at $\alpha = 1.24 \times 10^{-6}$ and only 1% missing data. A false negative rate of $\beta = 0.171$ was learnt under the infinite sites hypothesis with doublet modeling. The pertinent doublet rate $\tilde{\delta}$ is however quite elevated and disparate for the two model classes (55% under the infinite sites hypothesis and 28% without). Cells for sequencing were selected as those whose nuclei had divided (to provide twice the starting amount of DNA to amplify) which could be connected to an enrichment in doublet capture. The tree topology under both models consists of a linear chain of mutations on top of a rather branched structure further down (Supp. Figure 15). While the lower parts are essentially identical, the presence of a recurrent mutation changes the upper tree structure compared to the tree under the finite sites model. The BF of 2000 provides very strong evidence that the model with the lost mutation fits the data much better than the infinite sites model.

S.3 Parameter and tree inference in single-cell panel sequencing data

We tested for infinite sites violations in datasets from 6 acute lymphoblastic leukemia (ALL) patients (Gawad et al., 2014). This study provides single-cell sequencing data for a personalized panel for each patient. Since these data come without direct estimates of the error rates, we inferred both α and β assuming the infinite sites assumption holds and accounting for doublets. To learn the false positive rate α, we included all genomic positions covered by the panel, including those where no mutation was detected in any of the sequenced cells. Only the detected mutations inform the tree and the estimate of the false negative rate β. The values are summarized in Supp. Table 2 where we note that since the size of the total data matrix is in the order of 10^5 we cannot reliably infer false positive rates below 10^{-4}. Bearing this in mind we find more elevated false positive rates than for the exome data (Supp. Table 1), but still low in absolute numbers. Of course since the error rates are inferred under the infinite sites hypothesis, the infinite sites model is favored regardless of the error rate values. The false negatives are in the 15–30% range in line with the previous exome data, indicating that these errors mostly arise from the amplification process before sequencing.

The learnt doublet rates δ are high again (up to 100%) suggesting that samples preferentially attach lower in the tree, while the percentage of data $\tilde{\delta}$ involving mutations from different lineages and actually requiring doublet samples to explain is in the 0–20% range. This is high in one sense, but not unexpected for the microfluidic cell sorting employed in Gawad et al. (2014).

Turning to the actual test of the infinite sites assumption, we found that the data suggests it is heavily violated for all patient samples with Bayes Factors in the range of 10^5 to 10^{15} (Supp. Table 2), apart from patient 2 with a more modest preliminary BF estimate of 300. A more accurate estimate of 326 with a 95% confidence interval of $[305, 350]$ is obtained with the full MCMC sampling. The trees (Supp. Figures 16–21) show that for three patients the lost mutation is actually the first to arise in their trees while the parallel mutation in Supp. Figure 20 shows the two copies of the mutation occurring at the end of different lineages.

We further tested for infinite sites violations in the three single-cell ovarian cancer datasets studied by McPherson et al. (2016). From targeted single-cell sequencing, each genomic position in the panel was tested as to whether the reference or variant allele could be reliably detected. This allows us to distinguish between heterozygous and homozygous mutations (modeled following Kim and Simon, 2014; Jahn et al., 2016). The panels designed by McPherson et al. (2016) included ancestral mutations predicted to be present in normal tissue, but lost in the tumor or its main clones, to mark normal cells in the sample. Since including known mutational losses may confound testing the ISA, we ran the analysis excluding all mutations present in the normal cells. These data come without direct estimates of the error rates, so we inferred both α and β under the infinite sites model while accounting for doublets. The panels focus on mutations present in many of the cells, limiting the number of true negatives to inform the inference of the false positive rate α. The corresponding values (Supp. Table 4) are therefore elevated compared to the leukemia panel data, but being inferred under the infinite sites assumption can only favor that model. In modeling both heterozygous and homozygous mutations, the error rate β includes allelic dropout of the reference allele (which leads to the detection of homozygous mutations). Halving its value to only consider dropout of the variant allele, to compare to the false negative rate of the binary datasets, we find comparable values of around 20%. The percentage of data $\tilde{\delta}$ involving mutations from different

lineages and requiring doublets to explain is around 20% for patient 2 and below 10% for patient 9. For patient 3 however we observe a discrepancy between the two model classes, which is also reflected by differences in the tree structures (Supp. Figure 23), but the large BF also indicates that the infinite sites model is not appropriate for this data.

To check that the inclusion of mutations with known LOH does not overly affect our test of the ISA, we reran the analysis including all mutations from the normal cells. In all cases the highest scoring recurrent mutations on the subset remained the highest scoring recurrence on the full set of mutations in the panel. The results (Supp. Table 4) are in line with the previous analysis but for patient 9 the recurrent mutation changes from being a parallel mutation to a lost one.

S.4 Survey of LOH in ALL

We uncovered loss of mutations in 5 out of the 6 leukemia patients Gawad et al. (2014), which could be the result of LOH or back mutations of the individual bases. We therefore examined whether LOH at the loci we identified are common in ALL. To obtain such statistics, we performed a comparison with large scale (>5Mb) copy number deletions found in a large study of 142 children and 123 adults with ALL (Forero-Castro et al., 2016). Patients 1 and 2 had lost mutations on the 8q Chromosome, which were not observed in any of the study samples, except a whole loss of Chromosome 8 in one adult. Patient 3 had a lost mutation at Chromosome 2q, which also was not observed in any of the study samples. The lost mutations for patients 1–3 therefore also do not seem to match common large-scale deletion events, but could be the result of smaller losses. The 8q mutational loss of patient 1 and 2 are close to *MYC* which plays an important role in ALL (Forero et al., 2013). Patient 4's lost mutation was at 8p which was lost in 4 children and 4 adults, while patient 6's lost mutation was at 22q and Chromosome 22 was subject to large-scale deletion for 4 children and one adult in the ALL sample (Forero-Castro et al., 2016). Their large BFs could be related to these relatively common LOH events. Patient 6's lost mutation also happened to be near *IGL* which is rarely translocated with *MYC*.

