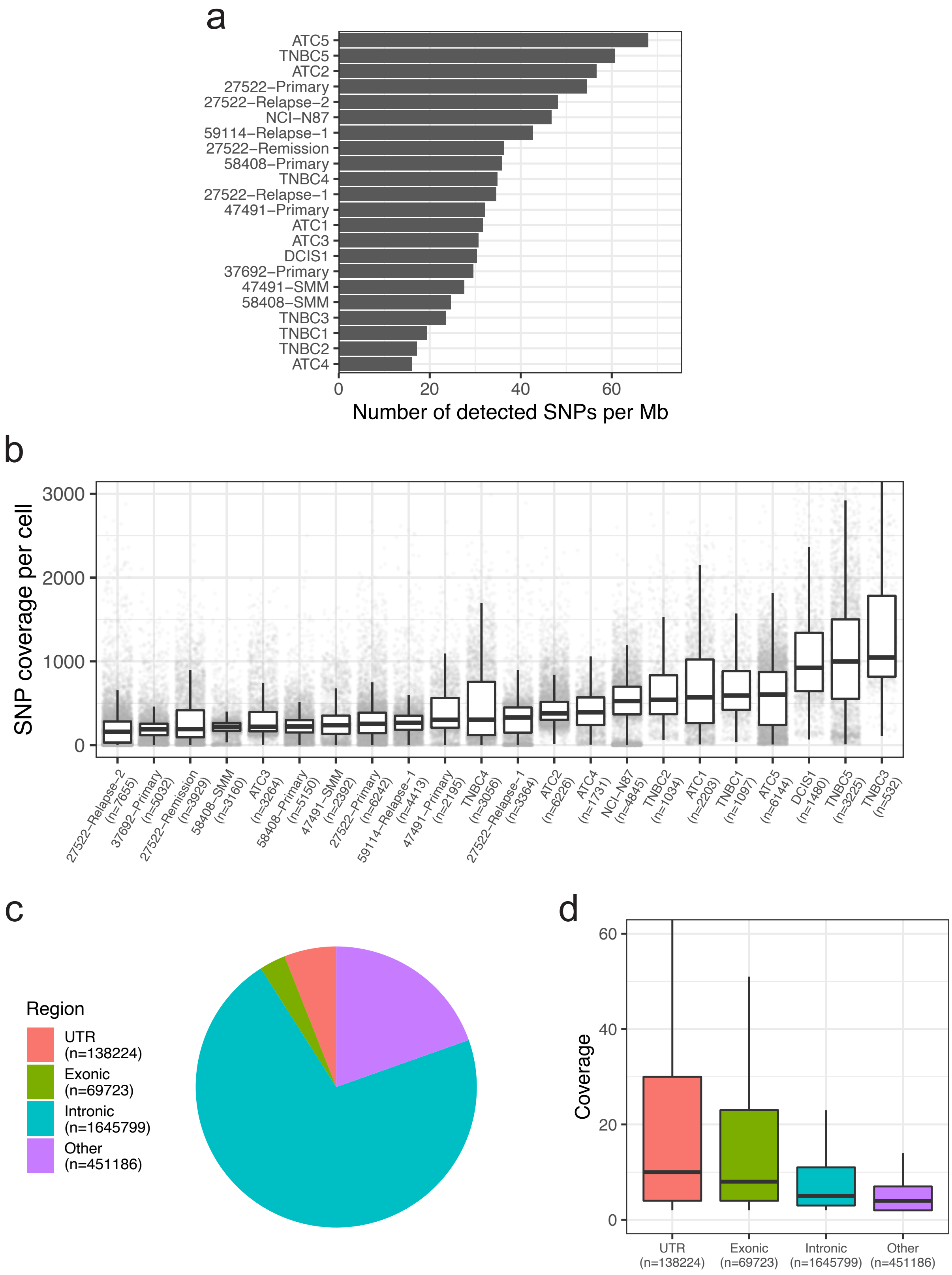
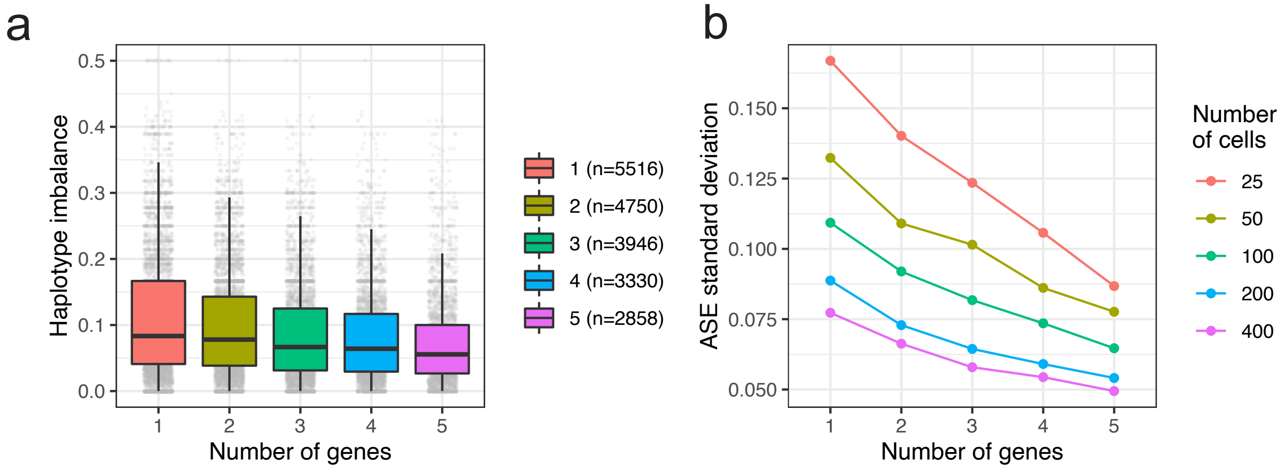
Haplotype-aware analysis of somatic copy number variations from single-cell transcriptomes

