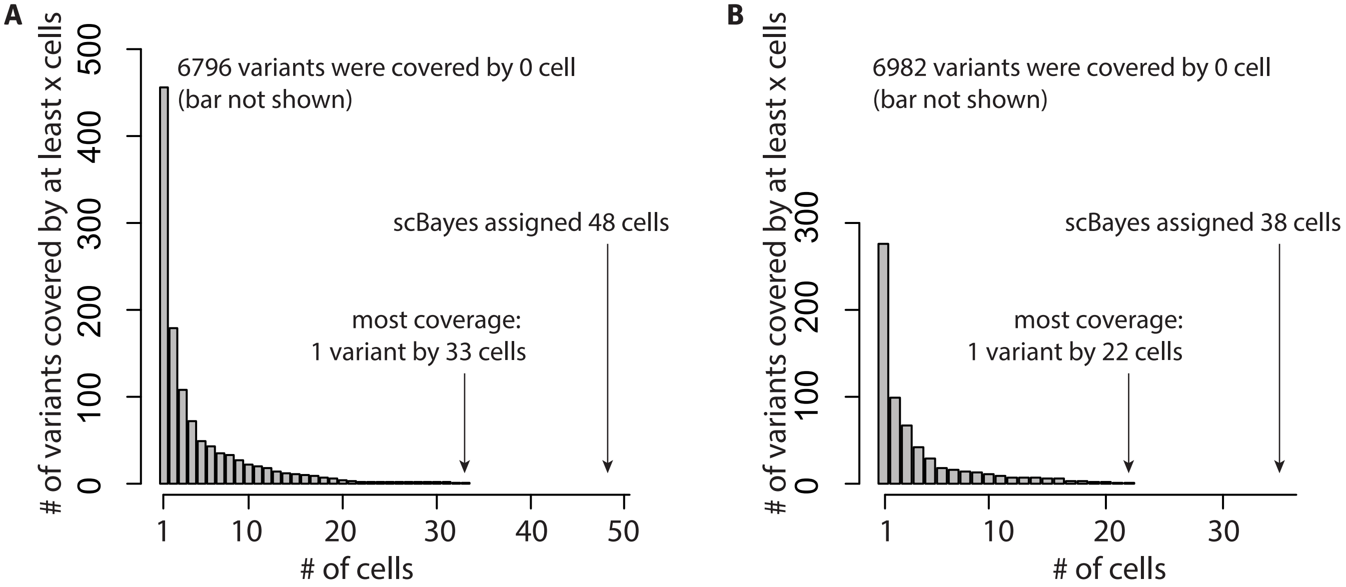
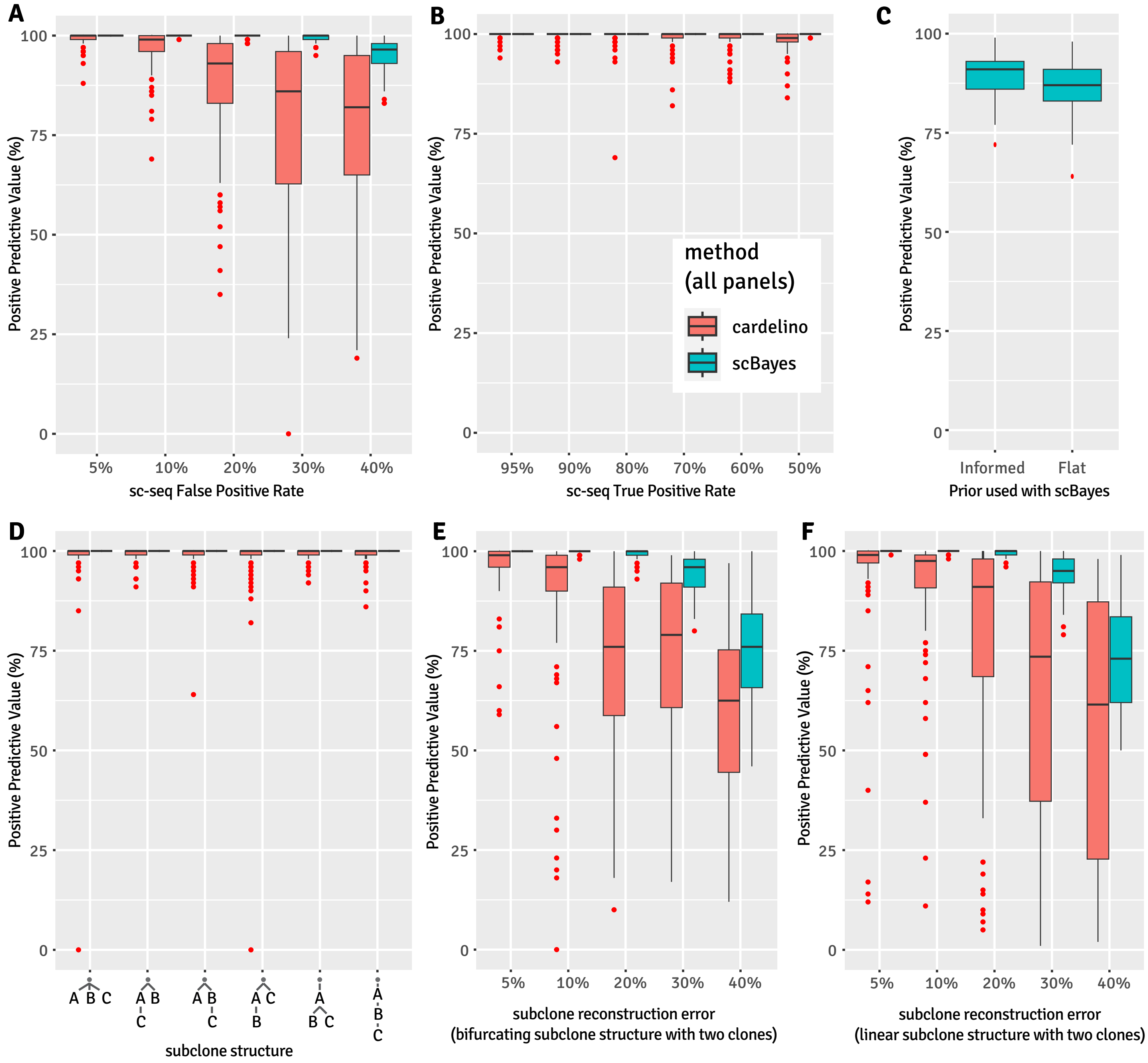
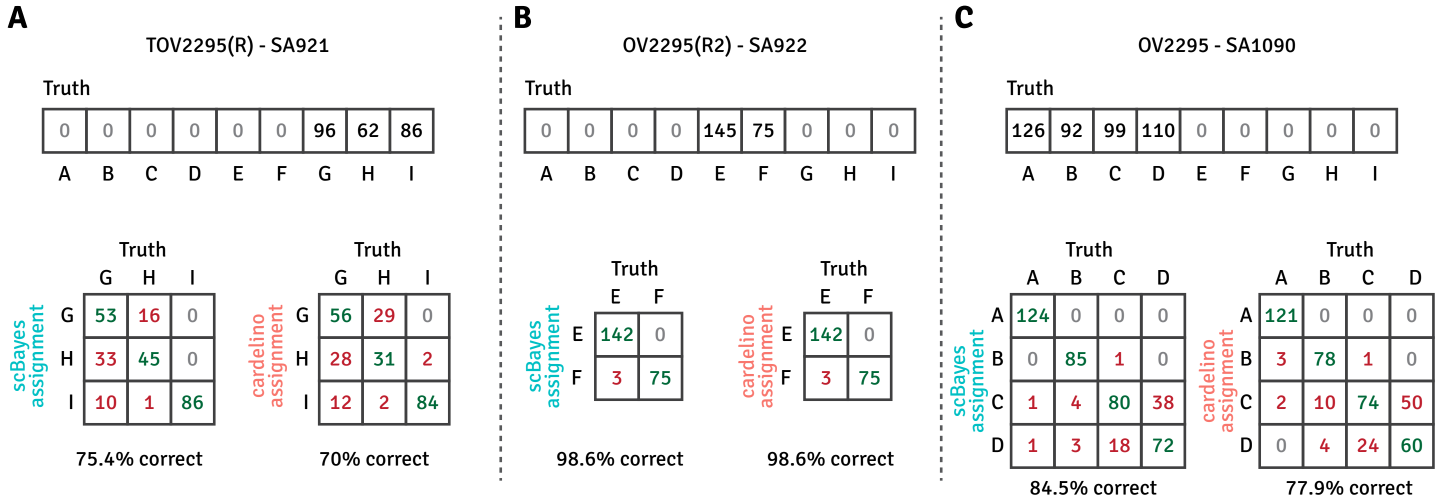
Supplemental Figures