S.5 Ruling out bias in the parallel mutation in patient 5

When testing the infinite sites hypothesis, the error rates are inferred under the infinite sites model, favoring that model. This excludes the possibility that overall inaccuracies in the error rate inference would suggest a violation in the infinite sites assumption, however it is conceivable that certain genomic positions have markedly different error profiles. Here we examine this possibility for the parallel mutation uncovered in patient 5 of the Gawad et al. (2014) data, which occurred on Chromosome 8 at position 172434428 (hg19) in the gene *C1orf105*. Examining the trees in Supp. Figure 20, the large improvement in fitting the data with a recurrent mutation could perceivably be the result of bias in the sequencing of that position in two ways: the recurrent mutation could be a trunk mutation with a very large false negative rate, or the mutation may appear to occur in two lineages due to a significant number of false positives.

To check these possibilities we extracted the coverage and number of reads supporting the variant at that genomic position for each cell sequenced by Gawad et al. (2014). Since allelic dropout should affect both alleles equally we set the number of false negatives with no supporting reads to be equal to the number of cells where all the reads supported the variant. Apart from these 15 cases, all cells with no variant reads were assigned as true negatives. The coverage in the cells with variant supporting reads was above 100 for all but 4 cells (which happened to have variant frequencies above 35%). Even assigning cells with variant frequencies below 10% (for coverages above 100) as negative would only result in a false negative rate of 0.268 which is still below the estimate rate of 0.291 for this data set. We can therefore safely conclude that there is no elevated false negative rate for the parallel mutation.

For the false positives we remove all cells with a variant frequency of 0% and 100% and assign everything else as evidence of the mutation. We look at the distribution of variant frequencies (Supp. Figure 25(a)). If all the positives start with one mutated allele which is subject to symmetric dropout, under amplification and sequencing errors we would expect to see a roughly symmetric variant frequency distribution. False positives on the other hand would start with no copies of the mutation and would lead to an excess in low frequency variants. In the histogram (Supp.

Figure 25(a)), this is not directly evident, but to amplify the signal we calculate the kernel density estimate $p(f)$ of the variant frequencies f and plot $p(f) - p(1-f)$ in Supp. Figure 25(b). This means subtracting the right hand side of the distribution from the left so that any systematic excess in low frequency variants would lead to a positive value for low f. Instead in the plot (Supp. Figure 25(b)) we see a small lack of low frequency variants and slightly fewer than we would expect from a symmetric process. Of course this asymmetry is allowed under the stochastic amplification and sequencing process, but this result certainly excludes the possibility of false positives affecting the strong evidence of a recurrent mutation at this genomic position.

When finding strong evidence for the parallel mutation in patient 5 of the Gawad et al. (2014) data, we model doublets and correct for their potential contamination. In the original publication, Gawad et al. (2014) clustered the mutations profiles of the single cells, without enforcing a phylogeny or the ISA, and then constructed a clonal history. Eight cells which appear to share mutations from clones in different lineages were then manually marked as doublets. As a final check we therefore recalculated the BF when these cells were excluded from the data. We still find a very high BF of 2.2×10^{13}, corroborating that the parallel mutation is not an artefact of doublet contamination.

S.6 Excluding the possibility of apparent parallel mutations from false negatives

One concern is that the high rates of allelic dropout in single-cell sequencing data can mimic signals of parallel mutations. To examine this possibility, we consider that false negatives are much more probable than false positives, $\alpha \ll \beta$, and we assume that the ISA holds. We can then ask, under the ISA, what is the probability that the parallel mutations detected could have arisen solely due to allelic dropout or false negatives.

For the parallel mutation of interest, M_X, we need to consider the cells which possess the mutation M_X and so attach below the mutation, but which do not possess any of its descendants in the tree under the ISA. These cells may potentially attach to a parallel copy of M_X, and we denote the number of such cells as m_X. Mutations present in any of its descendants of M_X would become false positive signals, were the cells with those mutations placed in a different lineage, and are therefore excluded from consideration.

Now we consider a second copy of M_X placed in a parallel lineage which branches off a distance u above M_X and with v mutations between the branch point and the second copy. If a cell which can attach below the original M_X misses all the u mutations between M_X and the branch point, then it may also be placed at or below the second copy of M_X without incurring any false positives. If k out of the m_X cells miss all u mutations, we can compute their contribution to the ratio of tree scores used for computing the BF in Equation (26) of the main text. First, if the mutation M_X has d_i descendants at a relative depth i, then the score of the infinite sites model tree includes the term

$$\prod_{j=1}^{k} L_j \left[1 + \sum_i d_i \left(\frac{\beta}{1-\beta} \right)^i \right] \tag{1}$$

where L_j is the score of cell j attached at mutation M_X and the sum accounts for attaching the cell to all descendants which become false negatives. Attaching the cell above M_X would make that mutation a false positive so that such terms are excluded by the approximation $\alpha \ll \beta$.