Supplementary Information

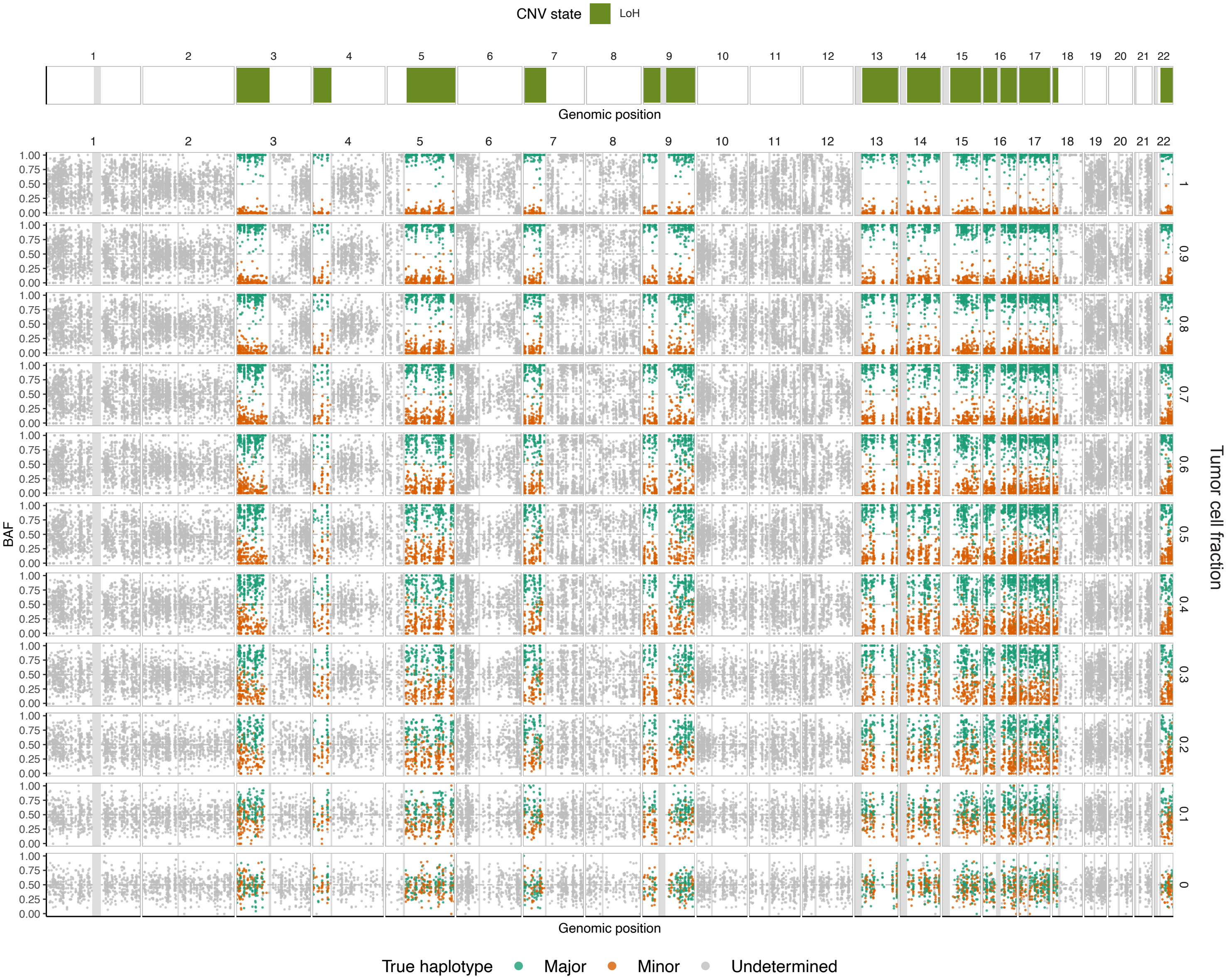
Table of Contents



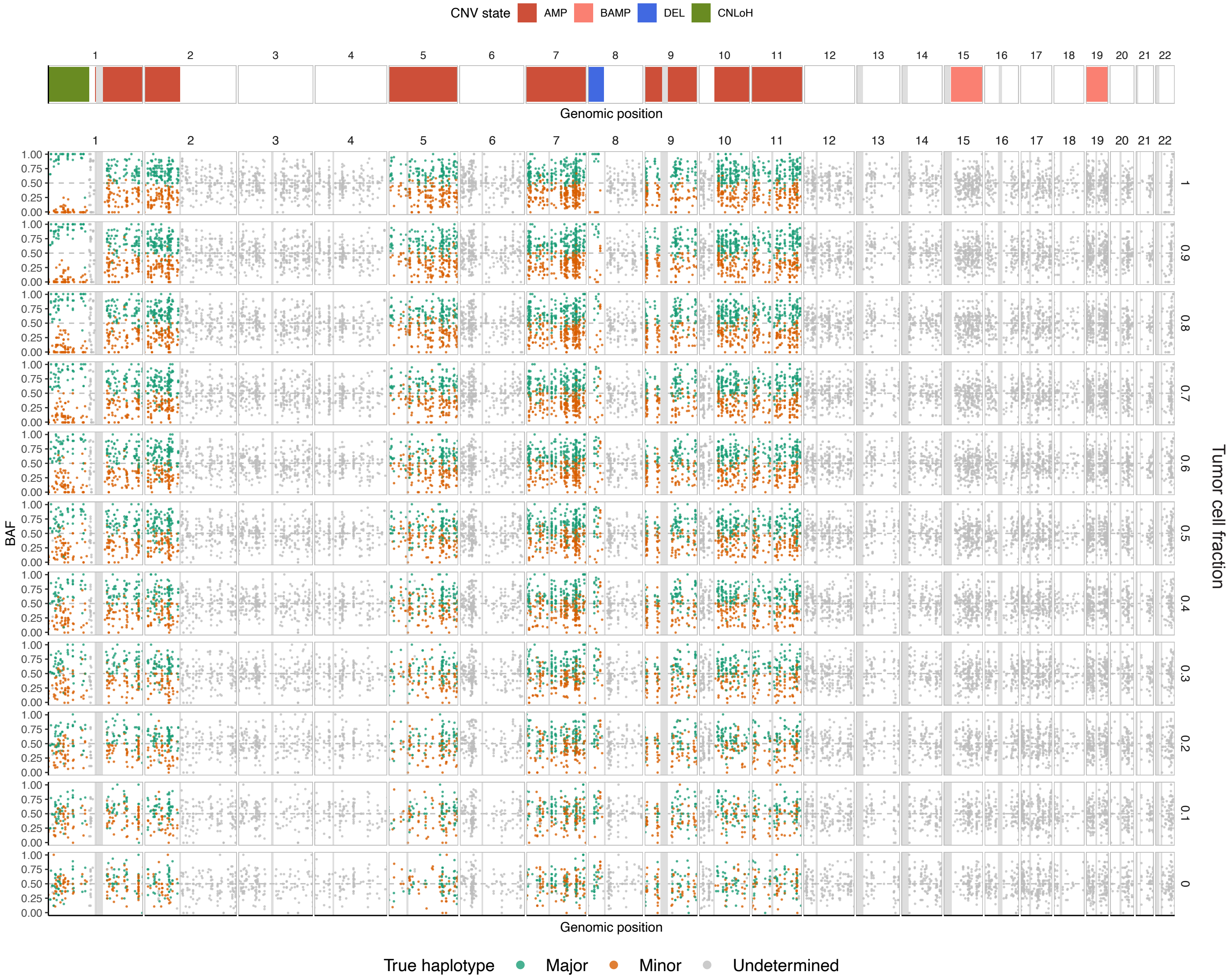
Supplementary Figure 1: Coverage metrics of germline heterozygous SNPs. a, Density of genotyped heterozygous SNPs for each sample. **b**, Total SNP coverage per cell for each sample. Each dot represents a distinct cell. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. **c**, Distribution of detected SNPs among genomic features. **d**, Coverage of SNPs by genomic features. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.



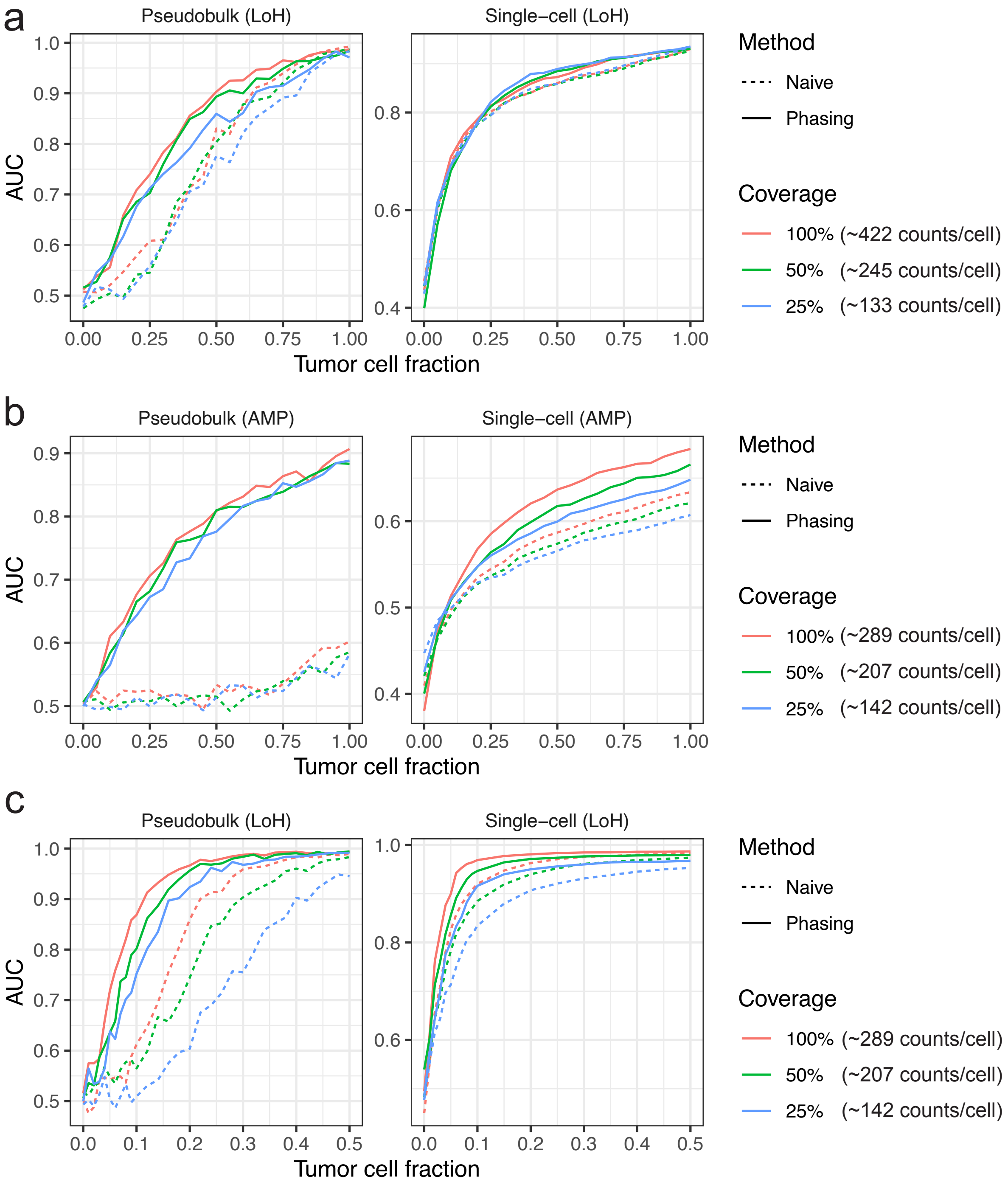
Supplementary Figure 2: Allele-specific expression in non-aberrant cells from TNBC4. a, Haplotype imbalance averaged across an increasing number of genes. Allele counts were created by aggregating 400 randomly sampled cells. Each dot represents one contiguous gene set. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range. **b**, Estimated standard deviation in allele-specific expression (modeled by a Beta distribution) when allele counts are aggregated across increasing number of genes. For all panels, ten replicates were performed for each condition with different random seeds.



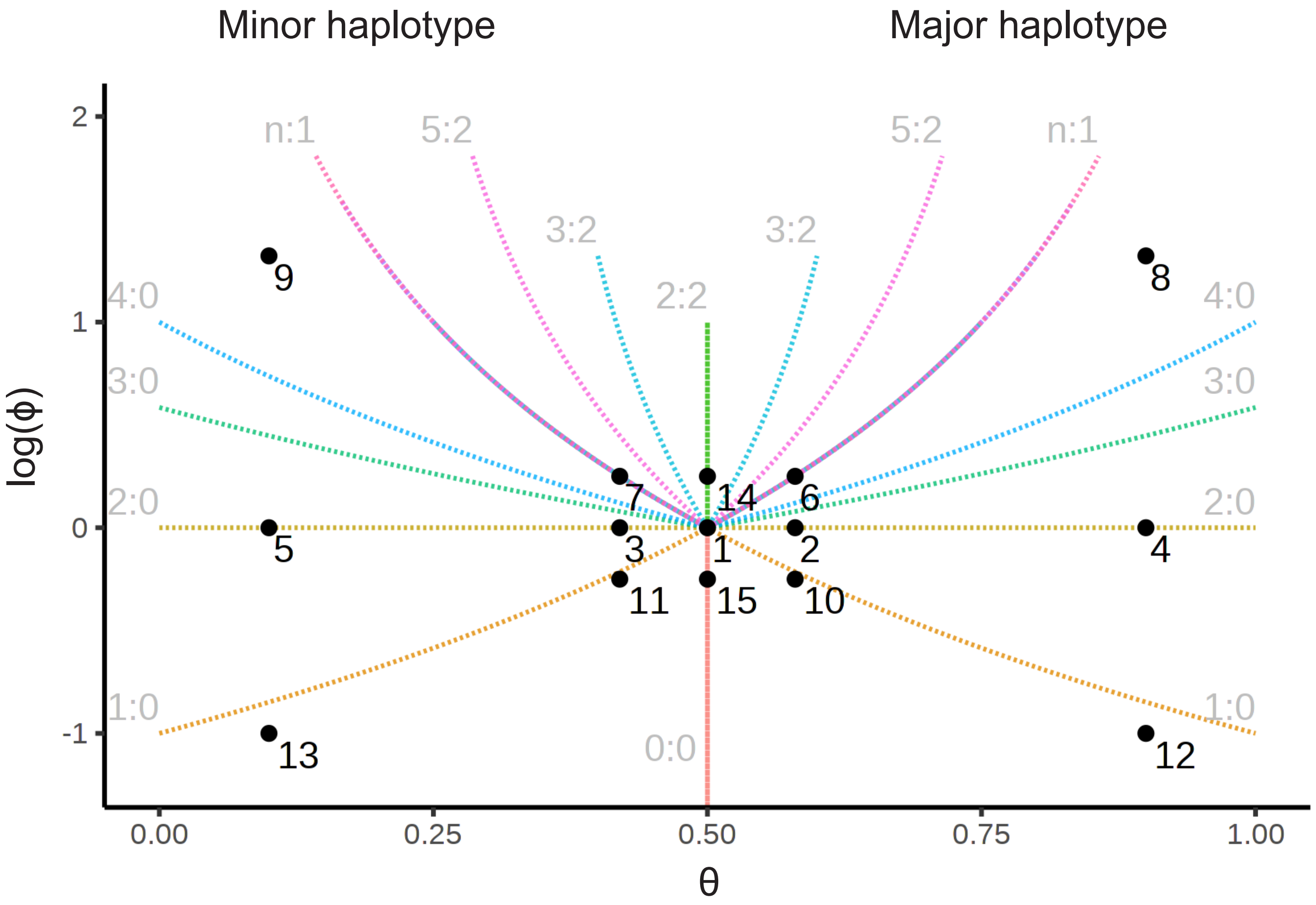
Supplementary Figure 3: ***In silico* serial dilution experiment of TNBC4 tumor and normal cells.** Top, chromosome arms with complete LoH (MAF > 0.95). Bottom, pseudobulk allele profiles of tumor-normal cell mixtures at various proportions. Heterozygous SNPs with complete LoH in the tumor are colored by their true haplotypes (major and minor; determined using observed allele counts in the tumor). The rest of the SNPs are colored in gray (undetermined). Gray vertical bars represent centromeres and gap regions.



