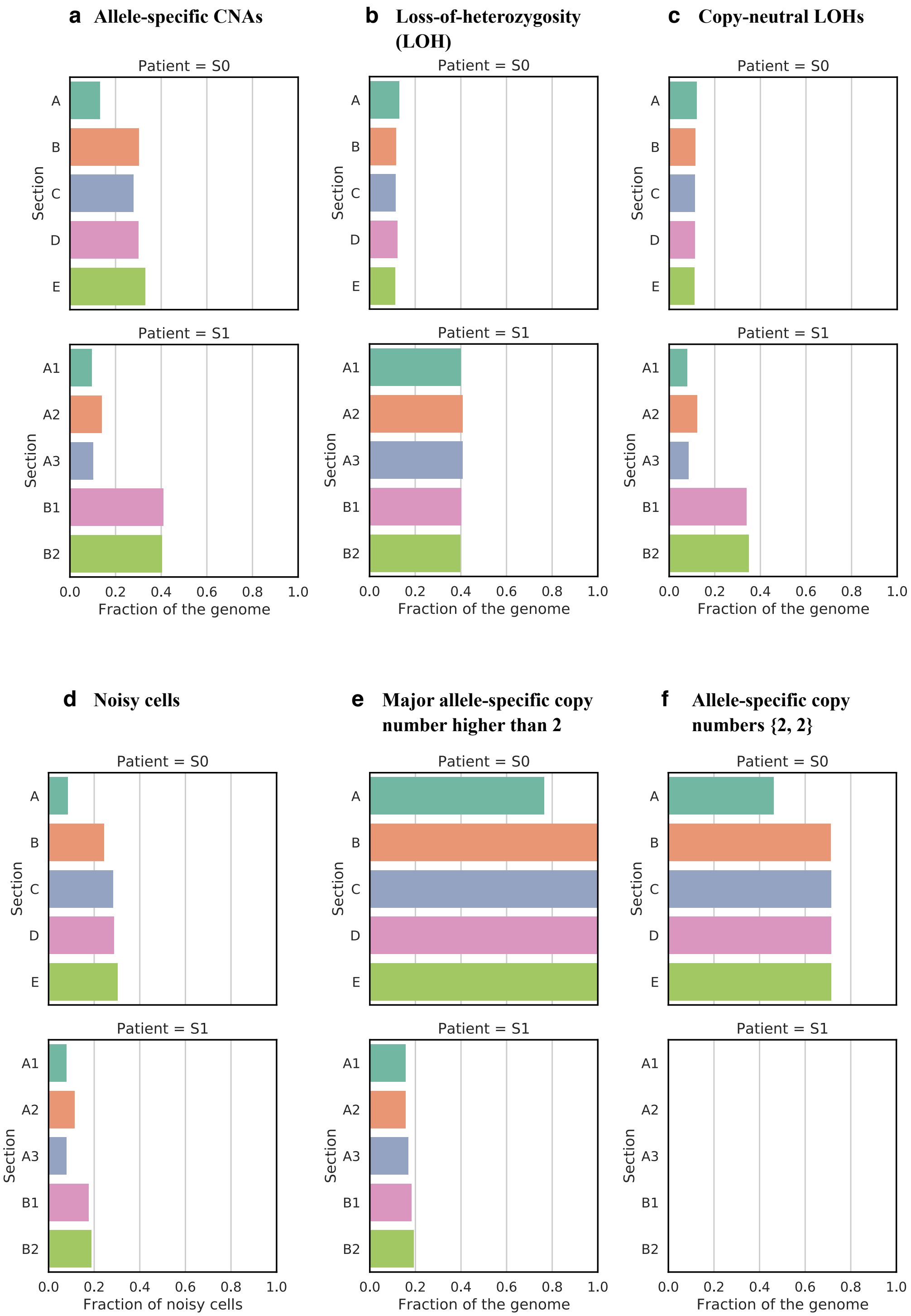
Supplementary Information

Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL

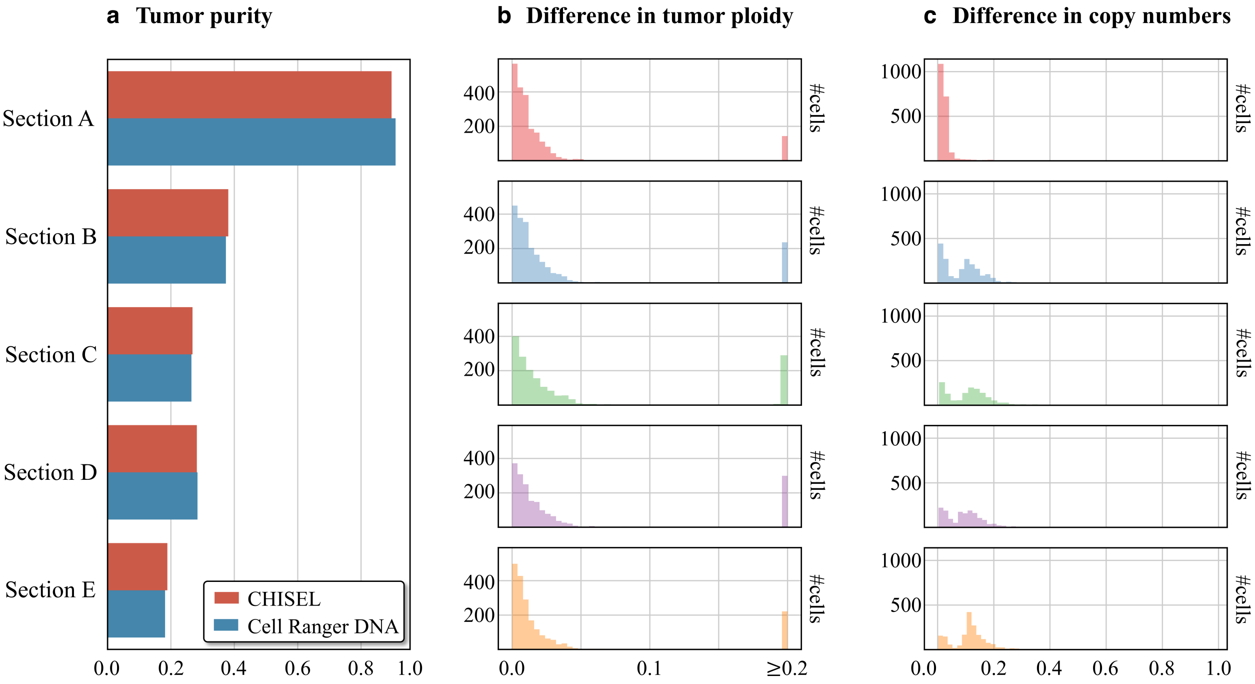
Contents

Supplementary Figures **3**

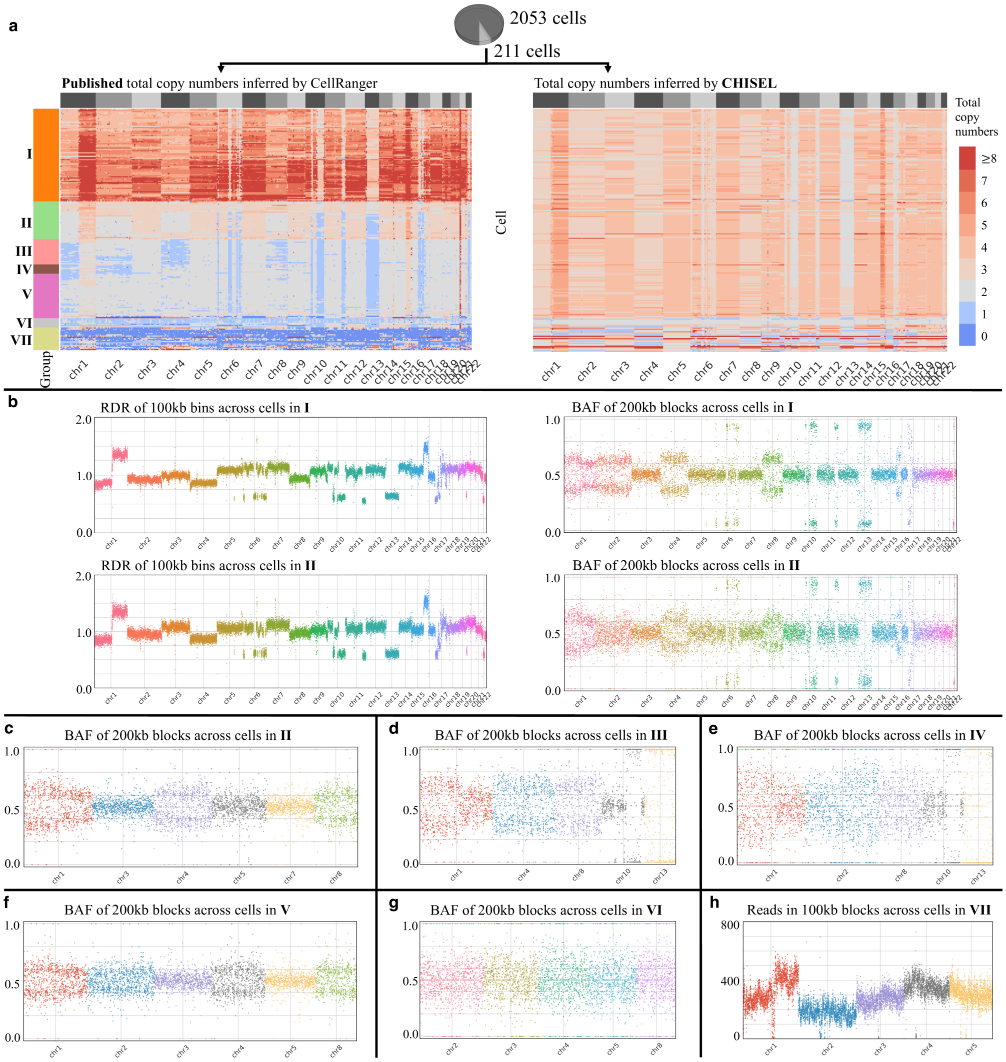
Supplementary Figures



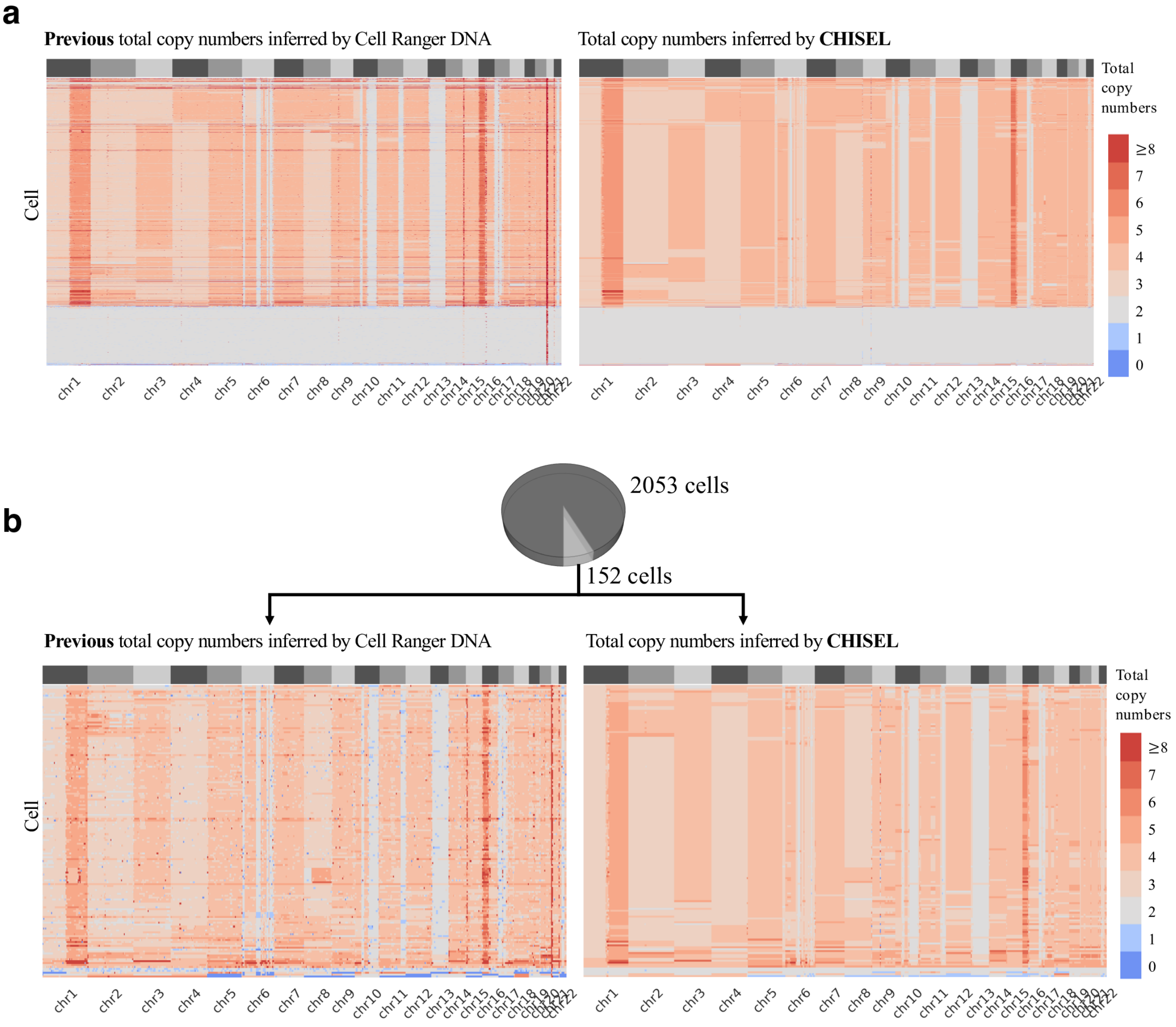
Supplementary Fig. 1: Summary of CHISEL's results across 10 single-cell datasets from 2 breast cancer patients.



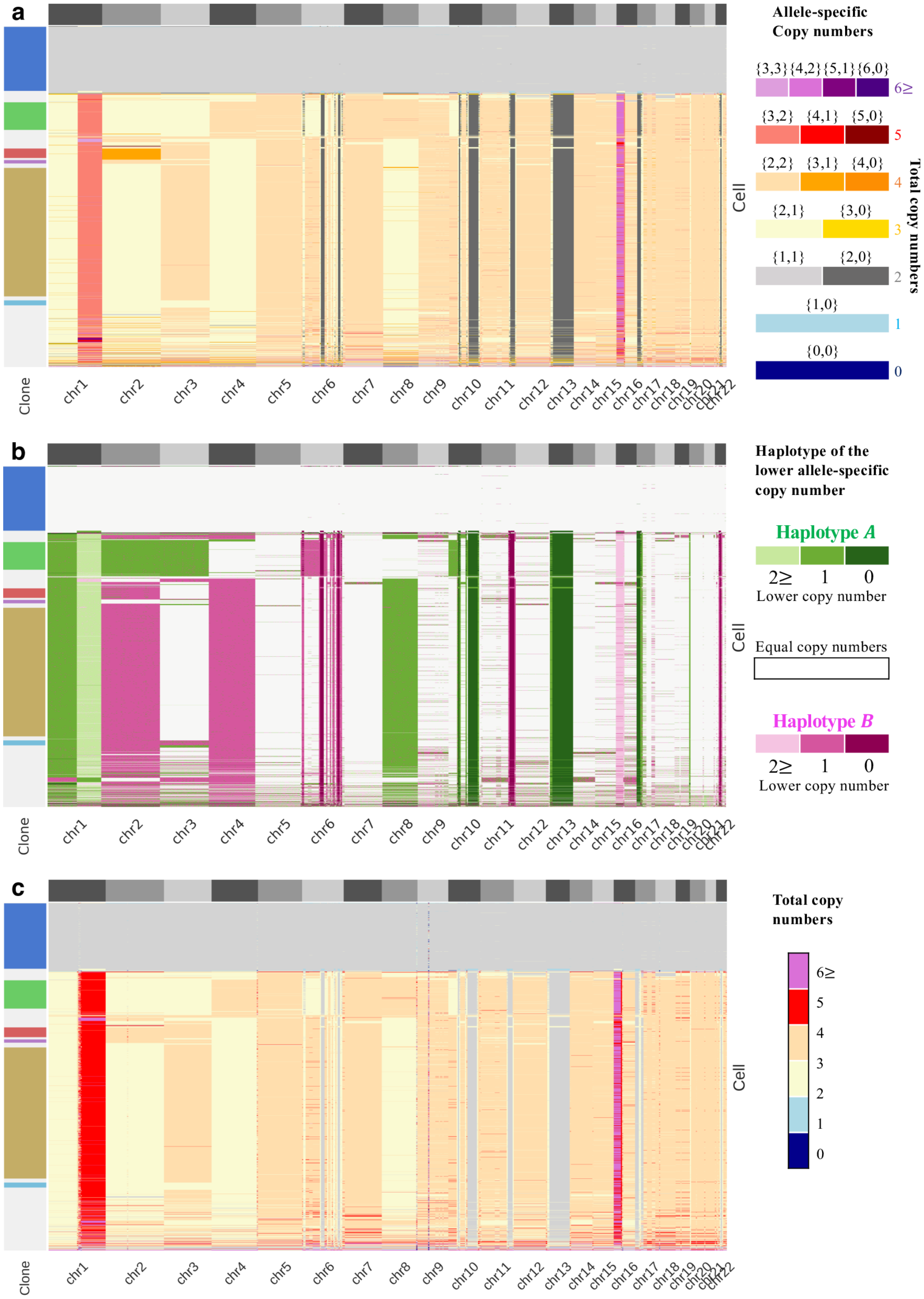
Supplementary Fig. 2: CHISEL's results are consistent with previous analysis for $\approx$90% cells. a, In every section of patient S0, CHISEL and Cell Ranger DNA infer