These cells will contribute the following to the score of the tree under the finite sites model

$$\prod_{j=1}^{k} L_j \left\{ 1 + \sum_i d_i \left(\frac{\beta}{1-\beta} \right)^i + \left(\frac{\beta}{1-\beta} \right)^{v-u} \left[1 + \sum_i d'_i \left(\frac{\beta}{1-\beta} \right)^i \right] \right\} \tag{2}$$

where d'_i are the number of descendants at relative depth i for the second parallel copy of M_X. The ratio of tree scores in Equation (26) of the main text is then replaced by the ratio of the scores of these k cells

$$\left(\frac{1 + \sum_i d_i \left(\frac{\beta}{1-\beta} \right)^i + \left(\frac{\beta}{1-\beta} \right)^{v-u} \left[1 + \sum_i d'_i \left(\frac{\beta}{1-\beta} \right)^i \right]}{1 + \sum_i d_i \left(\frac{\beta}{1-\beta} \right)^i} \right)^k \tag{3}$$

For each parallel mutation which we uncover in real data, we compute the minimum number of cells, k^*, needed to provide a BF larger than 1 utilizing Equation (26) of the main text. We then compute the probability that k^* or more cells out of the m_X lose all u mutations back to the branch point to also fit along the second lineage. The probability for each cell is $p = \beta^u$ and we perform a binomial test with m_X number of trials, probability p of success and a threshold of k^*.

For data where both heterozygous and homozygous mutations are recorded, we replace the $\frac{\beta}{1-\beta}$ by $\frac{\beta}{2(1-\beta)}$ in Supp. Equation (3) and set $p = \left(\frac{\beta}{2}\right)^u$.

For the data from the *JAK2*-negative myeloproliferative neoplasm (Hou et al., 2012), the parallel mutations are both at the end of lineages (Supp. Figure 13) so that $d_i = d_i' = 0$ and we have to compute the smallest k such that

$$\left[1 + \left(\frac{\beta}{2(1-\beta)}\right)^{v-u}\right]^k - n\mathrm{e} > \frac{\tilde{K}_\mathrm{F}}{K_\mathrm{I}} \tag{4}$$

Substituting $u = 4$, $v = 2$ from Supp. Figure 13 and $\beta = 0.455$ among with the other values from Supp. Table 1 we find $k^* = 5$. A total of $m_X = 10$ cells possess the *RETSAT* mutation and we compute the probability that 5 or more cells out of the 10 lose all 4 mutations from *RETSAT* to the branching point. From the binomial test with a probability $p \approx 2.7 \times 10^{-3}$ of success, the probability of the parallel signal for the myeloproliferative neoplasm (Hou et al., 2012) coming just from false negatives is 3.4×10^{-11}. We may further consider a type of multiple testing correction and multiply by n to arrive at 6.2×10^{-10}.

For patient 5 of the leukemia dataset of Gawad et al. (2014), we have very long branch lengths of $u = 35$ and $v = 18$ in Supp. Figure 20 while $m_X = 20$ cells possess the parallel mutation in the gene *C1orf105*. Due to the long lengths, only 1 cell would need to lose all 35 mutations with success probability $p = 1.7 \times 10^{-19}$ in the binomial test giving a result of 3.4×10^{-18} or a multiply corrected version of 3.6×10^{-16}. For patient 2 of the ovarian cancer dataset of McPherson et al. (2016), if we take the position of the parallel mutation in the gene *AC004538.3* from the infinite sites model (Supp. Figure 22(a)), there are no cells which do not possess any of the descendant mutations in the tree and which could move to a second lineage without creating false positives. Instead we can consider the placement under the finite sites model in Supp. Figure 22(b) leading to $u = 14$, $v = 5$ and $m_X = 99$. We would need at least 4 cells to lose all 14 mutations back to the branch point, each with a probability of success of 1.3×10^{-9}. When substituted into the binomial test, we obtain a result of 1.2×10^{-29} with a multiply corrected version of 4.6×10^{-28}.

For each dataset where we found a clear parallel mutation, it is therefore highly unlikely that the signal of recurrence could come just from false negatives. However, depending on the exact tree structure, the attachment of cells and the false negative rate, situations could arise where spurious signals of parallel mutations may be possible. Examining the formula in Supp. Equation (3), we would need a large number of cells which attach at the mutation and a low distance u of the second copy from the original lineage. A moderately low distance u from the original copy of M_X to the branch point would also be needed, although the case $u = v = 0$ is excluded from the finite sites model since it recreates the infinite sites model (Supp. Section S.9). In general, results with low u and v should be checked following the reasoning here.

S.7 Quantitative survey of somatic cell evolution refutes the infinite sites hypothesis

The idea that every genomic position mutates at most once over the lifetime of a human is easy to disprove when looking at the lifetime mutation counts in different tissues. Using the estimate for the mutation rate from Lynch (2010) and the cumulative number of stem cell divisions per lifetime in different human tissues fromTomasetti and Vogelstein (2015), we find that for fast-dividing cell types such as skin and blood cells, the number of life-time mutations is about three times the number of genomic positions necessitating the recurrence of many mutations. For the slowest dividing tissue type, the osteoblastic cells of the pelvis, we find that the total number of mutations makes up only 0.24% of the number of genomic sites. The probability of at least two mutations hitting the same site is still practically 1 (Supp. Table 6). This counter intuitive result can be obtained from a generalization of the birthday problem, a popular math puzzle which asks for the probability that two people in a group share the same birthday.