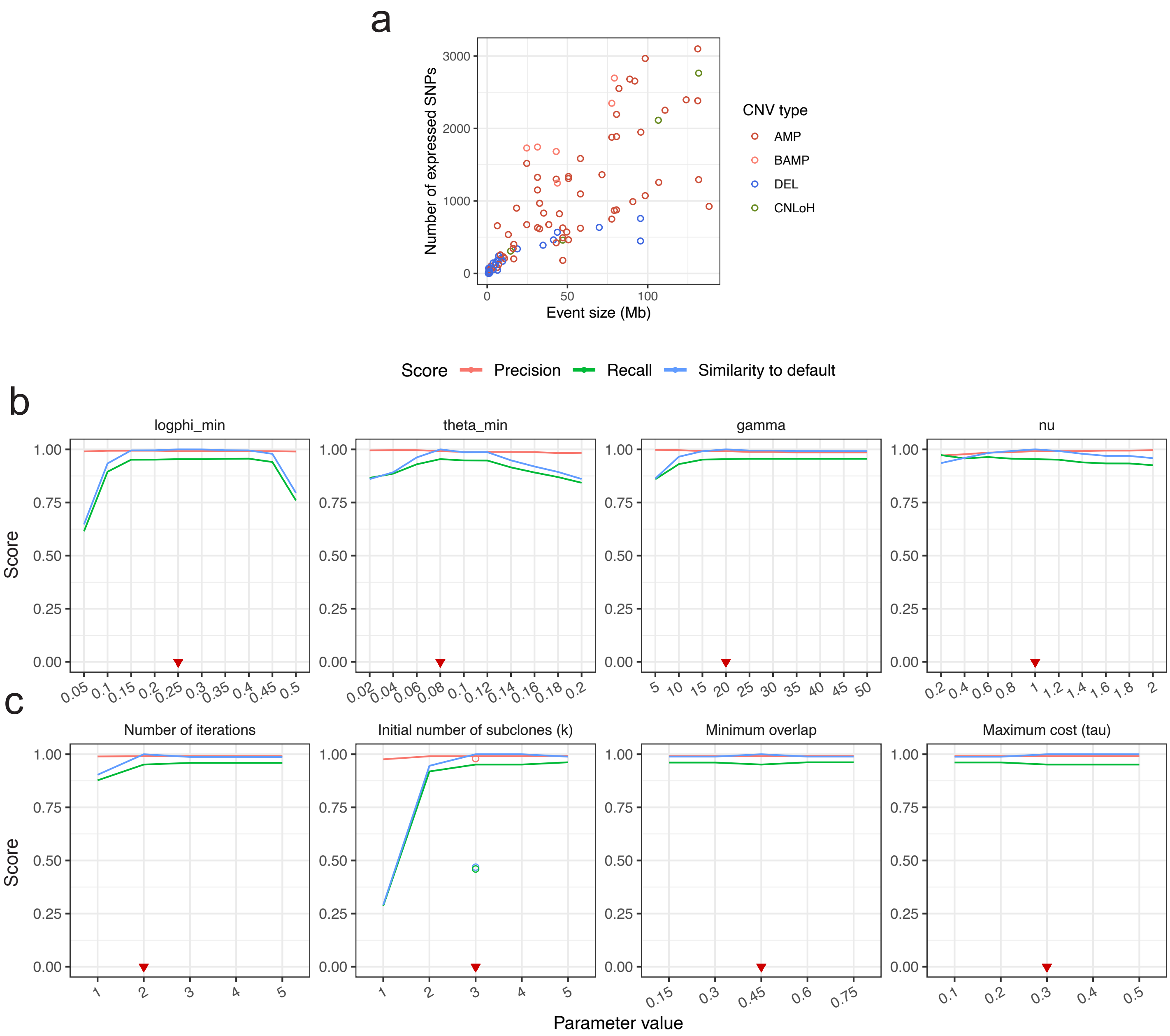
Supplementary Figure 4: ***In silico*** **serial dilution experiment of multiple myeloma (47491-Primary) tumor and normal cells.** Top, CNV events detected by WGS. Bottom, pseudobulk allele profiles of tumor-normal cell mixtures at various proportions. Heterozygous SNPs affected by allelic imbalances in the tumor are colored by their true haplotypes (major and minor; determined using observed allele counts in the tumor). The rest of the SNPs are colored in gray (undetermined). Gray vertical bars represent centromeres and gap regions.



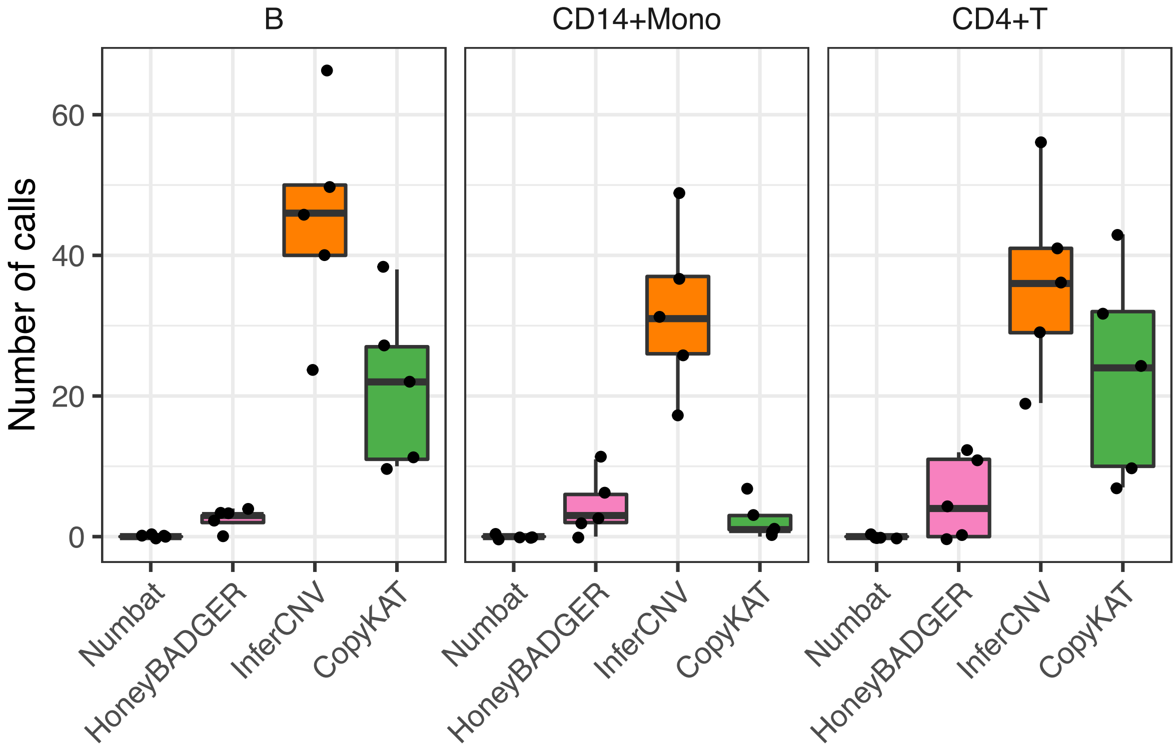
Supplementary Figure 5: Effect of population-based phasing on the detection of LoH and amplification events at different coverages. Performance of subclonal CNV detection from allele data in tumor-normal mixtures with and without haplotype phasing (“phasing” and “naive”). AUC, area under the ROC curve. **a**, Performance comparison for subclonal LoH detection in the TNBC4 dataset. **b**, Performance comparison for subclonal LoH detection in the multiple myeloma dataset. **c**, Performance comparison for subclonal amplification detection in the multiple myeloma dataset. Numbers in brackets denote mean SNP coverage per cell.



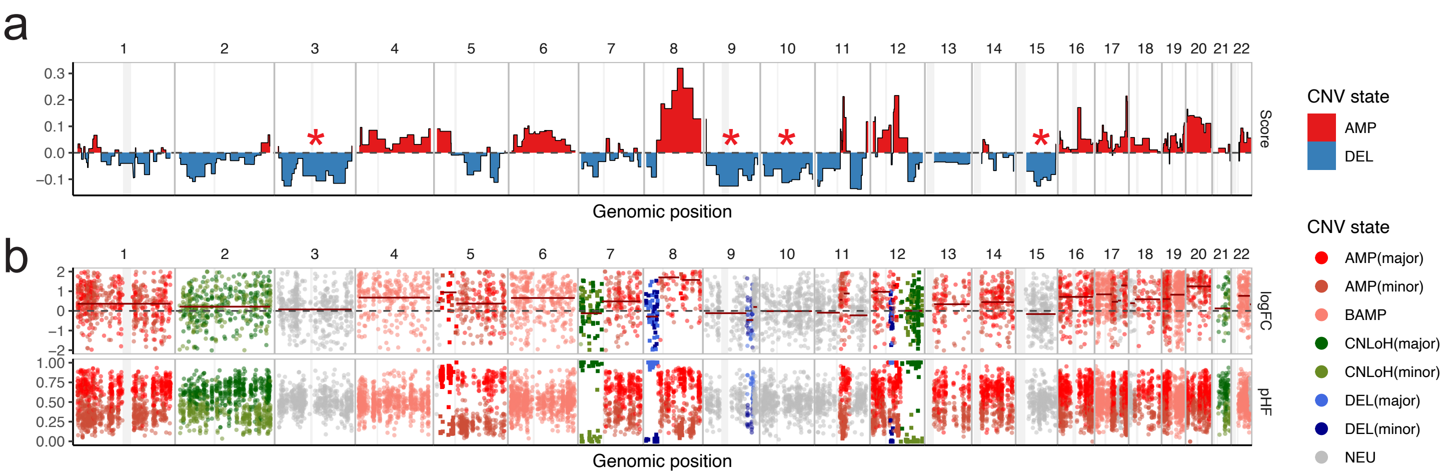
Supplementary Figure 6: Expected expression fold-change and allele fraction for different genotype configurations, cellular fraction, and haplotype state. Hidden states in the Numbat joint HMM and their respective parameter configurations are marked in black solid dots. Each dashed line represents the possible expression change and allele fraction for a given genotype depending on the cell fraction and haplotype state (major or minor). The genotype configuration corresponding to each line is marked in gray in the notation “paternal copies:maternal copies”. The homologous chromosome that has the higher number of copies is designated as the paternal chromosome.