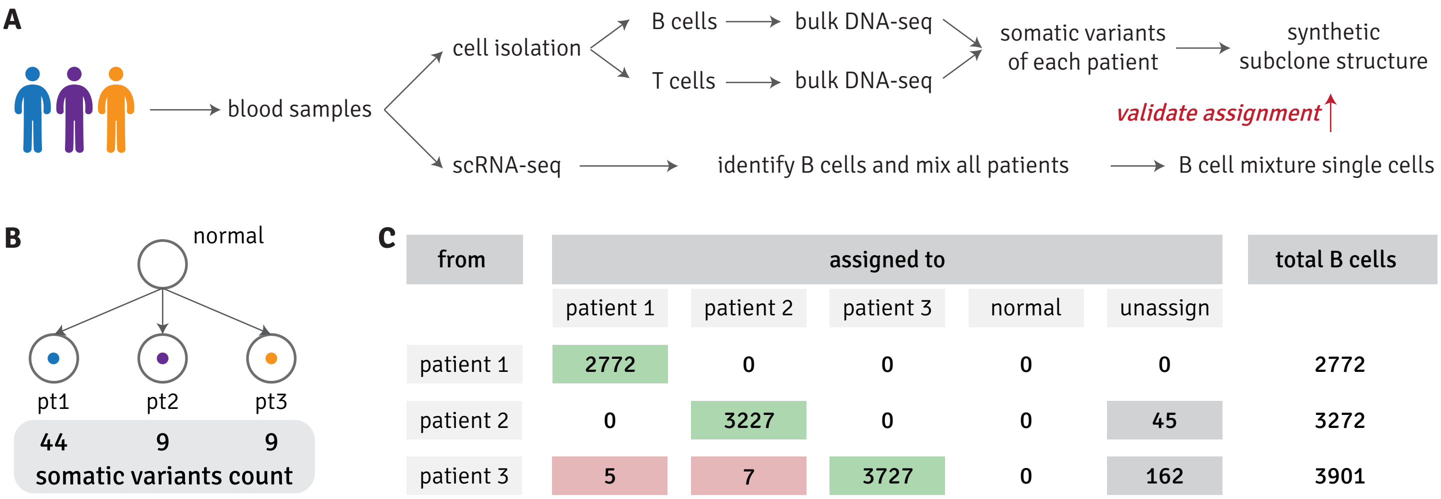
Supplemental Figure 1, The power of individual variants at delineating cell genotypes, calculated on the breast cancer dataset . Both figures show the number of variants (y-axis) having sequencing coverage (>= 1 read) in *at least* x (x-axis) number of cells. **A)** breast cancer pre-treatment sample (***Fig. 2, Supp. Fig. 7***). This figure indicates that the majority of the somatic variants (6796) did not have sequencing coverage in any cells. These variants will have no power at delineating cells between wild type and mutant. The likelihood of any variant to be covered by more cells is monotonically and drastically decreasing (e.g. only 22 variants are covered by at least 10 cells). The most covered variant is covered by 33 cells, a theoretical upper limit of single variant based cell assignment approach. scBayes was able to assign more (42) cells by using groups of variants according to subclones. **B)** breast cancer post-treatment sample (***Fig. 2, Supp. Fig. 7***)



Supplemental Figure 2, scBayes performance evaluation using simulation and comparison to cardelino. A) The effect of single cell sequencing false positive rate (error rate) between 5% and 40%. **B)** The effect of single cell sequencing true positive rate (pick-up rate) between 95% and 50%. **C)** The effect of using informed vs flat priors with scBayes (cardelino does not support custom priors). Data is simulated using 20% false positive rate and 30% subclone error rate to render the difference more obvious. **D)** The effect of different subclone structures. **E)** The effect of subclone reconstruction errors from 5% to 40% in a bifurcating subclone structure with two clones. We define subclone reconstruction error as the percentage of somatic mutations being erroneously attributed to the wrong subclone. **F)** The effect of subclone reconstruction errors from 5% to 40% in a linear subclone structure with two clones.