Let Z be the number of positions in a genome, and D the number of life-time cell divisions in a tissue, then the number of life time mutations N in this tissue can be calculated as

$$N = \mu D Z \tag{5}$$

for a given mutation rate μ. Assuming that every site is equally likely to mutate, the probability of drawing at least one of Z sites more than once in N draws can be approximated as

$$P_{\mathrm{F}} \approx 1 - \mathrm{e}^{\frac{-N^2}{2Z}} \tag{6}$$

Setting the mutation rate and the number of genomic positions gives the estimates in Supp. Table 6, while restricting the observable mutations to the exome changes the figures slightly, but the probability of a recurrent mutation is still practically 1 for all cancer types. This method of estimation is even on the conservative side as any differences in the mutation rates of the individual sites would further increase the likelihood of hitting the same site multiple times.

While the infinite sites assumption does not hold for a life-time of somatic cell evolution, we also want to understand whether recurrent mutations are likely to be observed in the sets of mutations detected in single-cell data.

S.8 Probabilities of recurrence among a set of mutations

We again consider a set of N mutations which occur somewhere along the Z possible base pair locations and calculate the probability that a mutation will reoccur when they are distributed uniformly across the genome. Any non-uniformity in the distribution will simply increase the probability of repeats.

For the computations, we will consider two cases. In the first we only examine a small subset of size n of all the N mutations observed. In the second we will examine the probability of recurrence when we consider all the observed mutations $n = N$. Since experimentally the mutations studied are selected since they have occurred, we actually need to calculate the probability of a duplication among the n selected mutations, given that we've seen all n selected mutations at least once. For this we start with the probability of obtaining all the n selected mutations.

Probability of observing all selected mutations To start we look at the case that k out of the total N mutations occurred among the n selected. The number of ways of filling up all n sites at least once is simply the Stirling number of the second kind

$$\left\{ {k \atop n} \right\} = \frac{1}{n!} \sum_{i=0}^{n} (-1)^i \binom{n}{i} (n-i)^k \tag{7}$$

giving a probability

$$\frac{n! \left\{ {k \atop n} \right\}}{n^k} \tag{8}$$

The total probability of having all n mutations filled is then

$$P_n = \sum_{k=0}^{N} \binom{N}{k} n! \left\{ {k \atop n} \right\} \frac{(Z-n)^{N-k}}{Z^N} = n! \left(\frac{Z-n}{Z} \right)^N \sum_{k=0}^{N} \binom{N}{k} \left\{ {k \atop n} \right\} z^k \tag{9}$$

with $z = \frac{1}{(Z-n)}$.

To tackle the central sum

$$H(z) = \sum_{k=0}^{N} \binom{N}{k} \left\{ {k \atop n} \right\} z^k \tag{10}$$

we use the *snake oil* method in combinatorics (Wilf, 2013) of introducing another generating variable

$$G(\lambda, z) = \sum_{n=0}^{\infty} \lambda^n \sum_{k=0}^{N} \binom{N}{k} \left\{ {k \atop n} \right\} z^k = \sum_{k=0}^{N} \binom{N}{k} z^k \sum_{n=0}^{k} \left\{ {k \atop n} \right\} \lambda^n \tag{11}$$

and treat z as arbitrary without keeping track of its n dependence. The n-th coefficient of the expansion of $G(\lambda, z)$ with respect to λ still recovers $H(z)$ with the correct n dependence in z:

$$[\lambda^n]G(\lambda) = H(z) \tag{12}$$

To perform the sum over n we use the following relation

$$\sum_{n=0}^{k} \left\{ {k \atop n} \right\} \lambda^n = \mathrm{E}\left[X^k\right] \tag{13}$$

where X is a Poisson random variable with parameter λ. This means that

$$G(\lambda, z) = \mathrm{E}\left[\sum_{k=0}^{N} \binom{N}{k} z^k X^k\right] = \mathrm{E}\left[(1+zX)^N\right] \tag{14}$$

Using the pmf of the Poisson gives

$$G(\lambda, z) = \sum_{k=0}^{\infty} (1+zk)^N \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \tag{15}$$

Next we Taylor expand the exponential

$$G(\lambda, z) = \sum_{k=0}^{\infty} (1+zk)^N \frac{\lambda^k}{k!} \sum_{l=0}^{\infty} \frac{(-\lambda)^l}{l!} \tag{16}$$

and collect all terms where λ has the power n

$$H(z) = \frac{1}{n!} \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (1+zk)^N \tag{17}$$

Finally the required probability is

$$P_n = \left(\frac{Z-n}{Z}\right)^N \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (1+zk)^N \tag{18}$$

Conditional probability of no recurrence With the probability of having all n selected mutations present, we can now calculate the conditional probability that at least one of them reoccurs. More simply we can work out the conditional probability that there are no duplications, which can only occur if $k = n$ and then in $n!$ ways. This is simply the $k = n$ term in the sum in Supp. Equation (9)

$$n! \left(\frac{Z-n}{Z}\right)^N \binom{N}{n} z^n \tag{19}$$

The conditional probability of no recurrence given all n were present is then this term divided by P_n or

$$P_{\mathrm{I}} = \frac{[N]_n \, z^n}{\sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} (1+zk)^N} \tag{20}$$

with $[N]_n$ the falling factorial. The conditional probability of recurrence is one minus this value, $P_{\mathrm{F}} = 1 - P_{\mathrm{I}}$.