approximately the same tumor purity, which is computed as the fraction of diploid cells. More specifically, a cell is defined as diploid when the allele-specific copy numbers are equal to $\{1, 1\}$ in at most 5% of the genome. **b**, In every cell, the difference in tumor ploidy is computed as the relative difference between the ploidy inferred by CHISEL and Cell Ranger DNA. **c**, In every cell, the difference in copy numbers is computed as the fraction of the genome with different total copy numbers inferred by CHISEL and Cell Ranger DNA.



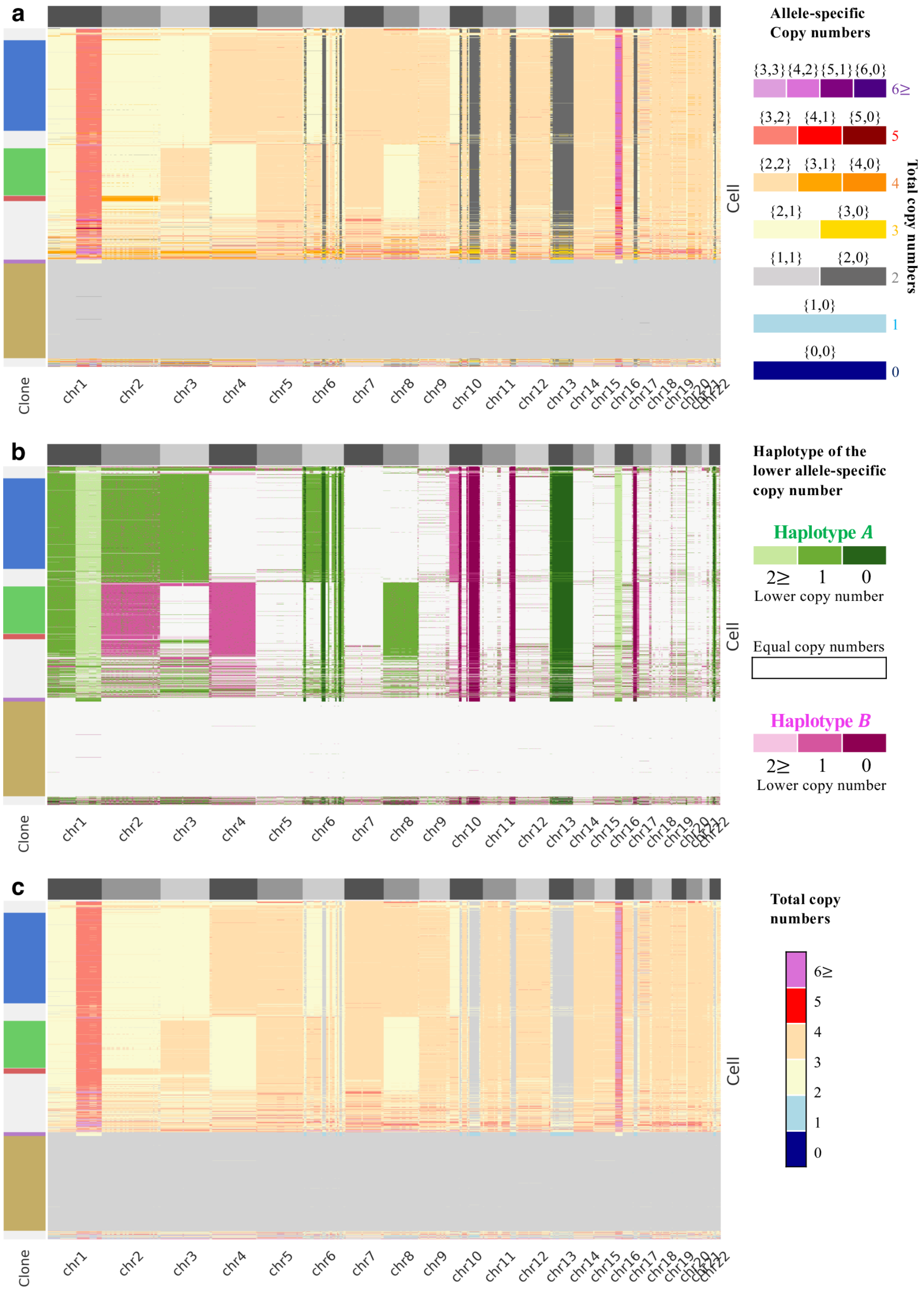