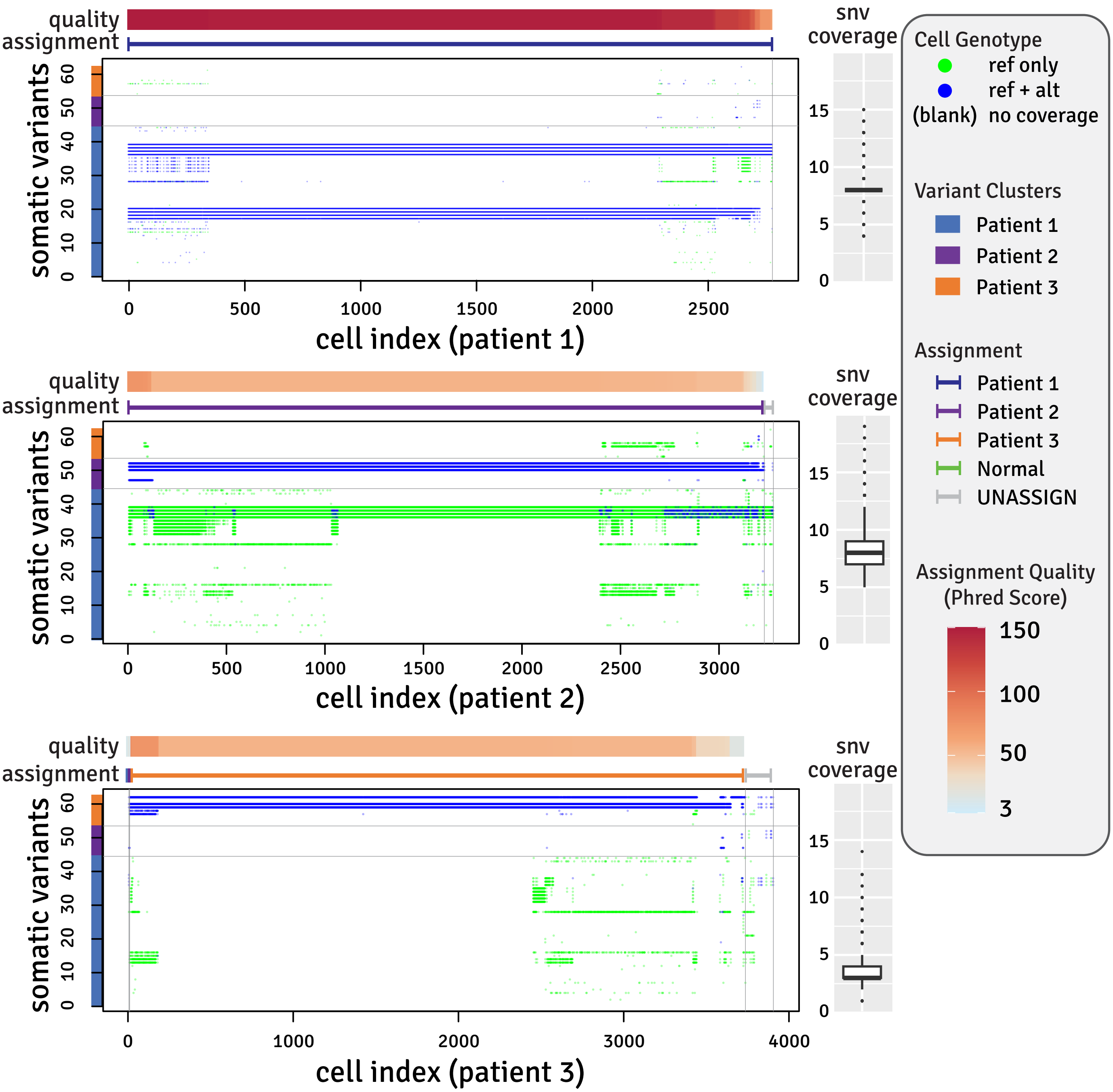


Supplemental Figure 3, Cell assignment performance using a published, single cell DNA sequencing derived pseudo-bulk dataset. A) assignment results on sample SA921. **B)** assignment results on sample SA922. **C)** assignment results on sample SA1090.

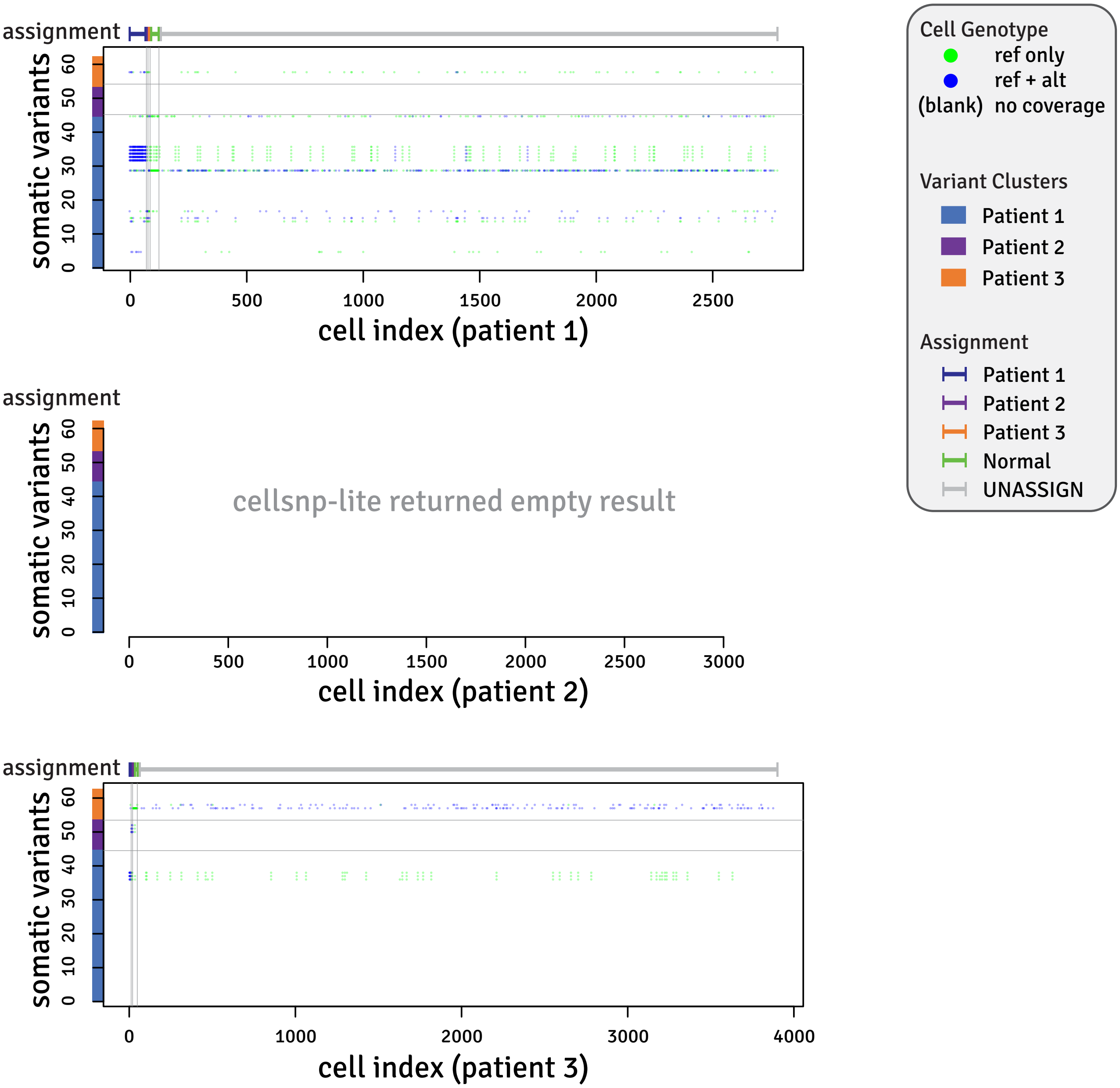


from	assigned to					total B cells
	patient 1	patient 2	patient 3	normal	unassign	
patient 1	2772	0	0	0	0	2772
patient 2	0	3227	0	0	45	3272
patient 3	5	7	3727	0	162	3901

Supplemental Figure 4, Validation of scBayes cell assignment algorithm using a synthetic dataset. A) Data generation. Three chronic lymphocytic leukemia patient samples are separately bulk DNA sequenced and single cell RNA sequenced. Bulk DNA sequencing is used to identify somatic mutations of each patient, and to construct a synthetic subclone structure. B cells from the single cell sequencing data are mixed together, and used to validate scBayes assignment. **B)** synthetic subclone structure that consists of one normal subclone having no mutations, and three cancer subclones each of which contains somatic mutations from one patient. **c**) scBayes assignment result.