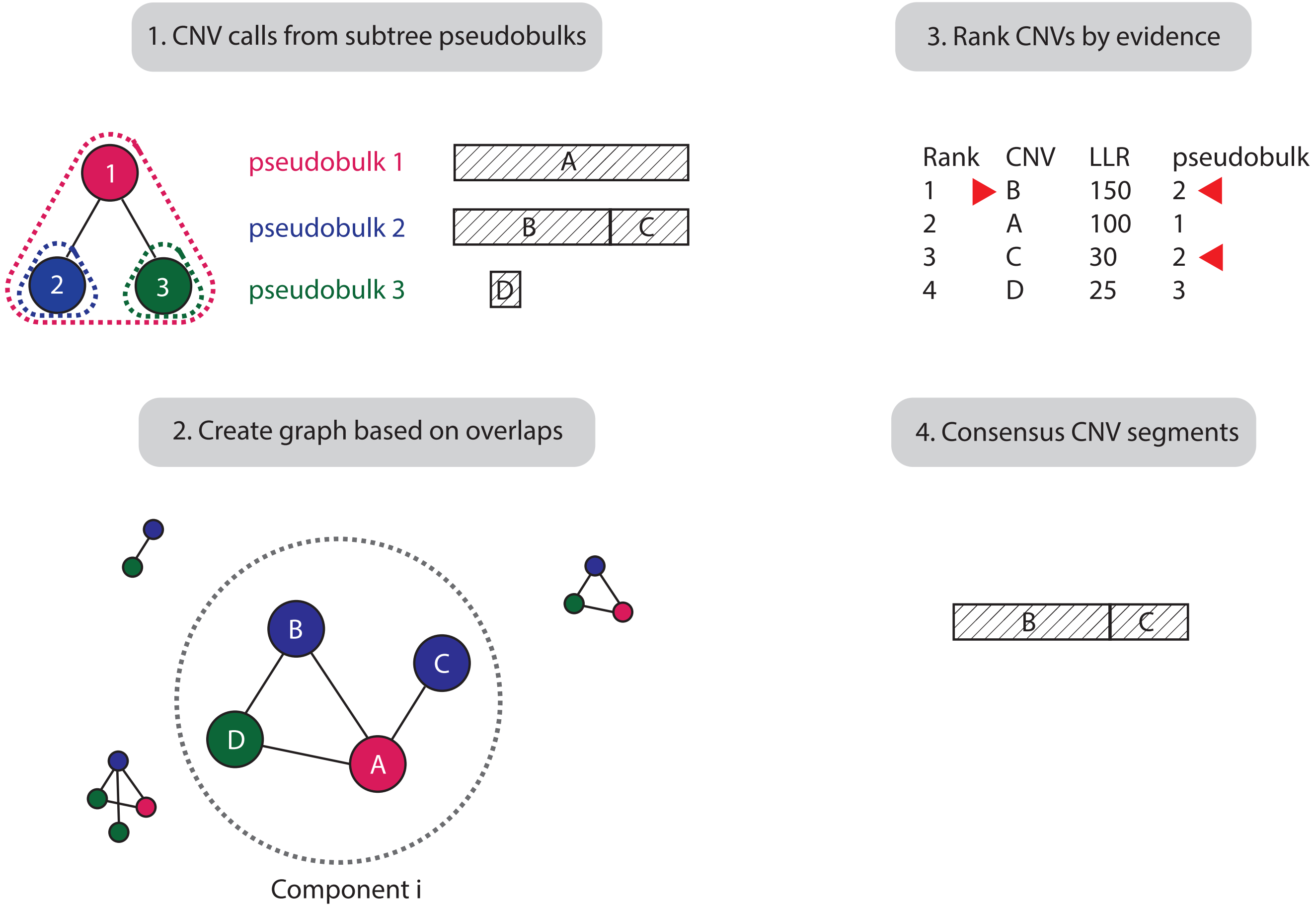
Supplementary Figure 7: Number of expressed SNPs per event and stability of joint HMM CNV calls with different parameter values in the multiple myeloma dataset. a, Number of expressed heterozygous SNPs per CNV region in the MM dataset. For each sample, CNV events were defined from matched WGS. Each dot represents a distinct CNV event. **b**, The effect of HMM-specific parameters. The joint HMM was run on pseduobulk profiles aggregating all tumor cells. **c**, The effect of parameters specific to iterative clonal decomposition. The full Numbat iterative algorithm was run on all cells (including tumor and normal cells). Jaccard similarity of CNV profiles was computed with respect to those of the default setting (marked by red triangles). Precision and recall were computed with respect to the ground-truth CNV profiles defined by WGS. Circles denote scores from initialization with a random tree.



Supplementary Figure 8: Number of false-positive CNV calls in non-aberrant cell populations in the multiple myeloma dataset. Each dot represents a distinct sample (n=5). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.

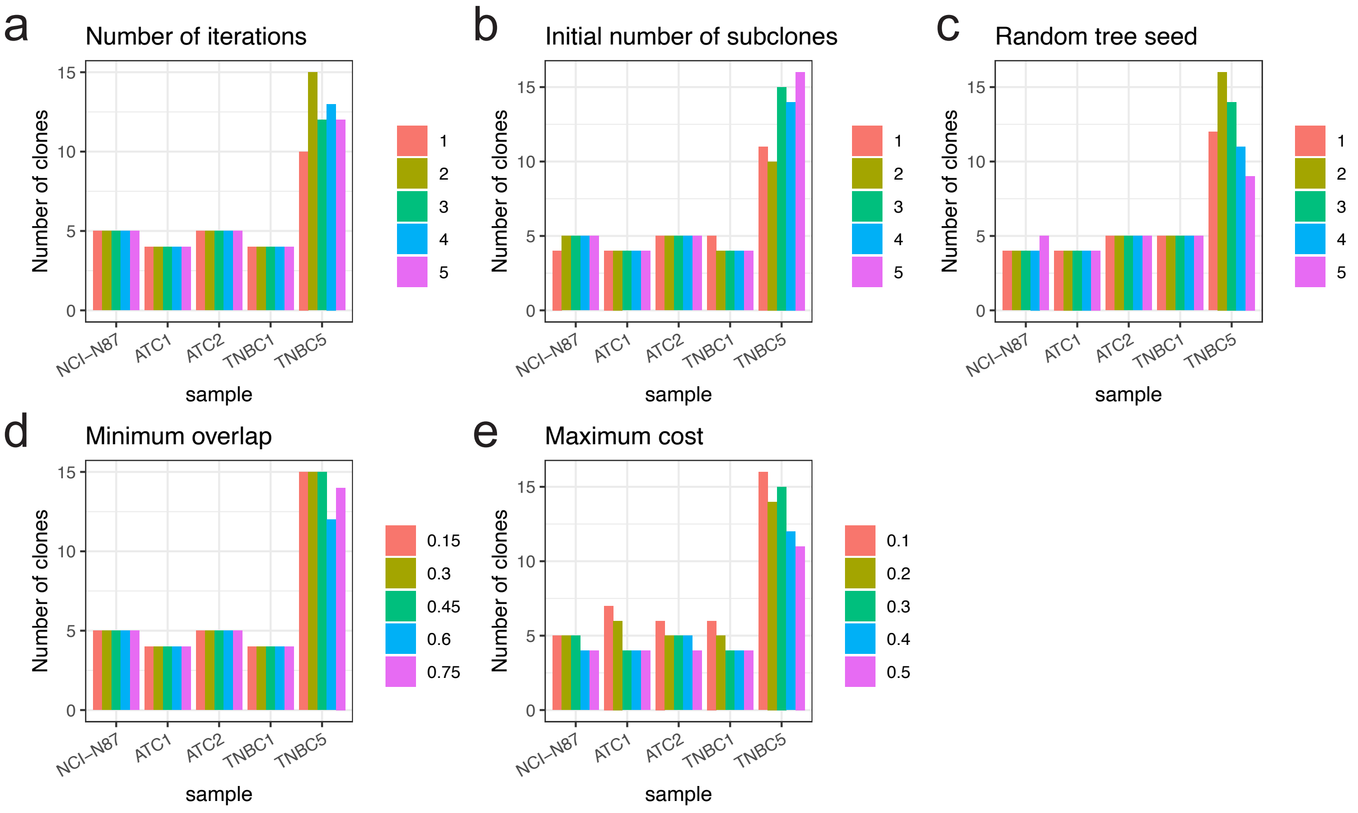


Supplementary Figure 9: Numbat and CopyKAT analysis of DCIS1 tumor cells. a, CNV profile inferred by CopyKAT. **b**, CNV profile inferred by Numbat joint HMM. BAMP, balanced amplification. Red asterisks mark diploid chromosomes that appear to have undergone a loss due to hyperdiploidy. Gray vertical bars represent centromeres and gap regions.