Supplementary Fig. 3: CHISEL infers more accurate copy numbers than previous analysis for ≈10% cells. **a**, The total copy numbers reported in previous total copy-number analysis (left) and CHISEL (right) are reported for the 211/2053 cells in section E of patient S0 with high difference in the inferred ploidy (⩾0.2). Previous analysis infers very different genome ploidy and total copy numbers for these 211 cells, while CHISEL infers copy numbers similar to those of the other cells in section E. Cells are thus separated into 7 groups according to the different genome ploidy inferred in previous analysis. RDRs and BAFs are computed in either 100-200kb bins across all chromosomes when pooling the sequencing reads from all the cells in each group. **b**, The RDRs and BAFs computed across all chromosomes in groups I and II are nearly identical and, hence, do not support the different tumor ploidy inferred in previous analysis. **c**, BAFs of group II indicate allelic imbalance in chromosomes 1, 4, and 8, and allelic balance in chromosomes 3, 5, and 7. In contrast, previous analysis suggests allelic balance for chromosomes 1, 4, and 8 with total copy number 2 (i.e. allele-specific copy numbers are $\{1, 1\}$), and allelic imbalance for chromosomes 3, 5, and 7 with total copy numbers 3 (i.e. with no potential balanced configurations of allele-specific copy numbers). **d**, BAFs of group III indicate the presence of both the alleles in chromosomes 1, 4, and 8 (i.e. BAF≠0 and BAF≠1), differently than chromosomes 10 and 13. In contrast, previous analysis report a total copy number 1 for chromosomes 1, 4, and 8, which correspond to the loss of one allele. **e**, BAFs of group IV indicate the presence of both alleles in chromosomes 1, 2, and 8 (i.e. BAF≠0 and BAF≠1), differently than chromosomes 10 and 13. In contrast, previous analysis reports a total copy number 1 for chromosomes 1, 2 and 8, corresponding to the loss of one allele. **f**, BAFs of group V indicate allelic imbalance in chromosomes 1, 2, 4, and 8. In contrast, previous analysis reports a total copy number 2 for these chromosomes, inconsistent with any imbalanced configuration of allele-specific copy numbers. **g**, BAFs of group VI indicate the presence of both alleles in chromosomes 2, 3, 4, 5 and 8 (i.e. BAF≠0 and BAF≠1). In contrast, previous analysis reports a total copy number 1 for these chromosomes, which correspond to the loss of one allele. **h**, Previous analysis reports a total copy number 0 for chromosomes 1, 2, 4, and 5, but a large number of reads is observed in these chromosomes in group VII.