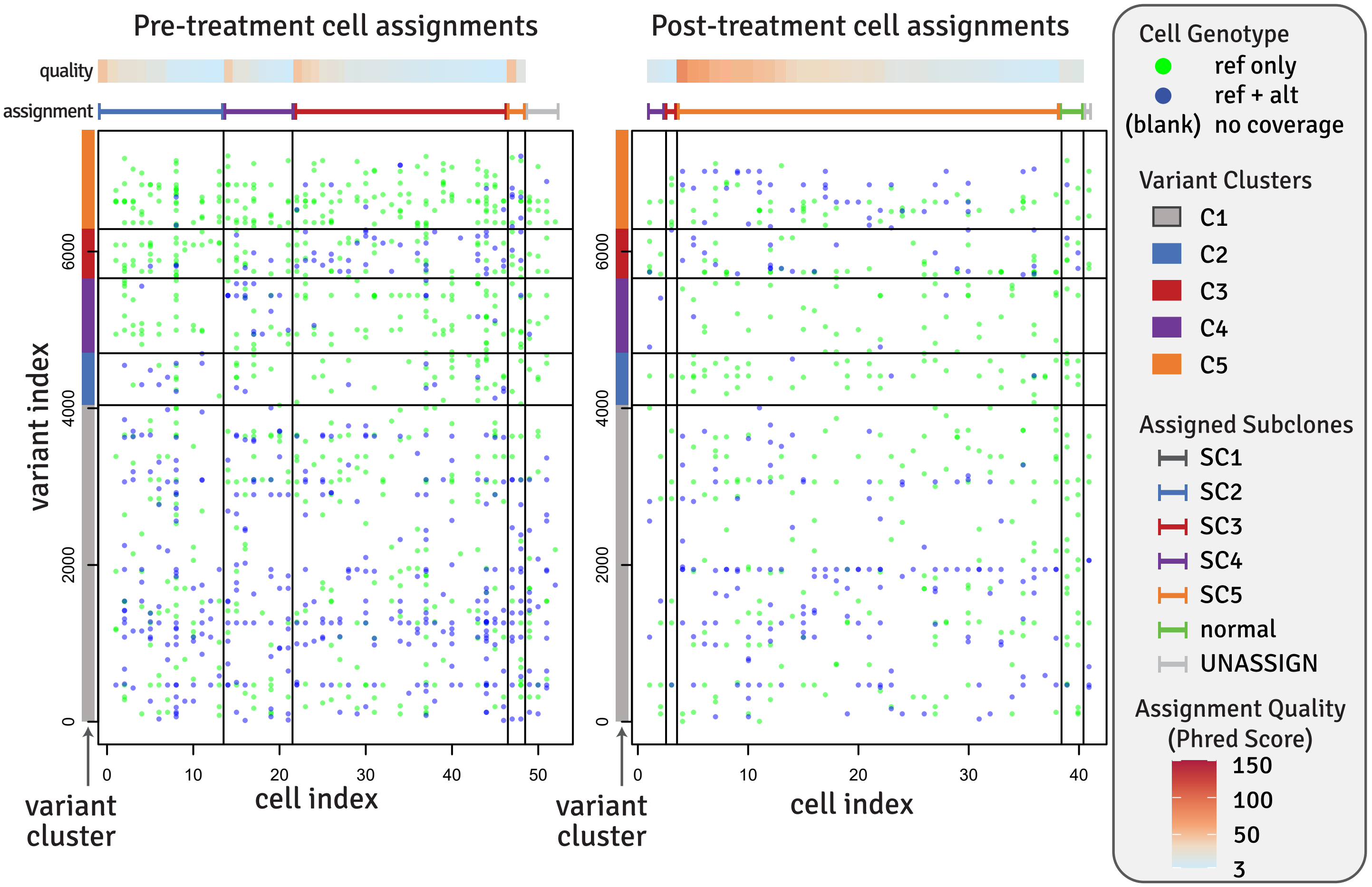


Supplemental Figure 5, Single cell assignment details with the synthetic subclone structure and B cells from three CLL patients. Each panel is a scatter plot in which the x-axis corresponds to individual cells, and the y-axis individual variants. A particular cell-variant coordinate is filled in when sequencing coverage is detected in that cell at the location of that variant. If all reads overlapping this location show the reference allele, a green dot is drawn; if at least one read shows the variant allele, a blue dot is drawn. The patient origin of each somatic variant, and the assignment results for each cell are shown along the x axis, and the top of the panels respectively. Cell assignment qualities (Phred scale of the maximum posterior probability) are indicated at the top of each panel.

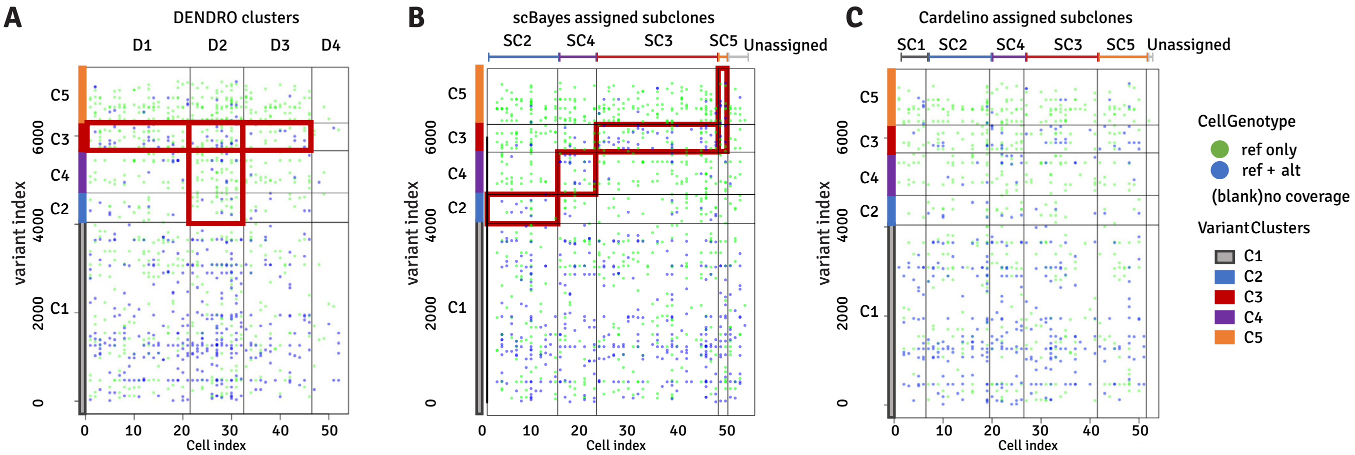


from	assigned to					total B cells
	patient 1	patient 2	patient 3	normal	unassign	
patient 1	73	7	9	38	2645	2772
patient 2	N/A	N/A	N/A	N/A	N/A	N/A
patient 3	11	9	0	28	3853	3901

Supplemental Figure 6, Single cell assignment details with the synthetic subclone structure and B cells from three CLL patients, using Cardelino and default cellsnp-lite filtering parameters. Cell assignment accuracies are 2.63%, N/A, and 0% respectively; and 95.42%, N/A, and 98.77% cells were unable to be assigned.



Supplemental Figure 7, Single cell assignment details with the longitudinal breast cancer dataset. Panels on the left and the right correspond to the pre-treatment cells and post-treatment cells respectively. Each panel is a scatter plot in which the x-axis corresponds to individual cells, and the y-axis individual variants. A particular cell-variant coordinate is filled in when sequencing coverage is detected in that cell at the location of that variant. If all reads overlapping this location show the reference allele, a green dot is drawn; if at least one read shows the variant allele, a blue dot is drawn. Cell assignment results and assignment quality (Phred scale of the maximum posterior probability) are indicated at the top of each panel.



A

Somatic mutations in each cell cluster

	D1	D2	D3	D4
C5	9	5	1	0
C3	12	12	6	1
C4	9	9	3	1
C2	5	6	4	0
C1	80	87	47	7

Somatic mutations in collision with assignment

Assignment	Collision	Expected
D1 -> SC5	14	101
D2 -> SC3	20	99
D3 -> SC2	10	51
D4 -> SC4	1	8
SUM	**45**	**259**
ACR	**0.15**	

B

Somatic mutations in each cell cluster

	SC1	SC2	SC4	SC3	SC5	Unassign
C5	0	2	0	6	6	1
C3	0	2	4	19	3	0
C4	0	1	13	6	1	1
C2	0	7	2	6	0	0
C1	0	63	46	79	19	21

Somatic mutations in collision with assignment

Assignment	Collision	Expected
SC1-> SC1	0	0
SC2-> SC2	5	68
SC3-> SC3	18	98
SC4-> SC4	6	59
SC5-> SC5	1	28
SUM	**30**	**253**
ACR	**0.106**	