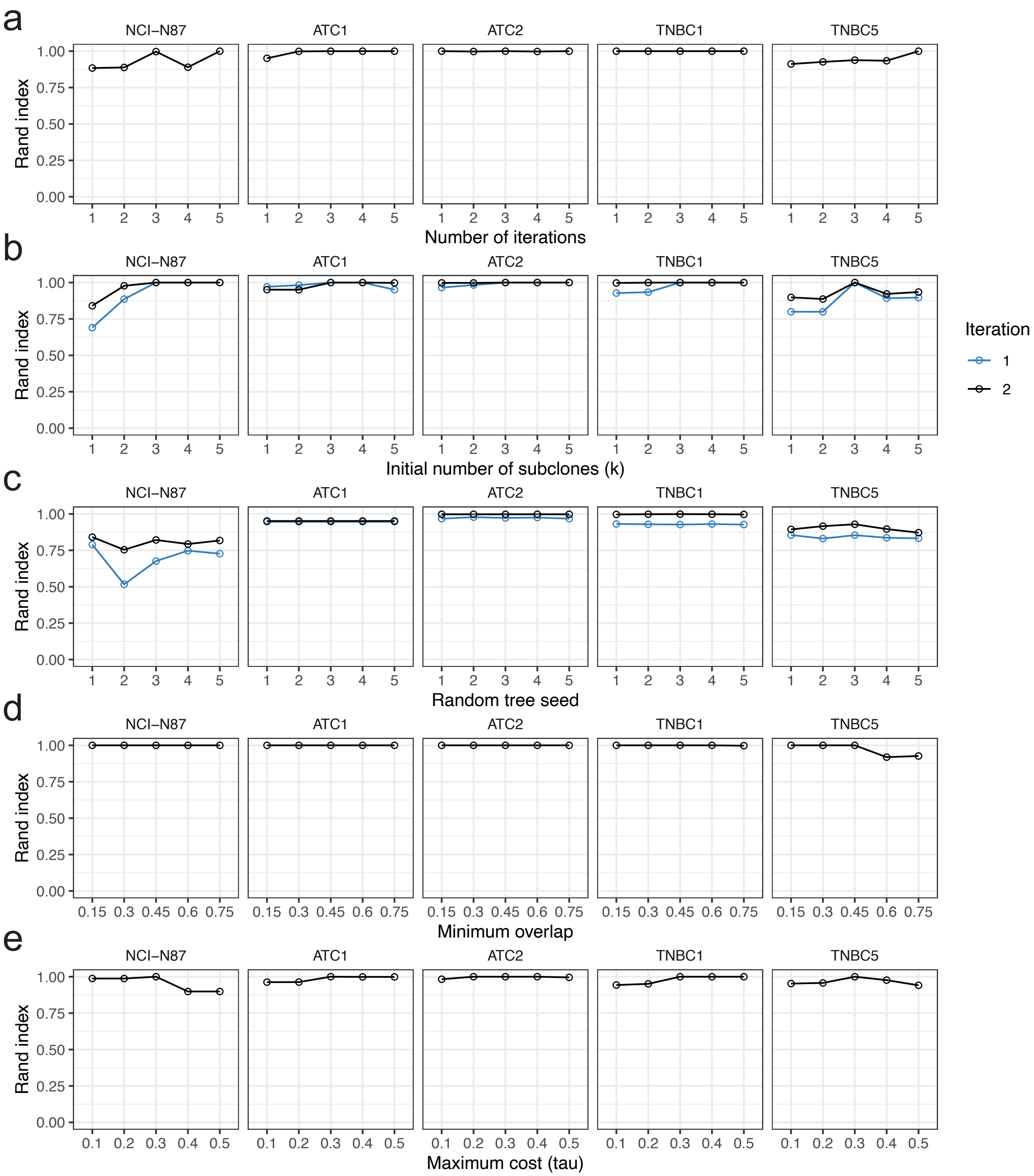

1. CNV calls from subtree pseudobulks
1
2
3
pseudobulk 1
pseudobulk 2
pseudobulk 3
A
B
C
D
2. Create graph based on overlaps
B
C
D
A
Component i
3. Rank CNVs by evidence
4. Consensus CNV segments
B
C

Rank		CNV	LLR	pseudobulk	
1	▶	B	150	2	◀
2		A	100	1	
3		C	30	2	◀
4		D	25	3	

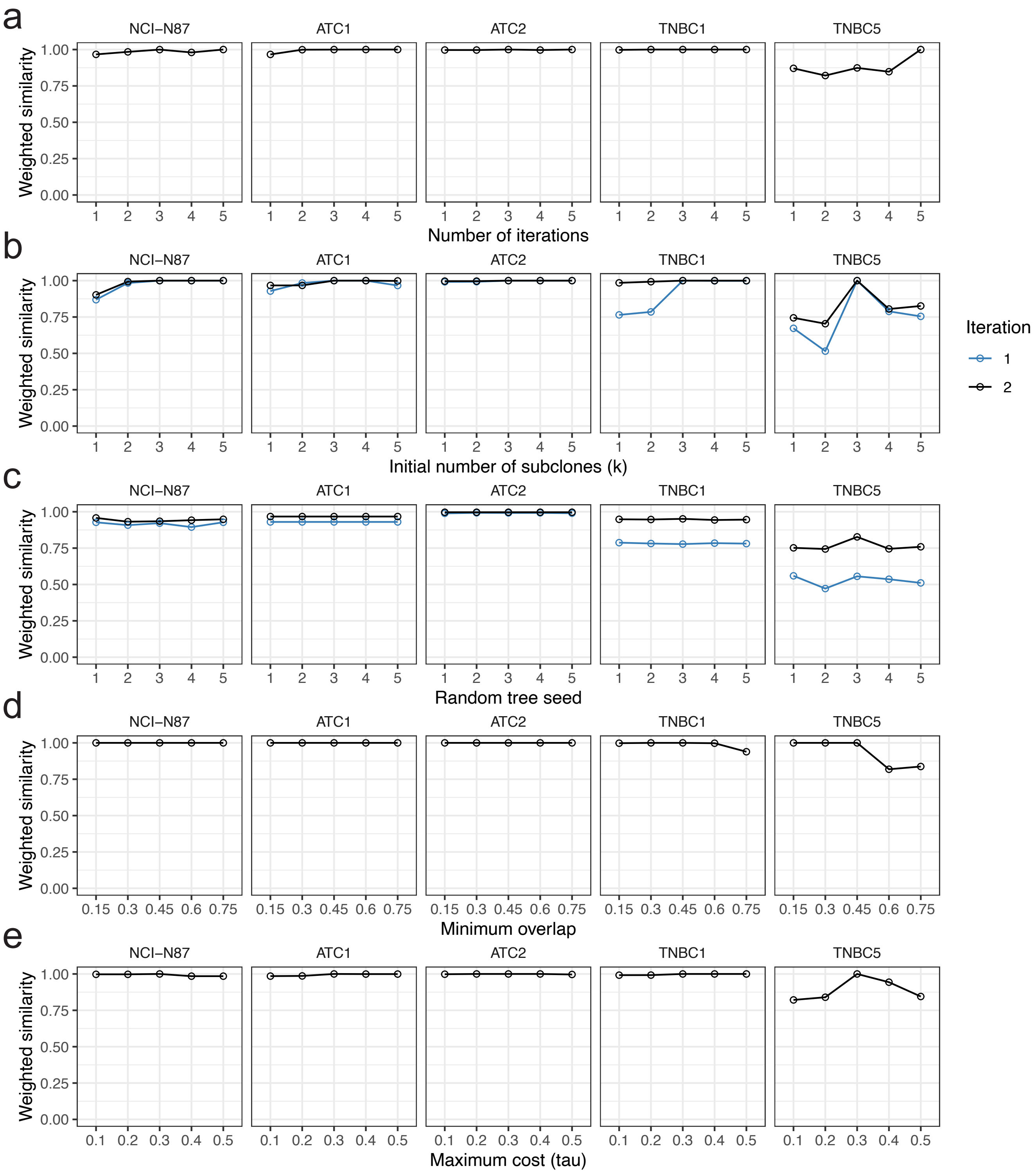
Supplementary Figure 10: Obtaining consensus CNV events from multiple cell populations. Step 1, HMMs are run independently on pseudobulk profiles formed from all possible subtrees in the phylogeny. Step 2, each detected CNV is represented as a node in a graph, and an edge is added between pairs of nodes if the two CNV segments significantly overlap. The nodes are then grouped by connected components of the resulting graph. Step 3, CNVs within the same component are ranked by likelihood evidence (combined LLR of expression and allele deviation). Step 4, all CNVs (within the same component) from the pseudobulk profile that harbors the top event are kept as part of the consensus segments.



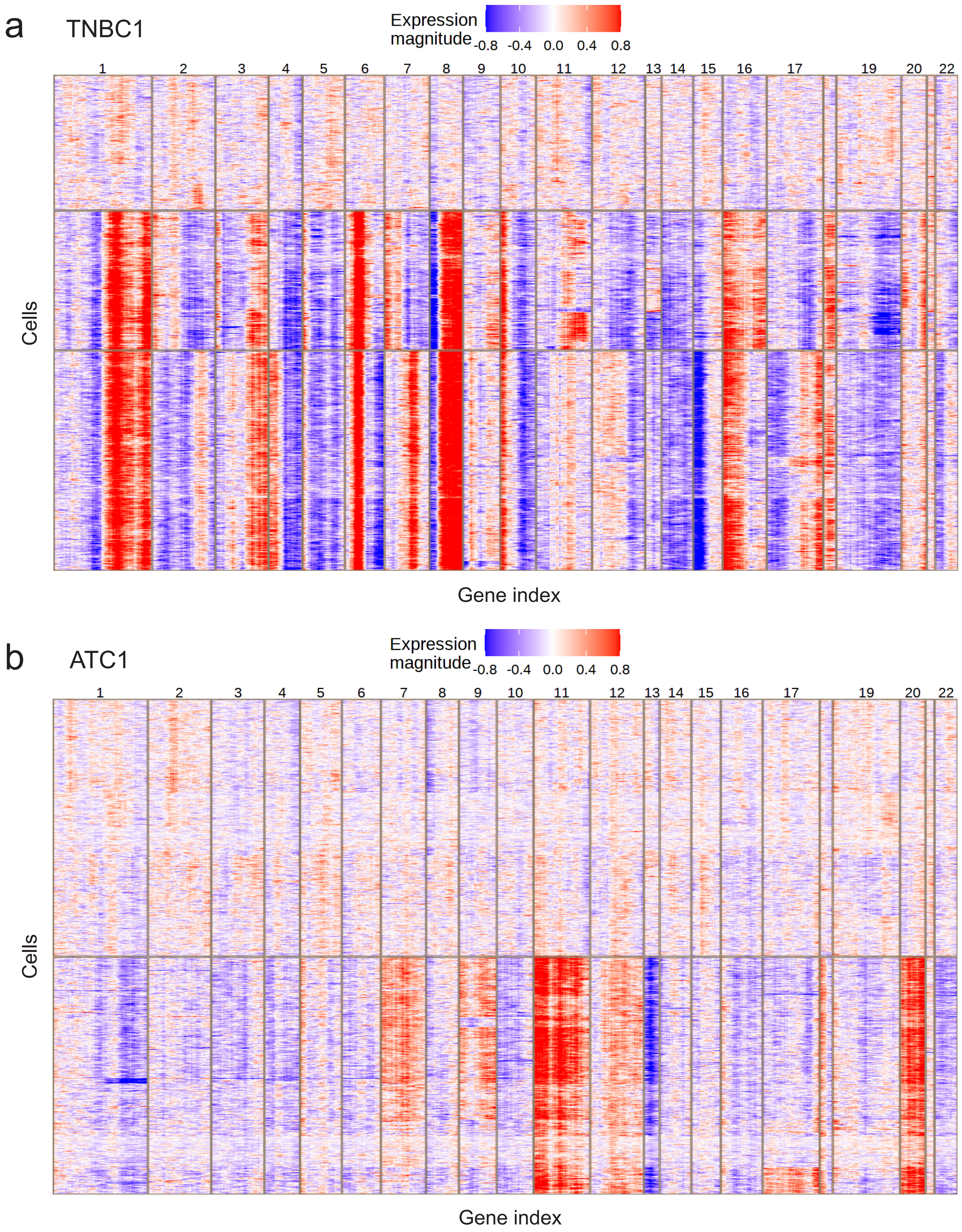
Supplementary Figure 11: Stability of the number of detected subclones with different initializations and run parameters. a, Effect of the number of iterations. **b**, Effect of the initial number of subclones (k). **c**, Effect of random initializations. Different random seeds were used to generate initial trees created from a random distance matrix. **d**, Effect of the minimum overlap threshold. **e**, Effect of the maximum cost parameter. For ATC2, the subsampled dataset was used in all experiments.