Approximation If $z \ll 1$ then only the $k = n$ term and the subsequent few will play any role in the sum in Supp. Equation (9). Taking just the first two, the probability of no recurrence is

$$P_{\mathrm{I}} \approx \frac{\binom{N}{n} \left\{ n \atop n \right\} z^n}{\binom{N}{n} \left\{ n \atop n \right\} z^n + \binom{N}{n+1} \left\{ n+1 \atop n \right\} z^{n+1}} = \frac{1}{1 + \frac{n(N-n)z}{2}} \tag{21}$$

and with the second term (necessarily) much smaller than the first, the probability of recurrence is simply

$$P_{\mathrm{F}} \approx \frac{n(N-n)}{2(Z-n)} \approx \frac{n(N-n)}{2Z} \tag{22}$$

So far this calculation has assumed that we start with N mutations spread uniformly across the Z positions and asks if any coincide. But if two were to coincide, we would observe $(N-1)$ mutations instead. Since for real data we know the number of observed mutations, not the actual number of mutations, we should modify the calculation appropriately.

Treating all observed mutations For most of the datasets we actually take the complete set of mutations, and the previous formula simply gives 0. If n mutations occurred in n positions then there is no chance of a duplication. Of course if a mutation had reoccured we would still observe the same set of n mutations and not notice the recurrence in that set. However we would have started with $(n+1)$ mutations, so we therefore need to consider starting with more mutations that happen to coincide.

We recall that if we have k mutations, the probability of observing all n selected mutations at least once, and no other mutations, is

$$n! \left\{ k \atop n \right\} \frac{1}{Z^k} \tag{23}$$

Now though, we need the probability of having k mutations to start with. Since the mutation rate is low, and we have many cell divisions, a Poisson distribution offers a good approximation, with the parameter set to the number of observed mutations n. The total probability of having all n mutations filled, and no others follows as

$$P_n = n! \sum_{k=0}^{\infty} \left\{ k \atop n \right\} \frac{n^k}{Z^k} \frac{\mathrm{e}^{-n}}{k!} = \left(\mathrm{e}^{\frac{n}{Z}} - 1\right)^n \mathrm{e}^{-n} \tag{24}$$

Given that we observe all n selected mutations, we can now calculate the conditional probability that at least one of them reoccurs. The conditional probability that there is no recurrence is just the $k = n$ term in the sum in Supp. Equation (24) giving

$$P_{\mathrm{I}} = \frac{\left(\frac{n}{Z}\right)^n \mathrm{e}^{-n}}{\left[\mathrm{e}^{\frac{n}{Z}} - 1\right]^n \mathrm{e}^{-n}} = \left[\frac{Z}{n}\left(\mathrm{e}^{\frac{n}{Z}} - 1\right)\right]^{-n} \tag{25}$$

Since $\frac{n}{Z} \ll 1$ we can expand the exponential

$$P_{\mathrm{I}} = \left[\frac{Z}{n}\left(\frac{n}{Z} + \frac{n^2}{2Z^2} + \ldots\right)\right]^{-n} = \left[1 + \frac{n}{2Z} + \ldots\right]^{-n} \tag{26}$$

and then the binomial

$$P_{\mathrm{I}} = 1 - \frac{n^2}{2Z} + \ldots \tag{27}$$

The conditional probability of recurrence becomes

$$P_{\mathrm{F}} \approx \frac{n^2}{2Z} \tag{28}$$

If we do not select the mutations beforehand, but choose those which appear, the intermediate probabilities change, but the conditional probabilities do not.

Treating a subset of all observed mutations Finally we can calculate the probability of recurrence when we observe N mutations and focus on just n of them. Taking Supp. Equation (28), the probability of a recurrence among the N is approximately $\frac{N^2}{2Z}$ while the chance that the recurrence occurs among the n is simply $\frac{n}{N}$ directly giving

$$P_{\mathrm{F}} \approx \frac{nN}{2Z} \tag{29}$$

Estimate for the different datasets The datasets considered involved whole exome sequencing of $Z \approx 3.2 \times 10^7$ positions. The first data set of Hou et al. (2012) found 712 SNVs of which we focussed on 18. Substituting these numbers into Supp. Equation (29) gives $P_{\mathrm{F}} \gtrsim 2.0 \times 10^{-4}$. This might be considered as the lower end estimate. To ensure high confidence in the calling, mutations were restricted to those found in at least 5 of the single cells. Also only about 70% of the exome was covered to a target depth of 5 and 58% to a depth of 20 while SNVs were only called if they had a support of 10×. This helps to explain the large amount of missing data in Hou et al. (2012) for sites where statistical power was lacking to determine mutational status, but also suggests that many more mutations might have occurred during the tumor evolution and which would increase P_{F}.

For the remaining exome datasets we took the full set of mutations detected (with confidence) so we employ Supp. Equation (28). For the data from Xu et al. (2012) along with the 35 mutations informative for the tree reconstruction, 15 mutations were found to occur in all the sampled cells, giving the estimate $P_{\mathrm{F}} \approx 2.7 \times 10^{-5}$. From the 40 mutations detected in at least two cells in Wang et al. (2014), we obtain the estimate $P_{\mathrm{F}} \approx 2.5 \times 10^{-5}$. The estimates for the whole exome data are also recorded in Supp. Table 7. Again these estimates would increase if mutations were missed in any of the cells or the mutations occur non-uniformly.

We also considered the 6 patient samples panel sequenced by Gawad et al. (2014). Although the panels were much shorter than the whole exome, the panels were designed based on mutation signal present in whole-exome bulk sequencing so essentially Z remains the same size. The panels detected different numbers of mutations giving the results summarized in Supp. Table 7. Finally for the panel sequencing of 3 ovarian cancers by McPherson et al. (2016), 43, 84 and 43 mutations were targeted from the 7009, 5576 and 3577 high-confidence mutations detected among the whole genome of the bulk samples, corresponding to $Z \approx 3.2 \times 10^9$ positions. Of the mutations in the panel, we excluded those present in the normal cells, leaving a total of 37, 60 and 37 and leading to the estimates in Supp. Table 7.