C

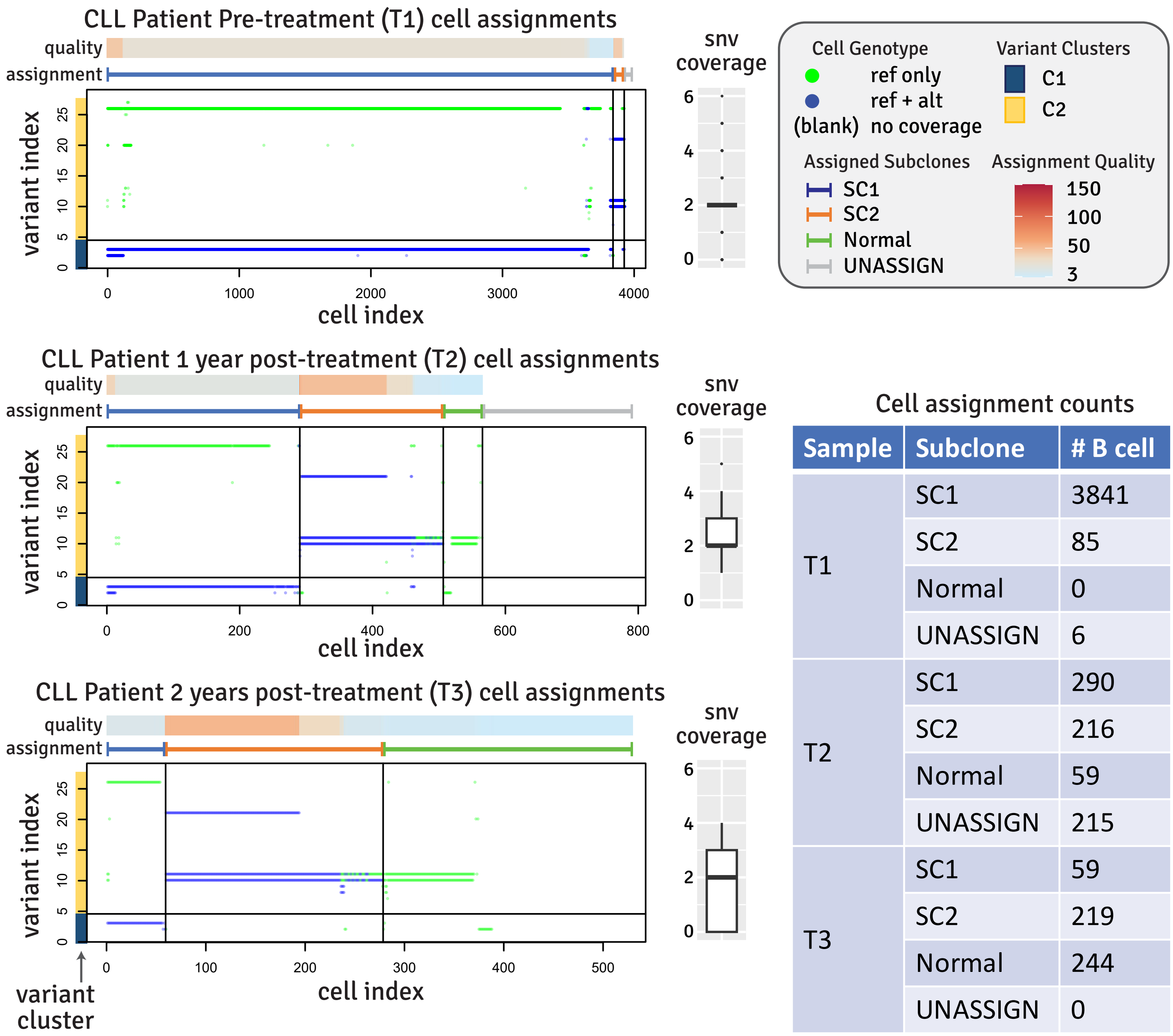
Somatic mutations in each cell cluster

	SC1	SC2	SC4	SC3	SC5	Unassign
C5	5	2	1	3	3	0
C3	1	3	7	6	5	0
C4	0	4	8	3	2	0
C2	0	4	3	2	1	0
C1	25	49	34	47	23	0

Somatic mutations in collision with assignment

Assignment	Collision	Expected
SC1-> SC1	6	25
SC2-> SC2	9	53
SC3-> SC3	8	53
SC4-> SC4	11	42
SC5-> SC5	3	31
SUM	**37**	**204**
ACR	**0.15**	

Supplemental Figure 8. Evaluating the subclone assignment results across DENDRO based cell clusters, scBayes, and Cardelino, with the breast cancer pre-treatment sample. A) DENDRO clustered the cells genetically into four clusters (D1-D4). Top panel shows Sequencing evidence for the presence of somatic mutations (blue dots) or reference only sequencing coverage (green dots) for groups of somatic mutations (C1-C5) that define genetic subclones (SC1-SC5, see ***Fig. 2***). C1 defines the founder clone SC1, therefore C1 mutations are expected to be present in all cancer cells. However C2, C3, and C4 each define SC2, SC3, and SC4 genetic subclones respectively, and are expected to be present exclusively in one cell cluster each. The horizontal red box highlights the fact that mutation cluster C3 were found in cells that DENDRO clustered into different groups (D1-D3); and the vertical red box highlights the fact that somatic mutations specific to C2, C3, C4, respectively were found within the same DENDRO cluster D2. Both of these observations are indications of discrepancy between DENDRO cell clusters and genetic subclones. Middle panel shows the number of somatic mutations of each mutation cluster detected in each cell cluster. This table is used to calculate how well a particular DENDRO cell cluster maps to a genetic subclone (see **Methods**). Since DENDRO does not provide an assignment out-of-the-box, we enumerated all possible DENDRO cell clusters to genetic subclones assignment schemes, and chose the best result to compare to scBayes. Bottom panel shows the best assignment scheme, as well as the amount of somatic mutations in collision with the assignment (lower is better) and relative ratio (**ACR**, see **Methods**). **B)** scBayes cell assignment results. Red boxes in the top panel highlight that subclone defining mutations C2, C3, C4, and C5 are largely exclusive to the scBayes-assigned scRNA-seq cell clusters (C3 is found in both SC3 and SC5 because SC5 is a subclone derived from SC3). Bottom panel shows that the assignment reported by scBayes has a lower collision rate than the best assignment scheme obtainable from DENDRO cell clusters. **C)** Cardelino cell assignment results. We were not able to observe any clear visual patterns that would either suggest correct or incorrect assignment results. Quantitative analysis with the calculation of ACR value revealed that the overall assignment quality is similar to DENDRO's best assignment scheme, and lower than scBayes.