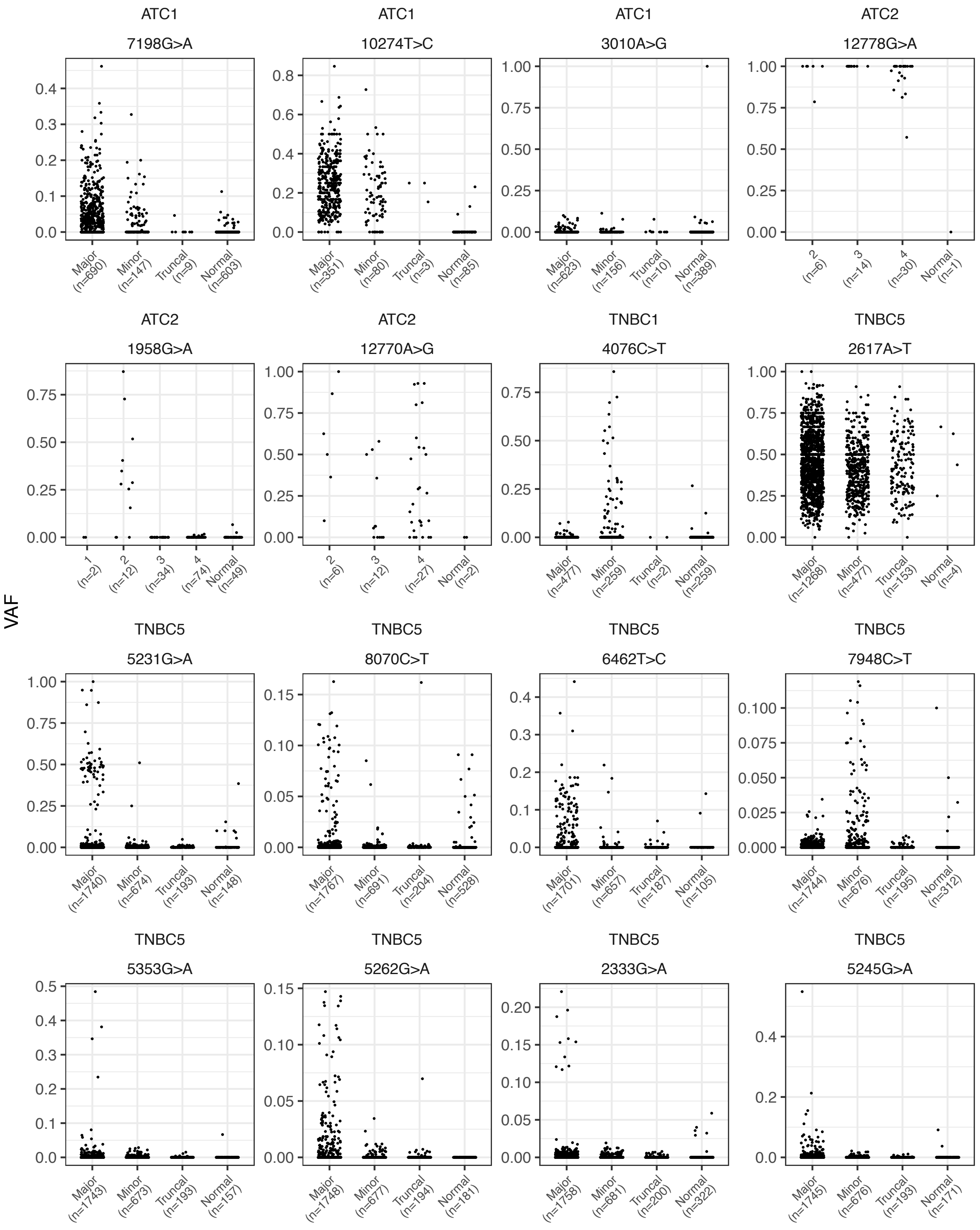
Supplementary Figure 12: Stability of subclone assignment with different initializations and run parameters. **a**, Effect of the number of iterations. Similarity of clone assignments was computed with respect to those of the last iteration. **b**, Effect of the initial number of subclones (k). Similarity of clone assignments was computed with respect to those of k = 3. **c**, Effect of random initializations. Different random seeds were used to generate initial trees created from a random distance matrix. Similarity of clone assignments was computed with respect to those obtained using the default initialization strategy (hierarchical clustering based on window-smoothed expression signals). **d**, Effect of the minimum overlap threshold. Similarity of clone assignments was computed with respect to those of minimum overlap = 0.45. **e**, Effect of the maximum cost parameter. Similarity of clone assignments was computed with respect to those of τ = 0.3. For ATC2, the subsampled dataset was used in all experiments.



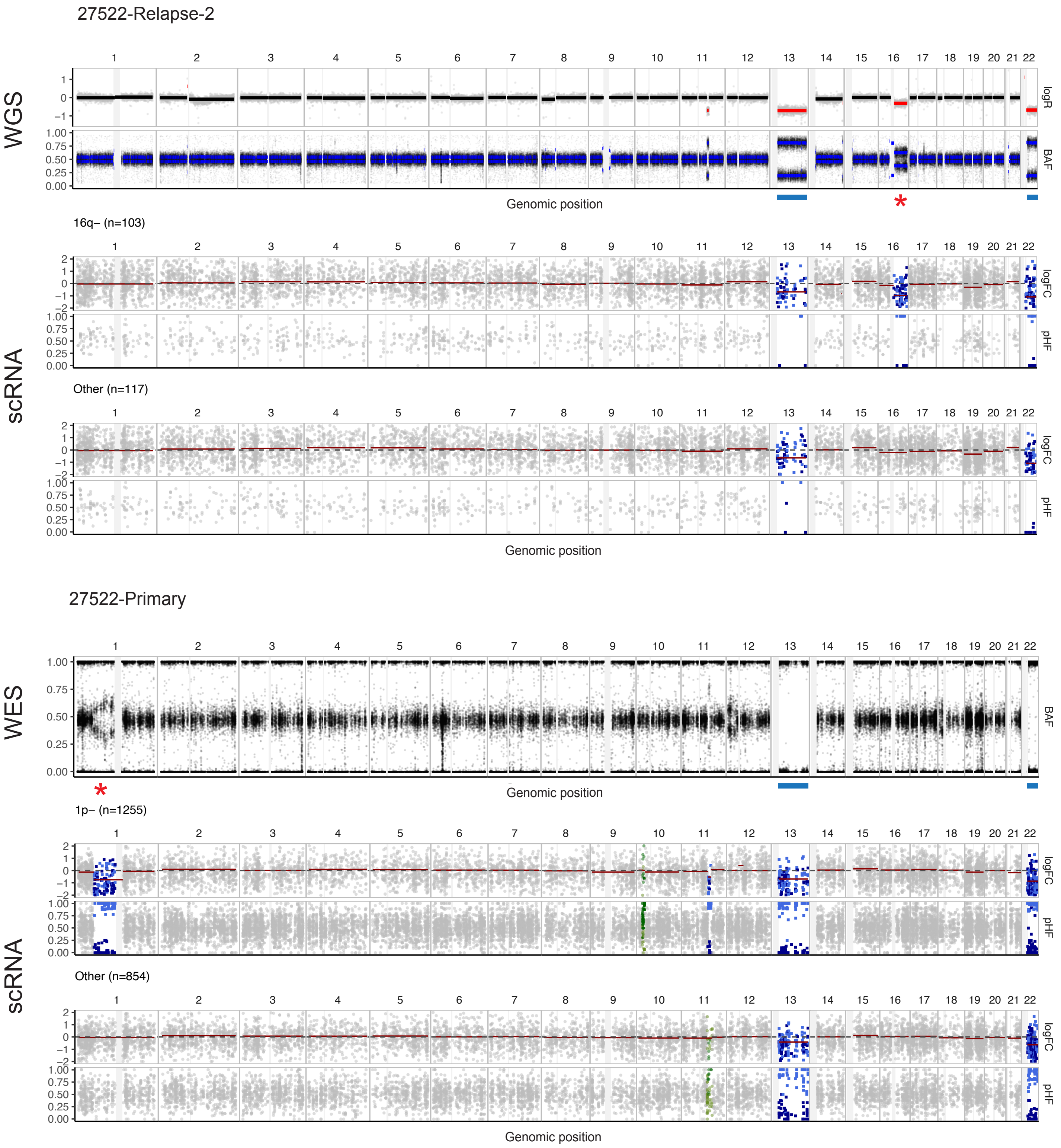
Supplementary Figure 13: Stability of subclone-specific CNV profiles with different initializations and run parameters. Weighted similarity was derived from pairwise comparisons between subclone CNV profiles from two different runs (Jaccard index), weighted by the proportion of cells assigned to the subclone pair in the respective runs. **a**, Effect of the number of iterations. Similarity of CNV calls was computed with respect to those of the last iteration. **b**, Effect of the initial number of subclones (k). Similarity of CNV calls was computed with respect to those of k = 3. **c**, Effect of random initializations. Different random seeds were used to generate initial trees created from a random distance matrix. Similarity of CNV profiles was computed with respect to those obtained using the default initialization strategy (hierarchical clustering based on window-smoothed expression signals). **d**, Effect of the minimum overlap threshold. Similarity of CNV calls was computed with respect to those of minimum overlap = 0.45. **e**, Effect of the maximum cost parameter. Similarity of CNV calls was computed with respect to those of τ = 0.3. For ATC2, the subsampled dataset was used in all experiments.



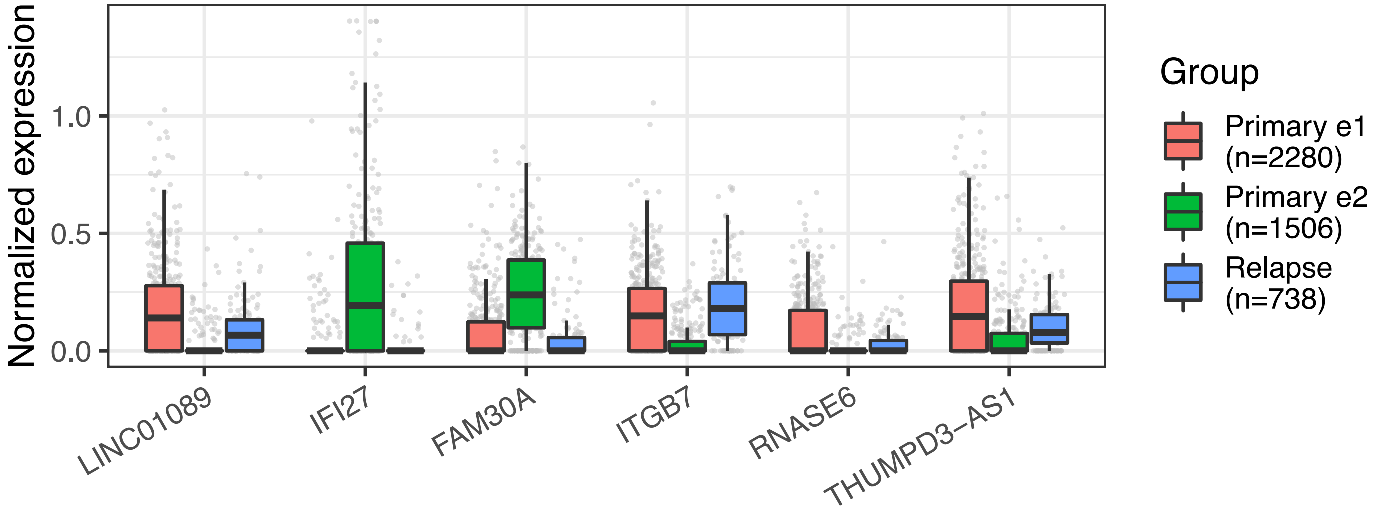
Supplementary Figure 14: Window-smoothed expression profile and hierarchical clustering of TNBC1 and ATC1 cells. a, Expression-based CNV profile of TNBC1. **b**, Expression-based CNV profile of ATC1**.** Cell clusters are marked by gray borders.