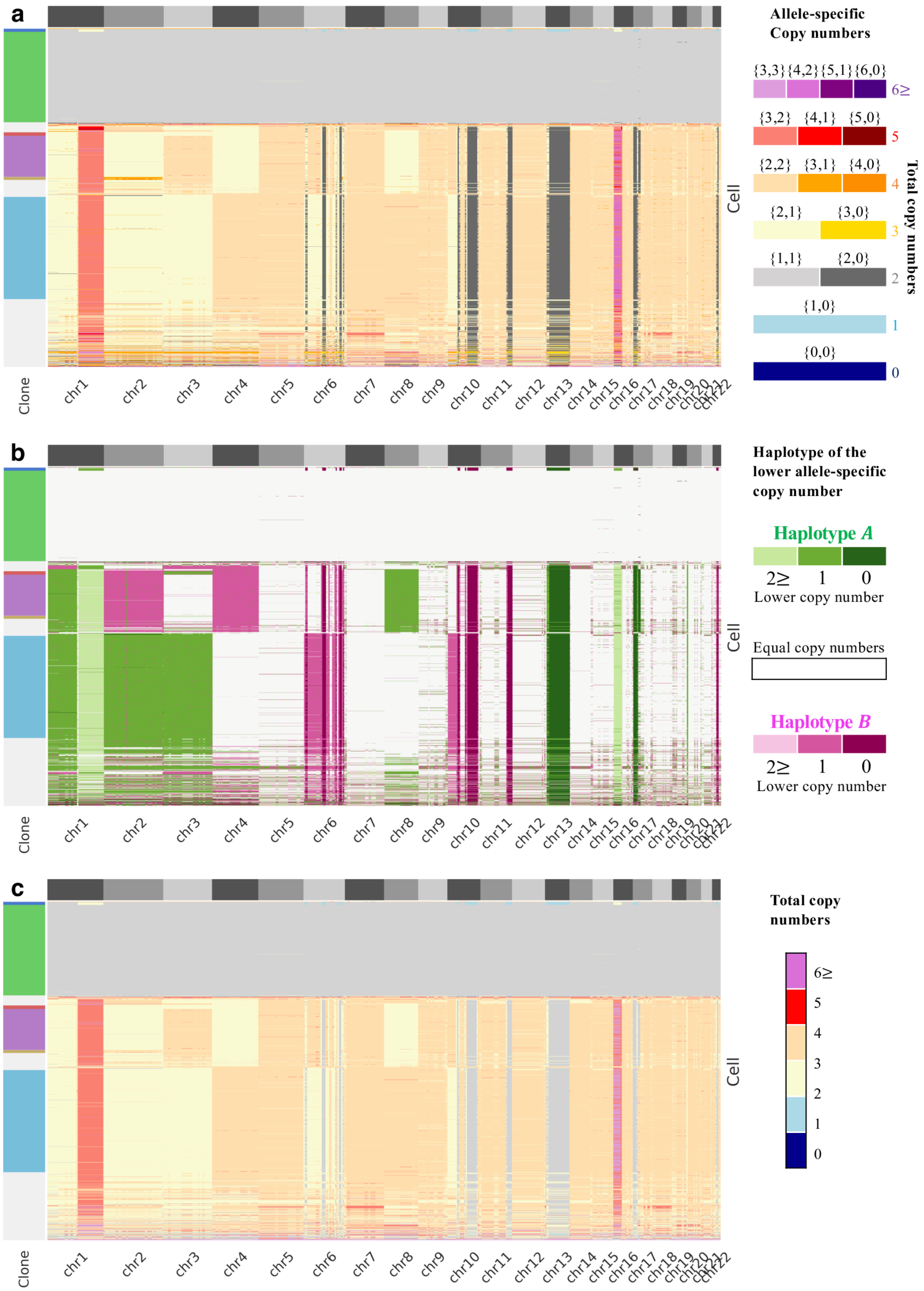
Supplementary Fig. 4: CHISEL reduces the number of outlying CNAs inferred by previous analysis. a, The total copy numbers reported in previous total copy-number analysis (left) and CHISEL (right) are reported for all the 2053 cells (rows in the same order) in section E of patient S0. The total copy numbers inferred in previous analysis are characterized by a high number of CNAs in small genomic regions and isolated in at most few cells. These events are outlying and likely noisy because they are shared by at most few cells. In contrast, only a limited number of outlying copy-number changes are present in the total copy numbers inferred by CHISEL. **b**, The outlying and noisy CNAs that are present in only few cells are especially clear in the 152 cells with the highest copy number difference between CHISEL and Cell Ranger DNA.