CLL Patient Pre-treatment (T1) cell assignments

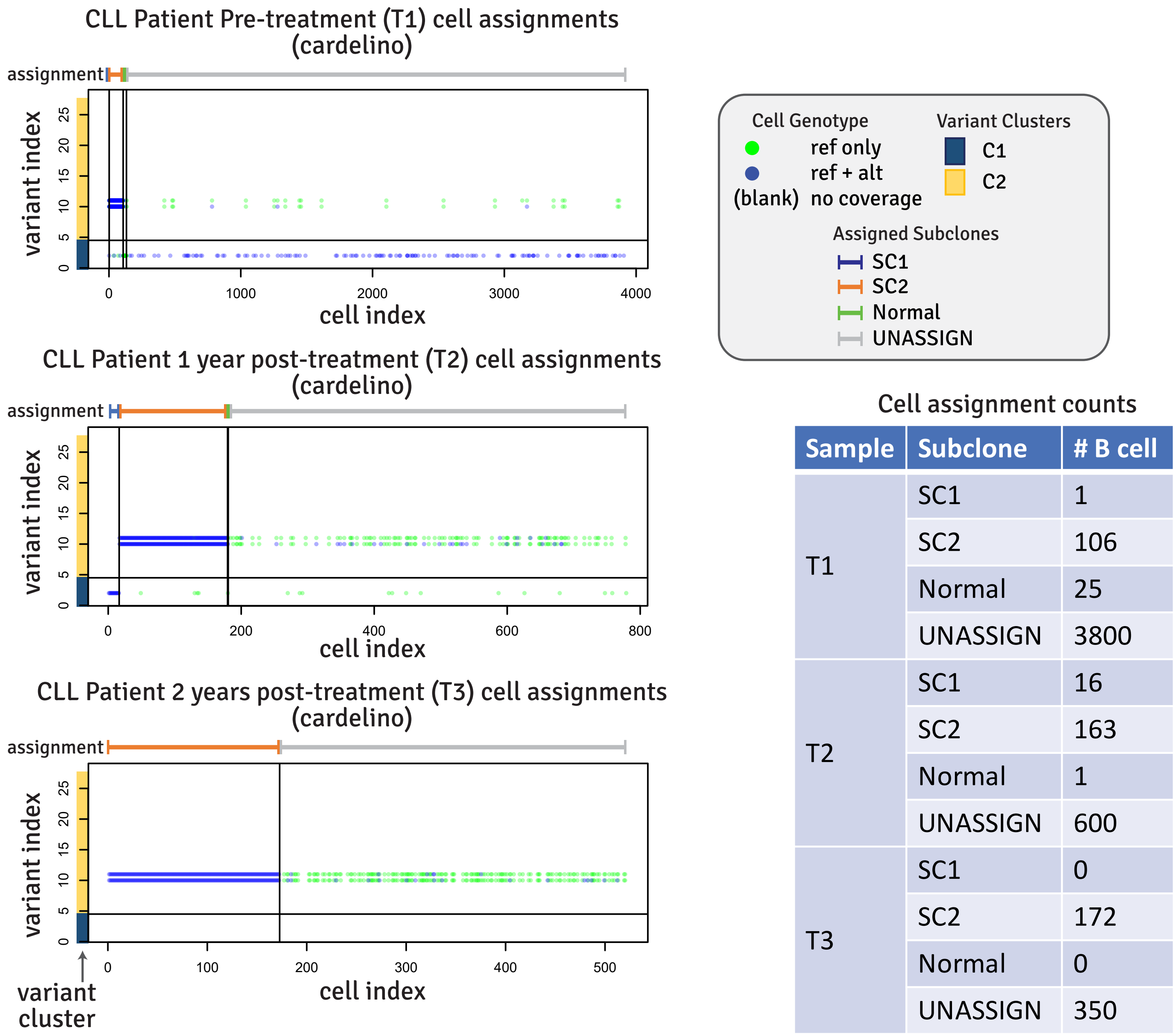
CLL Patient 1 year post-treatment (T2) cell assignments

CLL Patient 2 years post-treatment (T3) cell assignments

Cell assignment counts

Sample	Subclone	# B cell
T1	SC1	3841
	SC2	85
	Normal	0
	UNASSIGN	6
T2	SC1	290
	SC2	216
	Normal	59
	UNASSIGN	215
T3	SC1	59
	SC2	219
	Normal	244
	UNASSIGN	0

Supplemental Figure 9, Single cell assignment details with the longitudinal CLL dataset. The three panels from top to bottom correspond to pre-treatment (T1), 1 year after initiation of treatment (T2), 2 years after initiation of treatment (T3). Each panel is a scatter plot in which the x-axis corresponds to individual cells, and the y-axis individual variants. A particular cell-variant coordinate is filled in when sequencing coverage is detected in that cell at the location of that variant. If all reads overlapping this location show the reference allele, a green dot is drawn; if at least one read shows the variant allele, a blue dot is drawn. Cell assignment results and assignment quality (Phred scale of the maximum posterior probability) are indicated at the top of each panel. The numbers of cells assigned to each subclone are summarized in the table.



CLL Patient Pre-treatment (T1) cell assignments (cardelino)

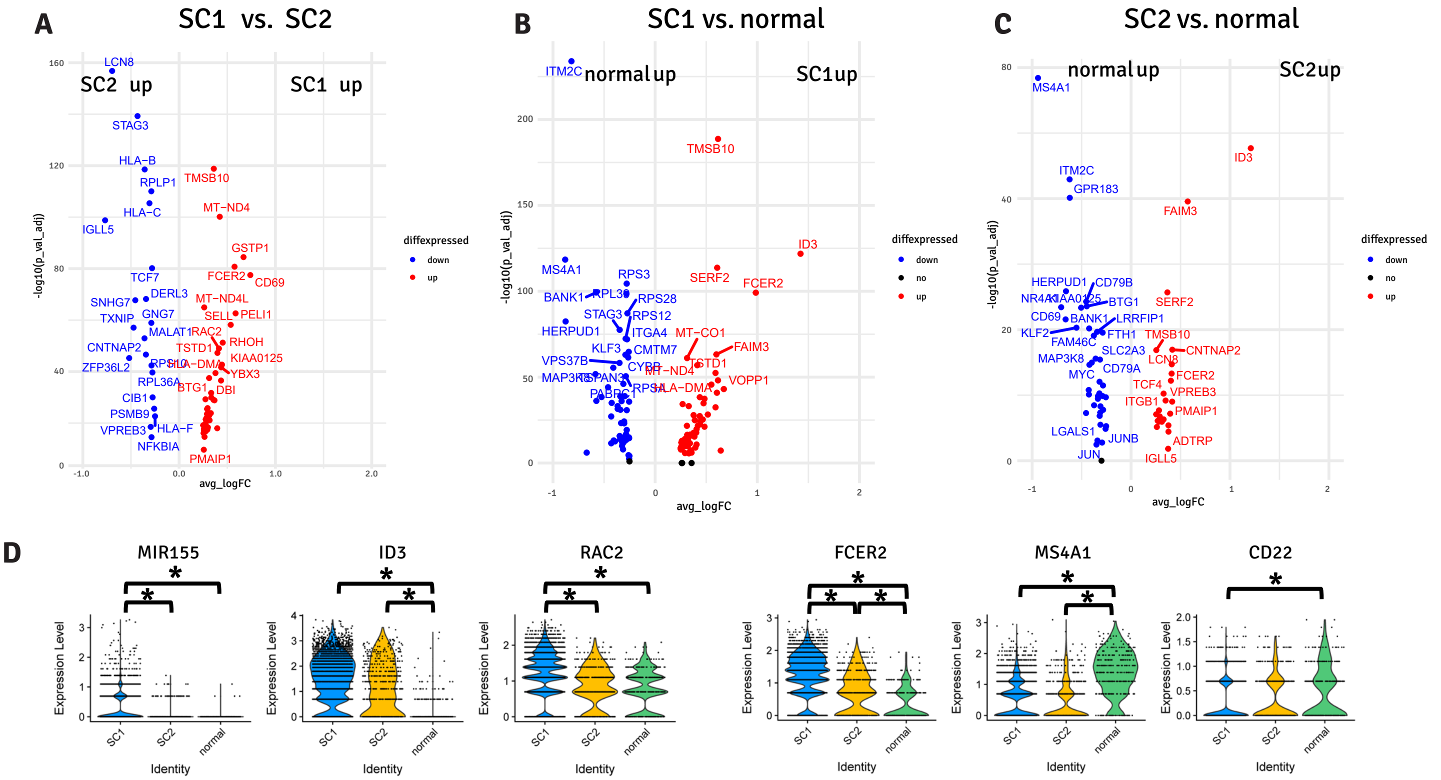
CLL Patient 1 year post-treatment (T2) cell assignments (cardelino)