Supplementary Figure 15: VAF distribution of mtRNA mutations by CNV subclones. Each dot represents a distinct cell. Only cells where the variant is covered by at least 10 reads are shown. For ATC2, results from the subsampled dataset are shown.



Supplementary Figure 16: Validated subclonal CNVs in two samples. Subclone-specific copy number profiles reconstructed by Numbat from scRNA data are juxtaposed with the DNA profiles. logFC, log expression fold-change. pHF, paternal haplotype frequency. logR, log coverage ratio. BAF, B-allele frequency. Blue bars mark regions with clonal deletions in the tumor. Asterisks denote regions with subclonal CNV, as demonstrated by incomplete loss of heterozygosity in the DNA allele profile. Gray vertical bars represent centromeres and gap regions.



Supplementary Figure 17: DE markers of expression clusters in multiple myeloma patient 27522. Each dot represents a distinct cell. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range.

Sample ID	Patient ID	Cancer type	Number of cells	Protocol	Sorted WGS	Study
58408-SMM	58408	MM	2374	10X 3’ v2	Yes	Liu et al. 2021
58408-Primary	58408	MM	3880	10X 3’ v2	Yes	Liu et al. 2021
47491-SMM	47491	MM	1709	10X 3’ v2	Low purity	Liu et al. 2021
47491-Primary	47491	MM	1465	10X 3’ v2	Yes	Liu et al. 2021
37692-Primary	37692	MM	3078	10X 3’ v2	Yes	Liu et al. 2021
59114-Relapse-1	59114	MM	3413	10X 3’ v2	Yes	Liu et al. 2021
27522-Primary	27522	MM	4220	10X 3’ v2	No	Liu et al. 2021
27522-Remission	27522	MM	727	10X 3’ v2	No	Liu et al. 2021
27522-Relapse-1	27522	MM	2275	10X 3’ v2	No	Liu et al. 2021
27522-Relapse-2	27522	MM	3499	10X 5’	Yes	Liu et al. 2021
ATC1	ATC1	ATC	2203	10X 3’ v3	No	Gao et al. 2021
ATC2	ATC2	ATC	6226	10X 3’ v3	No	Gao et al. 2021
ATC3	ATC3	ATC	3264	10X 3’ v3	No	Gao et al. 2021
ATC4	ATC4	ATC	1731	10X 3’ v3	No	Gao et al. 2021
ATC5	ATC5	ATC	6144	10X 3’ v3	No	Gao et al. 2021
TNBC1	TNBC1	TNBC	1097	10X 3’ v2	No	Gao et al. 2021
TNBC2	TNBC2	TNBC	1034	10X 3’ v2	No	Gao et al. 2021
TNBC3	TNBC3	TNBC	532	10X 3’ v2	No	Gao et al. 2021
TNBC4	TNBC4	TNBC	3056	10X 3’ v3	No	Gao et al. 2021
TNBC5	TNBC5	TNBC	3225	10X 3’ v3	No	Gao et al. 2021
DCIS1	DCIS1	DCIS	1480	10X 3’ v2	No	Gao et al. 2021
NCI-N87	NCI-N87	GC	3246	10X 3' v2	No	Andor et al. 2020

Supplementary Table 1: Sample information and sequencing characteristics of analyzed scRNA-seq datasets.

Comparison	Gene set	P value	Q value	Enrichment score	Edge value
e2g1 vs e1g1	TNFA SIGNALING VIA NFKB	1.00E-04	0.00049995	1.522687672	0.83414457
e2g1 vs e1g1	CHOLESTEROL HOMEOSTASIS	0.00119988	0.004285286	1.246954138	0.83414457
e2g1 vs e1g1	IL2 STAT5 SIGNALING	0.01179882	0.0268155	0.988969146	1.718667352
e1g2 vs e1g1	E2F TARGETS	1.00E-04	0.00049995	3.382174569	0.26140133
e1g2 vs e1g1	G2M CHECKPOINT	1.00E-04	0.00049995	3.168155605	0.74799209
e1g2 vs e1g1	MITOTIC SPINDLE	1.00E-04	0.00049995	2.115124706	0.682944793
e1g2 vs e1g1	SPERMATOGENESIS	1.00E-04	0.00049995	1.451127988	1.459131697
e1g2 vs e1g1	ESTROGEN RESPONSE LATE	0.00109989	0.004582875	1.20248879	0.525201849
e1g2 vs e1g1	EPITHELIAL MESENCHYMAL TRANSITION	0.01339866	0.03349665	0.96220813	0.258354859
e1g3 vs e1g1	INTERFERON GAMMA RESPONSE	0.0009999	0.0119988	-1.240737475	-0.294176916

Supplementary Table 2: List of significantly enriched pathways in multiple myeloma patient 27522. P value, unadjusted P values from gene set enrichment analysis (two-sided). Q value, adjusted P values using the Benjamini-Hochberg method.

	NEU	CNLOH1(major)	CNLOH1(minor)	CNLOH2(major)	CNLOH2(minor)	AMP1(major)	AMP1(minor)	AMP2(major)	AMP2(minor)	DEL1(major)	DEL1(minor)	DEL2(major)	DEL2(minor)	BAMP	BDEL
NEU	1-t	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
CNLOH1(major)	t*0.33	(1-t)*(1-p_s)	(1-t)*p_s	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
CNLOH1(minor)	t*0.33	(1-t)*p_s	(1-t)*(1-p_s)	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
CNLOH2(major)	t*0.25	t*0.25/2	t*0.25/2	(1-t)*(1-p_s)	(1-t)*p_s	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*2.5e-05	t*2.5e-11
CNLOH2(minor)	t*0.25	t*0.25/2	t*0.25/2	(1-t)*p_s	(1-t)*(1-p_s)	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*2.5e-05	t*2.5e-11
AMP1(major)	t*0.33	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	(1-t)*(1-p_s)	(1-t)*p_s	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
AMP1(minor)	t*0.33	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	(1-t)*p_s	(1-t)*(1-p_s)	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
AMP2(major)	t*0.25	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	(1-t)*(1-p_s)	(1-t)*p_s	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*2.5e-05	t*2.5e-11
AMP2(minor)	t*0.25	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	(1-t)*p_s	(1-t)*(1-p_s)	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*2.5e-05	t*2.5e-11
DEL1(major)	t*0.33	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	(1-t)*(1-p_s)	(1-t)*p_s	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
DEL1(minor)	t*0.33	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	t*0.33/2	t*0.33/2	t*3.3e-11/2	t*3.3e-11/2	(1-t)*p_s	(1-t)*(1-p_s)	t*3.3e-11/2	t*3.3e-11/2	t*3.3e-05	t*3.3e-11
DEL2(major)	t*0.25	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	(1-t)*(1-p_s)	(1-t)*p_s	t*2.5e-05	t*2.5e-11
DEL2(minor)	t*0.25	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	(1-t)*p_s	(1-t)*(1-p_s)	t*2.5e-05	t*2.5e-11
BAMP	t*0.25	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	1-t	t*2.5e-11
BDEL	t*0.25	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*0.25/2	t*0.25/2	t*2.5e-11/2	t*2.5e-11/2	t*2.5e-05	1-t

Supplementary Table 3: Transition probability matrix in the joint Numbat HMM. t, copy number transition probability. p_s, phase switch probability. BAMP, balanced amplification. BDEL, balanced deletion.

Supplementary Methods

Derivation of the gene expression count model. Given a reference expression profile $\overrightarrow{\lambda^*} = (\lambda_1^*, \lambda_2^*, \dots, \lambda_N^*)$, we can model the expression magnitudes of the observed cell or pseudobulk using the following additive equation:

$$\log \lambda_i = \mu + \log \lambda_i^* + \log \phi + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

where the factors that influence expression magnitudes include fixed components μ, λ_i^*, ϕ and the random noise component ϵ_i. Here λ_i^* represents the baseline expression magnitude. μ captures the systematic bias in expression magnitudes between the reference and observation (e.g., global up or down transcriptional regulation), and ϕ captures the effect of chromosomal dosage. σ^2 captures the expression variation in the comparison, which intuitively reflects the magnitude of differences between the reference and observation expression profiles (e.g., inter-individual differences if the reference is from the same cell type but a different subject, or cell type differences if the reference belongs to a difference cell type). Note that λ_i^* and ϵ_i are gene-specific, since each gene in the observation has its own baseline expression level and transcriptional noise in the measurement. ϕ is shared between genes in the same copy number segment. μ and σ^2 are shared across all genes and capture the global bias and variance between the reference and observation expression magnitudes. The model can be extended to incorporate gene-specific bias and variance when technical replicates of the reference and observation are available. Since we typically do not have access to technical replicates, we assume homoskedasticity and use the same set of μ and σ^2 for all genes.

We note that the expression magnitudes λ_i are not directly observed. Instead, we observe discrete counts X_i, which can be assumed to follow a Poisson distribution with rate λ_i:

$$X_i \sim \text{Pois}(l\lambda_i)$$

Here l is the total library size. This gives rise to the following Poisson-Lognormal (PLN) mixture model as stated in Equation (1):

$$X_i \sim \text{PoisLogNorm}(\mu + \log(l\lambda_i^*) + \log \phi, \sigma^2)$$

Reference gene expression profile. Since gene expression varies substantially across tissue and cell types, comparing the observed expression profile with the expected expression profile of the same tissue or cell type (without aneuploidy) can help reveal CNV signal. We denote the reference profile as $\overrightarrow{\lambda^*} = (\lambda_1^*, \lambda_2^*, \dots, \lambda_N^*)$ which pre-specifies the background expression magnitudes in the model described above. $\overrightarrow{\lambda^*}$ can be obtained from the expression magnitudes estimated from a single reference cell type or a collection of reference cell types, which we can represent as matrix Λ^* of dimension $N \times C$ where C is the number of reference cell types. While analyzing an observed pseudobulk expression profile, we can model the pseudobulk as a mixture of the reference cell types where the cell type proportions (denoted as $\vec{w}$) are unknown. We therefore create $\overrightarrow{\lambda^*}$ from a convex combination of Λ^* that minimizes the least squared error between the two expression profiles in log scale:

$$\vec{w}^{\dagger} = \text{argmin}_{\vec{w}} || \log(\Lambda^* \vec{w}) - \log(\vec{X}/l) ||_2^2$$
$$\text{Subject to: } |\vec{w}| = 1; w_k \geq 0, \forall k \in \{1 \dots C\}$$
$$\vec{\lambda}^* = \Lambda^* \vec{w}^{\dagger}$$

where only genes with non-zero expression in the pseudobulk and reference profiles are considered in the optimization. For CNV evaluation in single cells, we choose a single best reference cell type that maximizes the correlation with the observed expression magnitudes after $\log(x+1)$ transformation. By default, the Numbat package uses a collection of reference profiles from the Human Cell Atlas lung study[1]. In practice, the user may wish to provide custom reference profiles created from data generated from the same sequencing platform or within the same processing batch in order to minimize noise due to technical factors.