S.9 Counting trees with recurrent mutations

To perform the model comparison between trees respecting the infinite sites hypothesis and those with a recurrent mutation and the sum over trees in the Bayes factor

$$B_{\mathrm{FI}} = \frac{K_{\mathrm{I}}}{K'_{\mathrm{F}}} \left[\frac{\sum_i \sum'_{T \in \mathcal{M}_i} s(T)}{\sum_{T \in \mathcal{M}_{\mathrm{I}}} s(T)} \right] \tag{30}$$

we need to understand which tree structures with a duplication can mimic infinite sites trees. In particular we wish to avoid possibilities where the recurrence does not provide any additional mutation patterns for attached samples and remove them as indicated by the primes in Supp. Equation (30). Accounting for these cases provides the bound in Equation (24) of the main text:

$$B_{\mathrm{FI}} \geq \frac{K_{\mathrm{I}}}{\tilde{K}_{\mathrm{F}}} \left[\frac{\sum_i \sum_{T \in \mathcal{M}_i} s(T)}{\sum_{T \in \mathcal{M}_{\mathrm{I}}} s(T)} - n^2 \mathrm{e} \right] \tag{31}$$

An obvious possibility is where the original mutation and its repetition share the same parent. Then any subtrees attached to either can be freely moved between them without affecting the mutation pattern the tree predicts. The pattern is also identical to the case when the recurrent mutation is removed. Likewise if there is a direct parent-child relationship between the original and the duplication. Descendants of the pair would not exhibit the mutation due to the lost mutation and could achieve the same mutation pattern by being placed above them both.

Trees where mutations share a parent or are in a parent-child relationship For each tree with n nodes without a recurrent mutation, we therefore wish to know the number of ways of adding the duplication so it shares the same parent as the original mutation, or where it becomes its parent or child. The number of trees where the original mutation has degree k is

$$\binom{n-1}{k-1} n^{n-k} \tag{32}$$

since there are $(n+1)$ nodes including the root. Of the k links to the original mutation, one is on the path to the root and the remaining $(k-1)$ are descendant subtrees. First we detach these subtrees and then we can add the duplication in three different ways: as a sibling of the original mutation, as its new parent along the path to the root, or as its child. Finally we can reattach each of the $(k-1)$ subtrees to either the original mutation or its duplication in 2^{k-1} ways. The total number of trees follows as

$$3\sum_{k=1}^{n} 2^{k-1} \binom{n-1}{k-1} n^{n-k} = 3(n+2)^{n-1} \tag{33}$$

These trees need to be subtracted from the set of all $(n+2)^n$ trees with a recurrent mutation. We also divide by 2 to account for swapping the labels of the original mutation and its duplication. Finally the m sampled cells may be attached to any of the mutations (including the root and the recurrent mutation) giving a total number of possibilities of

$$K_i = \frac{1}{2}(n-1)(n+2)^{n-1+m} \tag{34}$$

Trees with no samples downstream of the recurrent mutation Along with the cases given above where the tree with a recurrent mutation would always recreate mutation patterns achievable under the infinite sites assumption, there will be further possibilities when sampled cells are placed so as to not cover the full set of possible patterns. In the extreme case, for example, that all samples are attached to the root, all trees with a recurrent mutation will recreate the mutation pattern which also come from any tree respecting the infinite sites hypothesis.

In particular, for any tree where the recurrent mutation or its original copy have no samples attached downstream, removing one of the copies would provide the same mutation patterns without needing to violate the infinite sites hypothesis.

To count such possibilities we start with the case where the original mutation may have descendant samples, but the duplication certainly does not. For now we exclude the original mutation from being a descendant of the duplication, which we let have k other mutations as descendants. These are selected from the remaining $(n-1)$ mutations and built into $(k+1)^{k-1}$ trees with the duplication as the root. The remaining $(n-k)$ mutations form a core tree, with an additional root, in $(n-k+1)^{n-k-1}$ ways. The m samples can be attached in $(n-k+1)^m$ ways. The tree with the duplication as the root can then also be attached to any of the core nodes, apart from the original mutation or its parent, in $(n-k-1)$ ways giving a total of

$$\sum_{k=0}^{n-1} \binom{n-1}{k} (n-k-1)(n-k+1)^{n-k-1+m}(k+1)^{k-1} \tag{35}$$

possibilities.

If we were to swap the labels of the duplication and the original mutation, we would obtain all the trees where the original mutation has no descendant samples. However trees where the original mutation and the duplication are in different lineages, and where neither have descendant samples, will appear both times and hence are currently counted twice. To remove the extra copy we build such trees by sharing the k descendants between the original mutation and the duplication. If i of the k become descendants of the original mutation and the remaining $(k-i)$ descendants of the duplication, then

$$\sum_{i=0}^{k} \binom{k}{i} (k-i+1)^{k-i-1}(i+1)^{i-1} = 2(k+2)^{k-1} \tag{36}$$

pairs of trees can be built and attached to the core tree in $(n-k)(n-k-1)$ ways, because they may not share a parent. The resulting sum is

$$2\sum_{k=0}^{n-1}\binom{n-1}{k}(n-k-1)(n-k)^{n-k-1+m}(k+2)^{k-1} \tag{37}$$

Now we can subtract half of Supp. Equation (37) from Supp. Equation (35), and we need to evaluate

$$S=\frac{1}{n}\sum_{k=0}^{n}\binom{n}{k}(n-k)(n-k+1)^{n-k-1+m}(k+1)^{k-2}\,(n-3k-1) \tag{38}$$

where we shifted the index of Supp. Equation (37) and rearranged.