CLL Patient 2 years post-treatment (T3) cell assignments (cardelino)

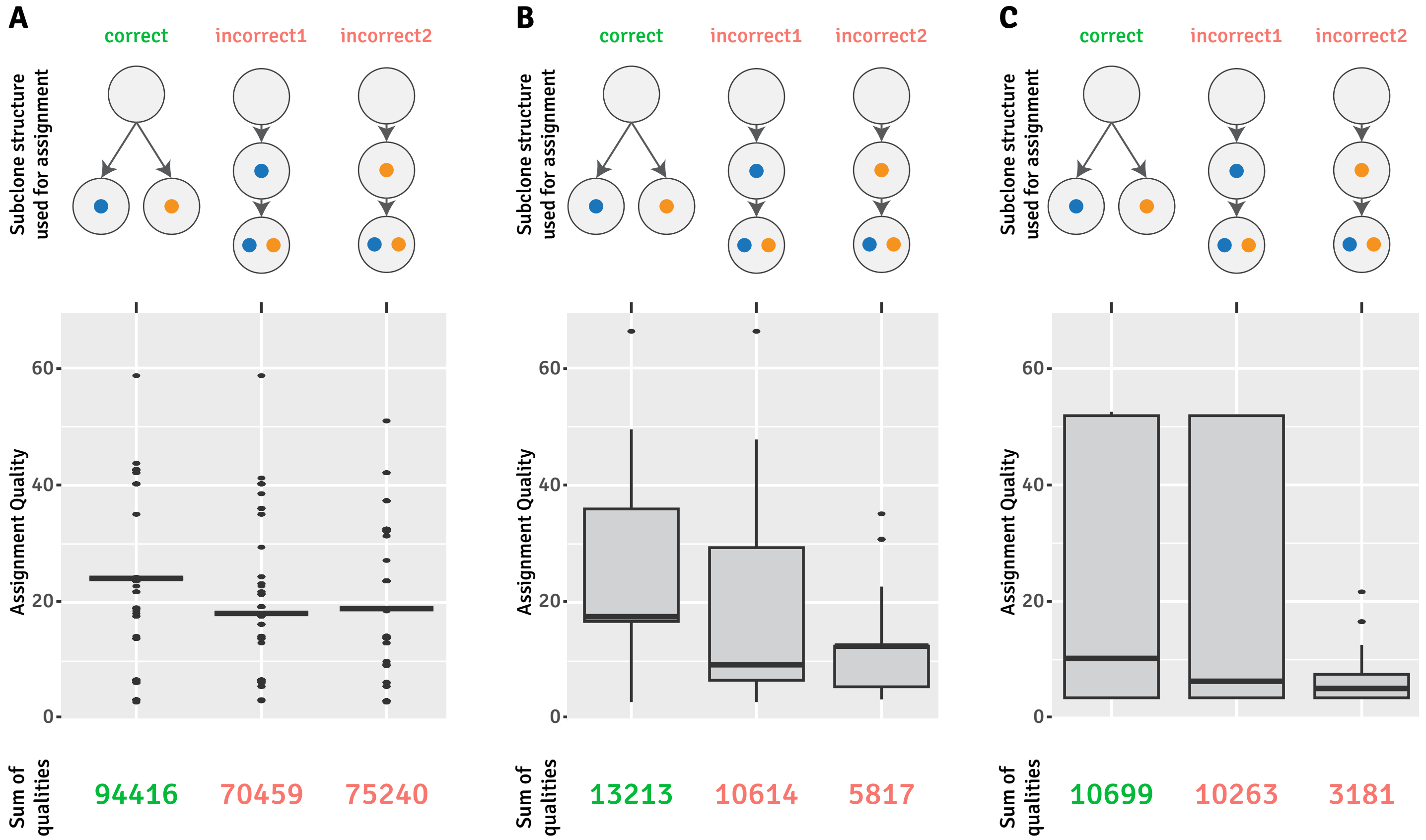
Cell assignment counts

Sample	Subclone	# B cell
T1	SC1	1
	SC2	106
	Normal	25
	UNASSIGN	3800
T2	SC1	16
	SC2	163
	Normal	1
	UNASSIGN	600
T3	SC1	0
	SC2	172
	Normal	0
	UNASSIGN	350

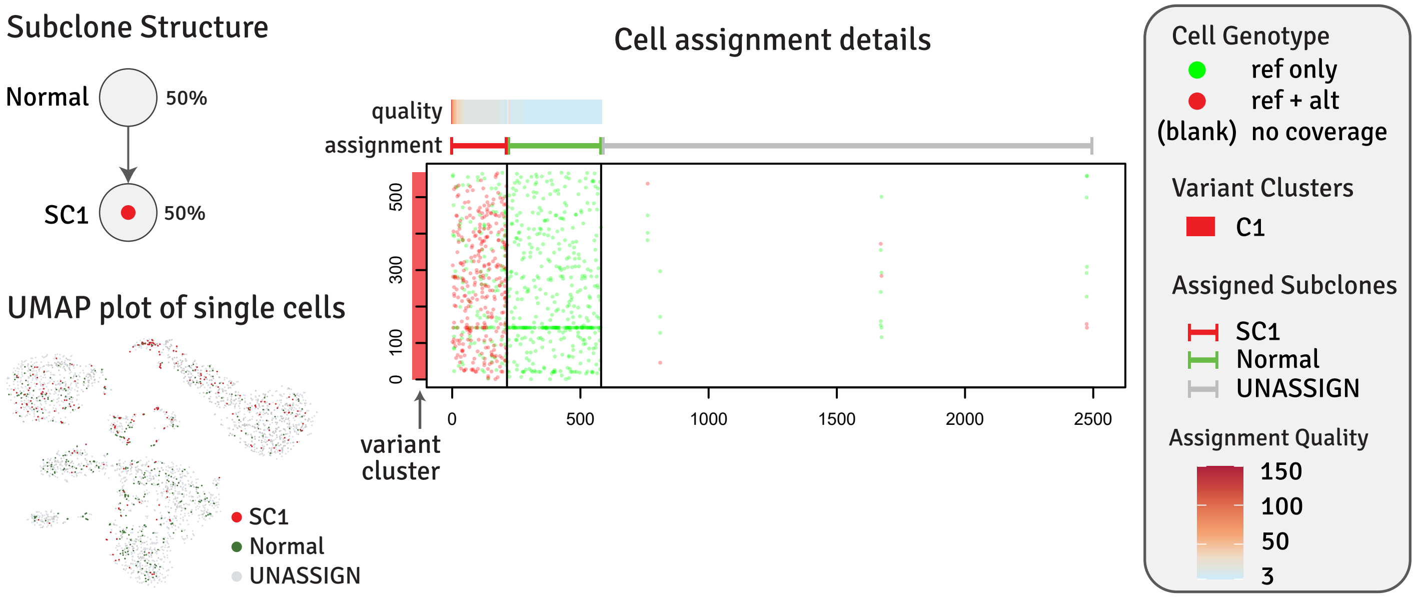
Supplemental Figure 10, Single cell assignment details with the longitudinal CLL dataset, using Cardelino. Comparing these results to scBayes (***Supp. Fig. 7***), we notice that 1) the number of assigned cells are significantly lower, which can negatively impact downstream, subclonal expression analysis; 2) The proportion of cells assigned to each subclone is significantly different from the subclonal fractions derived from bulk DNA sequencing, which are performed on the same bio-samples.