Overdispersion parameter in the allele count model. We note that although the degree of overdispersion in allele counts can vary between genes, we do not have technical replicates of the same cell or cell populations to estimate gene-specific overdispersion. Therefore, the γ parameter in Equation (3) is shared across all genes:

$$Y_j \sim \text{BetaBinom}\big(m_j, \theta\gamma, (1-\theta)\gamma\big)$$

If diploid regions are known, γ can be estimated using maximum likelihood for specific pseudobulk profiles. However, inference of diploid regions relies on detection of allelic imbalance. We therefore fix $\gamma = 20$ which is suitable for most pseudobulk sizes, as estimated from the normal cells in the TNBC4 dataset.

Configuration of the Hidden Markov model. The 15 states in the haplotype-aware HMM and their respective parameter configurations are listed in the table below.

State ID	Copy number state	Cell fraction	Haplotype state	$\log(\phi)$	θ
1	Neutral	-	-	0	0.5
2	CNLoH	Low	Major	0	$0.5 + \theta_{\text{min}}$
3	CNLoH	Low	Minor	0	$0.5 - \theta_{\text{min}}$
4	CNLoH	High	Major	0	0.9
5	CNLoH	High	Minor	0	0.1
6	Amplification	Low	Major	$\log \phi_{\text{min}}$	$0.5 + \theta_{\text{min}}$
7	Amplification	Low	Minor	$\log \phi_{\text{min}}$	$0.5 - \theta_{\text{min}}$
8	Amplification	High	Major	$\log(2.5)$	0.9
9	Amplification	High	Minor	$\log(2.5)$	0.1
10	Deletion	Low	Major	$-\log \phi_{\text{min}}$	$0.5 + \theta_{\text{min}}$
11	Deletion	Low	Minor	$-\log \phi_{\text{min}}$	$0.5 - \theta_{\text{min}}$
12	Deletion	High	Major	-1	0.9
13	Deletion	High	Minor	-1	0.1

14	Balanced amplification	-	-	$\log \phi_{\min}$	0.5
15	Homozygous deletion	-	-	- $\log \phi_{\min}$	0.5

The set of parameters for each state is chosen to capture one or multiple copy number configurations that exhibit similar changes in expression magnitude and allele frequency at a given cell fraction and haplotype state. States with haplotype imbalance appear in pairs (e.g., states 2 and 3) that have opposite deviations of haplotype fractions, which correspond to major and minor haplotype states. For example, state pairs (2,3) and (4,5) respectively correspond to CNLoH at low and high cell fractions, whereas state pair (8,9) is aimed to capture multiple amplified states with high allelic imbalance (e.g., 3:1, 4:1, 3:0, 4:0). In addition, we introduce fixed prior abundances to each of the states: $\pi(z) = \{1 = 0.25, 2 = 0.125, 3 = 0.125, 4 = 0.125 \cdot 10^{-10},\ 5 = 0.125 \cdot 10^{-10},\ 6 = 0.125,\ 7 = 0.125,\ 8 = 0.125 \cdot 10^{-10},$ $9 = 0.125 \cdot 10^{-10},\ 10 = 0.125, 11 = 0.125, 12 = 0.125 \cdot 10^{-10},\ 13 = 0.125 \cdot 10^{-10},\ 14 = 0.25 \cdot 10^{-4},$ $15 = 0.25 \cdot 10^{-10}\}$.

The allele-only HMM and the expression-only HMM are special cases of the joint HMM defined above. The allele-only HMM includes a subset of the states (1-5) and only uses the allele counts, whereas the expression-only HMM only uses the gene expression counts and does not allow transition between haplotype states.

Detecting regions with clonal deletion. In samples with high tumor purity (e.g., tumor cell lines) without matched normal cells, heterozygous SNPs are challenging to identify in regions of LoH, leading to decreased power of detection. We therefore provide a separate HMM module to identify clonal deletions based on heterozygous SNP density when normal diploid cells are not available. The HMM allows two hidden states (neutral and clonal deletion), where each gene emits an expression read count X_i and heterozygous SNP count V_i. The emission probabilities of the expression read counts (X_i) follow the same definition in the joint HMM described before. We model the heterozygous SNP count per gene (V_i) using a Negative Binomial distribution:

$$V_i | Z_i = z \sim \mathrm{NBinom}(u_z \cdot h_i, \sigma^2)$$

where h_i is the gene length in Mb, u is the heterozygous SNP density per Mb, and σ^2 is the variance in heterozygous SNP density along the transcriptome. We first fit a null model using the observed heterozygous SNP counts in all genes using maximum likelihood to obtain estimates $\hat{u}$ and $\hat{\sigma}^2$. We then let $u_{\mathrm{neu}} = \hat{u}$ and $u_{\mathrm{del}} = 5$. The transition probabilities are determined by a single parameter t.

Identifying diploid regions. We use a graph-based clustering approach to identify genomic regions in the diploid (neutral) state from a given pseudobulk profile. First, regions of allelic imbalance are identified by the allele-only HMM and excluded. The remaining allelically balanced segments are assumed to be in even-valued copy number states[2]. We then perform a pairwise comparison of the log expression fold-change (logFC) of the balanced segments using Student's t-test. We construct a graph where the nodes represent the balanced segments and the edges are determined by the adjusted *P* values (alpha level of 10^{-4}) from the previous step. An edge connecting two segments means that their expression magnitudes are not significantly different, and that they likely occupy the same total copy number state. We note that a clique in such a graph would be a collection of segments that are most coherent in total copy number. Consequently, the diploid segments can be inferred as the maximum clique with the lowest average logFC[3]. An extension of this procedure can be used to identify shared diploid regions in multiple pseudobulk profiles (e.g., representing different subclonal populations), where balanced segments are compared within each pseudobulk profile and a Simes' test is used to determine edges in the graph.

Initial approximation of single-cell phylogeny and subclonal structure. By default, Numbat builds an initial phylogeny using window-smoothed expression signals. First, normalized gene expression levels in transcript per million are $\log(\mathrm{x}+1)$ transformed. The expression signals are then smoothed by a running mean procedure using a window size of 101 genes[3]. Hierarchical clustering (Ward's minimum variance method) is then applied on the smoothed single-cell expression profiles to obtain a hierarchical cluster tree. Initial subtrees and subclones are determined by cutting the tree (*cutree* function in R) into k groups (default: $k = 3$). The initial number of subclones k can be set based on prior knowledge of the number of subclones, which can be used to obtain more accurate approximations in the initial iteration.

Obtaining consensus CNV events from cell subpopulations. To obtain consensus CNV segments from different cell subpopulations while leveraging their phylogenetic relationships, we apply the following heuristic procedure (Supplementary Fig. 10). First, HMMs are run independently on pseudobulk profiles formed from all possible subtrees in the phylogeny. Each CNV is represented as a node in a graph, and an edge is added between pairs of nodes if the two CNV segments significantly overlap (e.g., length of the overlap is more than 45% in either segment). The nodes are then grouped by connected components of the resulting graph. Within each component, the CNVs are ranked by likelihood evidence (combined LLR of expression and allele deviation). All CNVs (within the same component) from the pseudobulk profile that harbors the top event are kept as part of the consensus segments. For instance, let there be three cell populations where population 1 is the ancestor of population 2 and 3 (Supplementary Fig. 10). We aggregate cells by subtrees defined by each node (i.e., subpopulation) in the lineage hierarchy, so that pseudobulk 1 contains cells from all three populations, pseudobulk 2 contains cells from only population 2, and pseudobulk 3 contains cells from only population 3. Thus, HMM on each pseudobulk achieves joint copy number segmentation of cells under the same lineage, which may harbor common CNV breakpoints. Let A be an event called from pseudobulk 1 and B, C be events called in pseudobulk 2 and D be an event called in pseudobulk 3. This situation can arise when higher-resolution segmentations are produced by lower subtrees, which are purer in terms of clonal composition (e.g., pseudobulk 2). Event D could be an artifact that does not have strong evidence. The four events form a connected component in the graph due to the overlaps between (B,D), (A,D), (A,C), and (A,B). Since event B from pseudobulk 2 has the highest likelihood evidence and C is called from the same pseudobulk, we include both B and C in the consensus CNV call set. Lastly, we retest the remaining regions (detected as aberrant in any subtree but not covered by the consensus calls) in each subtree to obtain the final consensus CNV profile.

Mutation placement on the phylogeny. Although the mutation placement on the single-cell phylogeny by maximum likelihood is optimal in terms of the goodness of fit, it is not necessarily the most parsimonious (i.e., it tends to produce intermediate genotypes for which there is not enough evidence in order to gain better fit to the observed data). We therefore adopt the following refinement procedure to enforce evolutionary parsimony. After identifying the optimal mutation placement on the tree by maximum likelihood, we iteratively simplify the mutational history (i.e., reduce the number of evolutionary steps) by re-assigning a mutation to the same branch as another mutation that occurred directly upstream or downstream, effectively collapsing an internal branch. The cost for such re-assignment is defined by the corresponding decrease in the genotype likelihood. We iteratively perform the least costly reassignment until a prespecified maximum cost threshold is reached. For example, if the maximum likelihood mutation assignment infers a mutation history of A → B → C, we simplify the mutation history as A,B → C or A → B,C (i.e., reassigning mutation B to the same branch as A or C, effectively collapsing the A → B branch or B → C branch) if there is no sufficient likelihood evidence to ascertain that mutation B was acquired after A or before C. In practice, we determine the maximum reassignment cost as a function of the number of cells ($n \cdot \tau$), where τ reflects the degree of parsimony in simplifying the phylogeny (default: $\tau = 0.3$). A higher τ produces a more simplified mutational history with fewer evolutionary steps and intermediate genotypes.

The final phylogeny specifies not only the lineage relationship between cells but also the genotypes of each subclone (defined as a paraphyletic or monophyletic group in the phylogeny sharing the same mutation profile). The tumor lineage can be identified as the clade in the phylogeny with the highest mutation burden (total number of mutations across all cells in the clade).

Supplementary References

1. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).
2. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
3. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).