The formula in Supp. Equation (38) is the exact number of trees we wish to exclude when testing trees which violate the infinite sites hypothesis. For calculating the BFs in the model comparison we would need to exclude all of them from our MCMC scheme which is computationally prohibitive since it would interfere with the independence used for speeding up the marginalization. Instead it is simpler to retain such trees (apart from the basic cases above) and to remove their contributions to the BFs separately. Their contributions appear in two places in Supp. Equation (30), once in the numerator (weighted by their scores) and once in the denominator. For the denominator we only need their number and we can use Supp. Equation (37) directly. This is unwieldy to calculate for large trees, so instead we can find a simple lower bound which is easier to calculate. Any inexactness will favor the infinite sites hypothesis

A lower bound for S We separate out the $k=0$ term and consider the sign of the rest of the sum in Supp. Equation (38). We split the final factor into two parts: $(n-2k)$ and $-(k+1)$. For the second part we have

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{n-1}\binom{n}{k}(n-k)(n-k+1)^{n-k-1+m}(k+1)^{k-1} \\
&<n^{m-1}\sum_{k=1}^{n-1}\binom{n}{k}(n-k)(n-k+1)^{n-k-1}(k+1)^{k-1} \\
&=n^m\left[(n+2)^{n-1}-(n+1)^{n-1}\right]
\end{aligned} \tag{39}$$

For the $(n-2k)$ terms, for $1<k<\frac{n}{2}$ we combine the k-th term and the $(n-k)$-th term

$$\begin{aligned}
&\binom{n}{k}(n-k)(n-k+1)^{n-k-1+m}(k+1)^{k-2}\left(1-\frac{2k}{n}\right) \\
&+\binom{n}{n-k}k(k+1)^{k-1+m}(n-k+1)^{n-k-2}\left(1-\frac{2(n-k)}{n}\right) \\
&=\binom{n}{k}\left(1-\frac{2k}{n}\right)\Big[(n-k)(n-k+1)^{n-k-1+m}(k+1)^{k-2} \\
&\qquad -k(k+1)^{k-1+m}(n-k+1)^{n-k-2}\Big]
\end{aligned} \tag{40}$$

Since $k<\frac{n}{2}$ we have

$$\begin{aligned}
\frac{n-k+1}{k+1}&>1 \\
\left(\frac{n-k+1}{k+1}\right)^{m+1}&>1 \\
(n-k+1)^{m-1}(k+1)^{-2}&>(k+1)^{m-1}(n-k+1)^{-2} \\
(n-k+1)^{n-k-1+m}(k+1)^{k-2}&>(k+1)^{k-1+m}(n-k+1)^{n-k-2} \\
(n-k)(n-k+1)^{n-k-1+m}(k+1)^{k-2}&>k(k+1)^{k-1+m}(n-k+1)^{n-k-2}
\end{aligned} \tag{41}$$

and that the terms in Supp. Equation (40) are positive making that part of Supp. Equation (38) also positive in total.

The next term to consider is for $k = 1, n-1$ whose combination is

$$(n-2)\left[\frac{1}{2}(n-1)n^{n-2+m} - 2^m n^{n-3}\right] \tag{42}$$

Subtracting the bound in Supp. Equation (39) leaves

$$\frac{1}{2}(n-1)(n-2)n^{n-2+m} - (n+2)^{n-1}n^m + (n+1)^{n-1}n^m - (n-2)2^m n^{n-3} \tag{43}$$

Taking out a factor of n^m tidies this to

$$\frac{1}{2}(n-1)(n-2)n^{n-2} - (n+2)^{n-1} + (n+1)^{n-1} - (n-2)n^{n-3}\left(\frac{2}{n}\right)^m \tag{44}$$

which is positive for $n > 7$.

Since the terms for $k > 0$ sum to a positive integer for $n > 7$ the whole sum Supp. Equation (38) is larger than its $k = 0$ term or

$$S > (n-1)(n+1)^{n-1+m} = (n-1)K_{\mathrm{I}}\,, \qquad n > 7 \tag{45}$$

This bound can be intuitively understood by starting from an infinite sites tree where for $m \gg n$ we would expect each mutation to have descendent samples. The only place to add the duplication would be to attach it below a mutation node in the tree so it remains without any attached or descendent samples. Since the duplication may not be added to the original copy or its parents, this can be achieved in $(n-1)$ ways for each of the K_{I} infinite sites trees.

From the result in Supp. Equation (45) we have

$$K'_{\mathrm{F}} \leq n\left[K_i - (n-1)K_{\mathrm{I}}\right] = \tilde{K}_{\mathrm{F}} \tag{46}$$

providing one component of the bound in Supp. Equation (31)

An upper bound For the numerator in Supp. Equation (30) we need not only the number of trees which recreate infinite sites patterns, but also weighted by their scores. Since this is highly complicated, we take a different tack. For each tree respecting the infinite sites hypothesis, we find the maximum number of trees with a recurrent mutation that create the same data matrix and likelihood. The sum of all the scores of the trees with a duplication will then be less than this maximum number times the sum of infinite sites trees.

The direct possibilities where the duplication has the same parent as the original mutation, or where it is the parent or a child of it, are already excluded from the recurrent mutation tree space. However, as above, one can still recreate infinite sites trees if for example the duplication is placed below any mutation so that it affects no attached sample.

For a given infinite sites tree with attachments, we first prune any mutation nodes that have no attached samples as descendants. Say we remove k mutations, then any reattachment below the remaining $(n-k+1)$ tree nodes will result in them not having any attached samples as descendants, and the same likelihood.