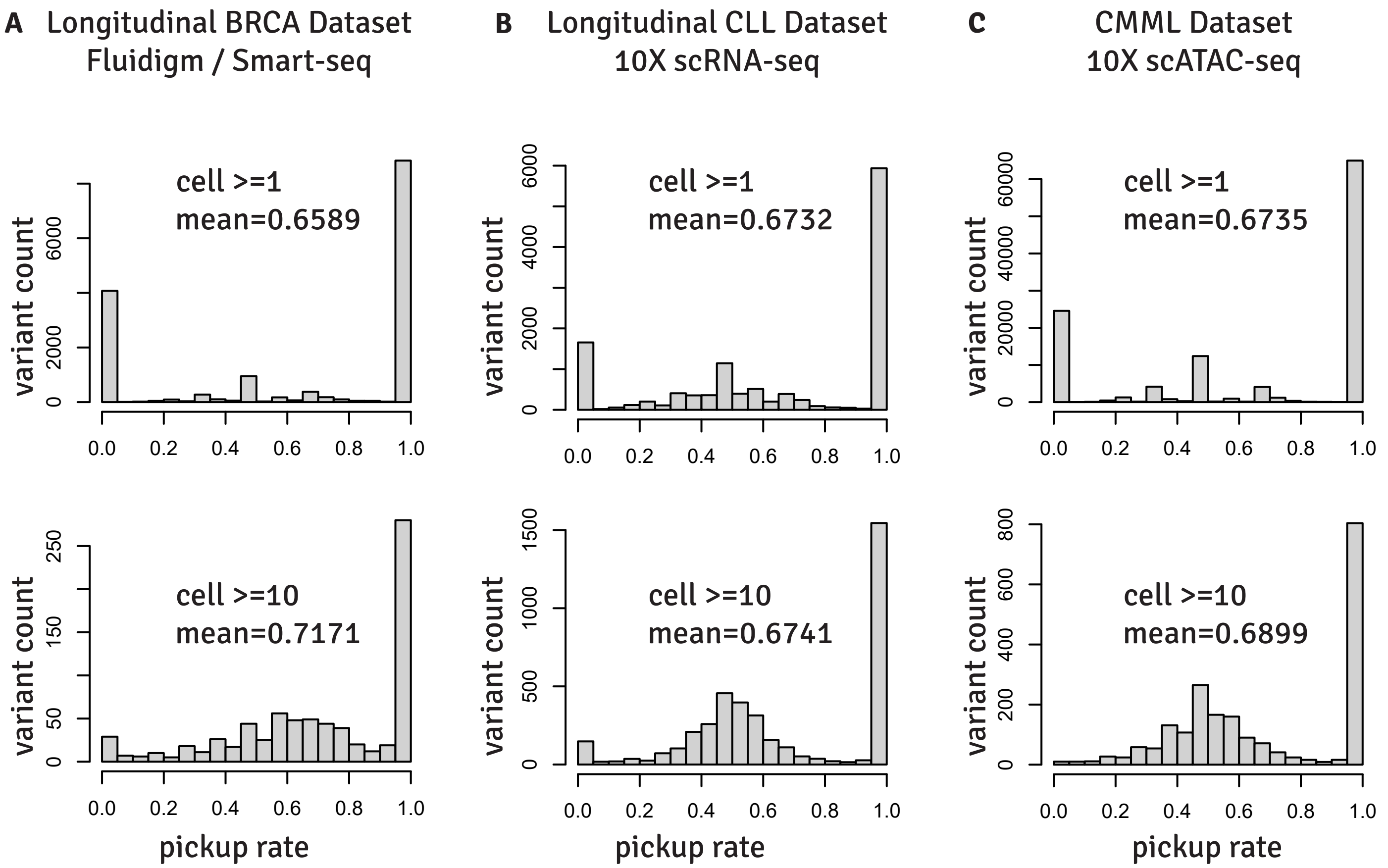
Supplemental Figure 11, Differential expression analysis of cells assigned to SC1, SC2, and normal clones from the CLL patient. A) We performed genome-wide differential expression analysis between cells assigned to SC1 and SC2, and found 44 up-regulated (red) and 22 down-regulated (blue) genesin SC1 relative to SC2. **B)** We found 96 up-regulated (red) and 56 down-regulated (blue) genes in SC1 relative to normal. **C)** We found 23 up-regulated (red) and 39 down-regulated (blue) genes in SC2 relative to normal. The x-axis shows average log2(fold change) of cells in different subclones; the y-axis shows the -log10(adjusted P value) in the volcano plots. Significantly expressed genes were defined as adjusted P <0.05. **D)** CLL malignancy relevant genes MIR155, ID3, RAC2, and FCER2 were overexpressed in SC1 relative to normal; and B cell markers CD22 and MS4A1 were underexpressed in SC1 relative to normal. The expression levels of these genes in SC2 cells were between SC1 and normal. * indicates P<0.005 and FDR<0.05.



Supplemental Figure 12, Using scBayes to evaluate alternative subclone structures. To assess if we can use the cell assignment quality scores from scBayes to select the correct subclone structure across alternative structures, we carried out an exercise in which we manually altered the correct subclone structure as presented in ***Figure 3*** to two incorrect versions, and assigned the single B cells to all three subclone structures separately. **A)** Distribution and sum of assignment qualities of B cells from sample T1. **B)** Distribution and sum of assignment qualities of B cells from sample T2. **C)** Distribution and sum of assignment qualities of B cells from sample T3.



Supplemental Figure 13, Proof-of-concept application of scBayes to a chronic myelomonocytic leukemia (CMML) patient blood sample analyzed with bulk DNA whole genome sequencing and single cell ATAC sequencing (10x Genomics protocol). Bulk whole genome DNA sequencing on sorted mononuclear cells using the skin sample as normal control from a CMML patient yielded a subclone structure that has one cancer clone (top left). scBayes was used to assign single cells from scATACseq data (bottom left) to identify cells of the cancerous SC1 population vs normal population. Cell assignment details (right) are shown with a scatterplot in which the x-axis corresponds to individual cells, and the y-axis individual variants. A particular cell-variant coordinate is filled in when sequencing coverage is detected in that cell at the location of that variant. If all reads overlap this location show the reference allele, a green dot is drawn; if at least one read shows the variant allele, a red dot is drawn.



Supplemental Figure 14, Estimation of single cell sequencing variant pick up rate. A) results from the longitudinal breast cancer dataset generated with the Fluidigm / Smart-seq platform. If we include variants at whose locations at least one cell had sequencing coverage, we get a variant pickup rate of 0.6589 on average. This is currently the default parameter in scBayes, but customizable. For example, if we increase the variant filtering criteria to be at least ten cells having sequencing coverage, the pick-up rate increases to 0.7171. **B)** results from the longitudinal CLL dataset generated with the 10x Genomics single cell RNA sequencing platform. **C)** results from the longitudinal CMML dataset generated with the 10x Genomics single cell ATAC sequencing platform.

Supplemental Table 1, A summary of studies investigating tumor heterogeneity, their cancer type, cohort size, and average number of subclones per tumor as well as their average number of somatic mutations per subclone.

Study	Cancer	Patients	Sequencing technology	Average number of subclones per tumor	Average number of somatic mutations per subclone
Gundem et al. Nature, 2015	Metastatic Prostate Cancer	10	55X WGS	2-8(4.6)	569
Hong et al, Nature Communication, 2014	Prostate Cancer	4	WGS and Custom capture sequencing	1-5(2.5)	3230
Hoadley et al. PloS Med. 2016	Breast Cancer	2	WGS (33X-70X)	1-5 (4)	42
Savas et al. PloS Med, 2016	Breast Cancer	4	WES	1-6 (3.4)	35