To build this equivalence class, we add an artificial root to the k mutations and build any rooted tree. If the root has degree i then when we cut at the root we have i subtrees which we attach to any of the $(n-k+1)$ core tree nodes. The number of rooted trees with $(k+1)$ nodes, whose top node has degree i is

$$\binom{k-1}{i-1} k^{k-i} \tag{47}$$

giving a total number of possibilities

$$\sum_{i=1}^{k}(n-k+1)^i \binom{k-1}{i-1} k^{k-i} = (n-k+1)(n+1)^{k-1} \tag{48}$$

When we also add the duplication below the core tree and start with $(k+2)$ nodes we have an equivalence class of size

$$\sum_{i=1}^{k+1}(n-k+1)^i \binom{k}{i-1} (k+1)^{k-i+1} = (n-k+1)(n+2)^k \tag{49}$$

For simplicity, this result does not include any restriction on the placement of the duplication.

Now we distinguish two cases: the node to be duplicated has descendant attached samples and remains in the core tree or it is removed as one of the k. In the first case swapping the labels of the duplication and the original mutation leads to a distinct tree in the larger space. We take one copy and have the ratio

$$\frac{(n+2)^k}{(n+1)^{k-1}} < n\mathrm{e}\,, \qquad n \geq 2 \tag{50}$$

for the maximum number of trees with duplications which recreate an infinite sites tree. To prove the bound, consider

$$\log\left[\frac{(n+2)^k}{(n+1)^k}\right] = k\log\left[1+\frac{1}{n+1}\right] < \frac{k}{n+1} \leq \frac{n-1}{n+1} \tag{51}$$

because $k < n$ so that

$$\frac{(n+2)^k}{(n+1)^{k-1}} < (n+1)\mathrm{e}^{\frac{n-1}{n+1}} < \left(n-1+\frac{2}{n}\right)\mathrm{e} \leq n\mathrm{e}\,, \qquad n \geq 2 \tag{52}$$

In obtaining the bound above, trees where the duplication is attached to the original mutation or its parent were not excluded to simplify the combinatorial considerations. Such restrictions just reduce the number of equivalent trees with a recurrent mutation and the numerator in Supp. Equation (50), and hence do not affect the validity of the bound.

In the second case, with both the original mutation and the duplication below the nodes in the remaining core tree, swapping their labels recreates a tree already counted in the equivalence class. Each tree is then counted twice, apart from when both the original mutation and the duplication are leaves below the same node. Such trees can be built by removing both the original mutations and its duplication during construction of the tree and then later placing them on any of the other n mutations (including the root) giving

$$n\sum_{i=1}^{k-1}(n-k+1)^i \binom{k-2}{i-1} (k-1)^{k-i-1} = (n-k+1)n^{k-1} \tag{53}$$

possibilities with $k \geq 1$. However trees where the original mutation and its duplication share a parent can be excluded from consideration (along with trees with a direct parent-child relationship between the original and the duplication) so we simply subtract these possibilities before dividing by two, leaving a total of

$$\frac{1}{2}(n-k+1)\left[(n+2)^k - n^{k-1}\right] \tag{54}$$

Taking the ratio with Supp. Equation (48) easily gives a lower bound than before

$$\frac{1}{2}\left[\frac{(n+2)^k - n^{k-1}}{(n+1)^{k-1}}\right] < \frac{1}{2}n\mathrm{e} < n\mathrm{e}\,, \qquad n \geq 2 \tag{55}$$

When $n = 1$, all trees with a duplicate are excluded, and the bound $n\mathrm{e}$ extends to all n. This bound allows us to express the sum in the numerator of Supp. Equation (30) as

$$\sum_{T\in\mathcal{M}_i}{}' s(T) \geq \sum_{T\in\mathcal{M}_i} s(T) - n\mathrm{e}\sum_{T\in\mathcal{M}_\mathrm{I}} s(T) \tag{56}$$

and obtain the other component of the bound in Supp. Equation (31).

Bibliography

Forero, R. M., Hernández, M., and Hernández-Rivas, J. M., 2013. Genetics of acute lymphoblastic leukemia. In Guenova, M. and Balatzenko, G., editors, *Leukemia*, pages 1–37. InTech.

Forero-Castro, M., Robledo, C., Benito, R., Abáigar, M., Martín, A. Á., Arefi, M., Fuster, J. L., de las Heras, N., Rodríguez, J. N., Quintero, J., *et al.*, 2016. Genome-wide DNA copy number analysis of acute lymphoblastic leukemia identifies new genetic markers associated with clinical outcome. *PloS ONE*, **11**:e0148972.

Gawad, C., Koh, W., and Quake, S. R., 2014. Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, **111**:17947–17952.

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., *et al.*, 2012. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell*, **148**:873–885.

Jahn, K., Kuipers, J., and Beerenwinkel, N., 2016. Tree inference for single-cell data. *Genome Biology*, **17**:86.

Kim, K. I. and Simon, R., 2014. Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, **15**:27.

Lynch, M., 2010. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences*, **107**:961–968.

McPherson, A., Roth, A., Laks, E., Masud, T., Bashashati, A., Zhang, A. W., Ha, G., Biele, J., Yap, D., Wan, A., *et al.*, 2016. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nature Genetics*, **48**:758–767.

Tomasetti, C. and Vogelstein, B., 2015. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science*, **347**:78–81.

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., and Nowak, M. A., 2015. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, **525**:261–264.

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., *et al.*, 2014. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, **512**:155–160.

Wilf, H. S., 2013. generatingfunctionology.

Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., *et al.*, 2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**:886–895.