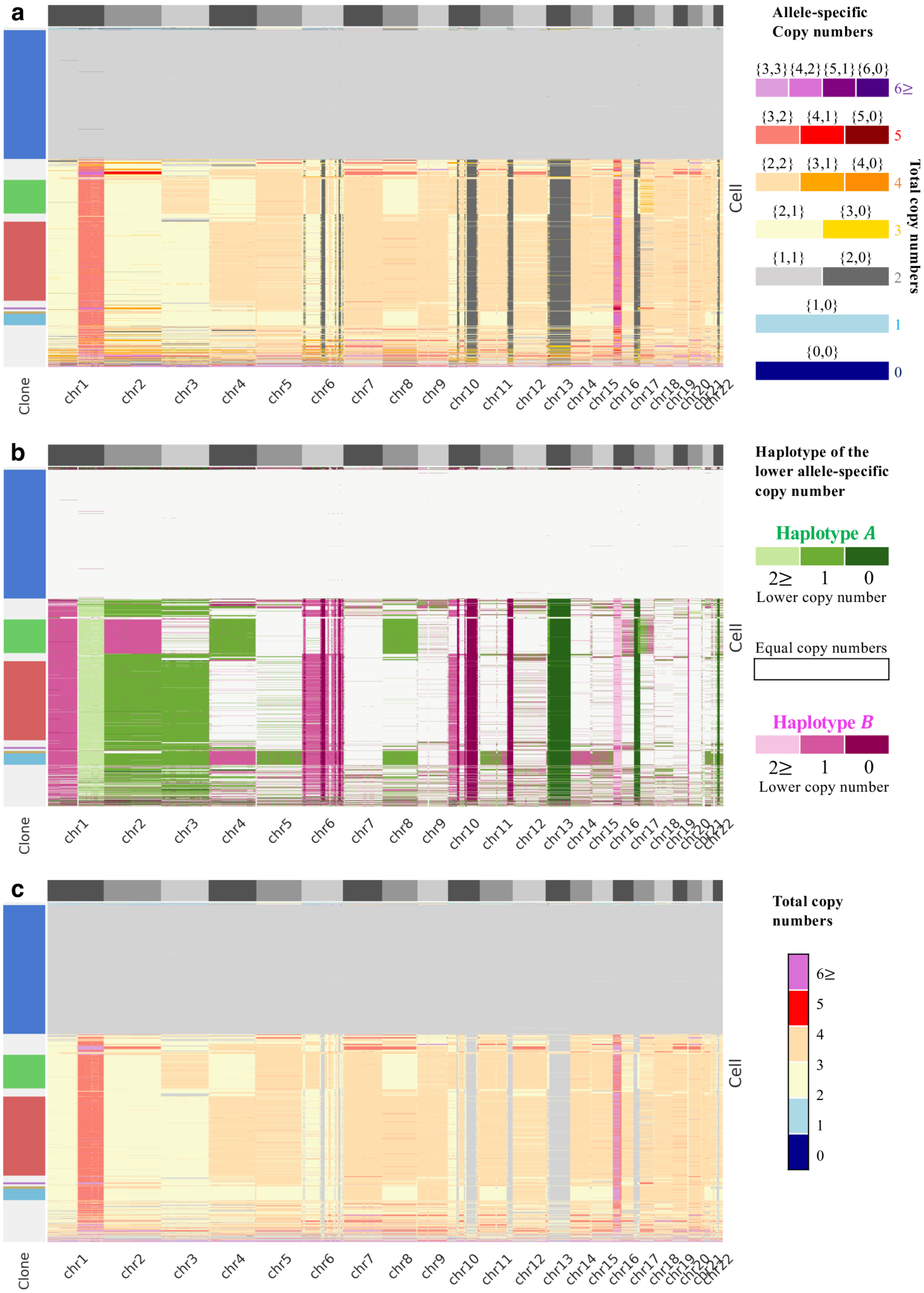
Supplementary Fig. 5: CHISEL's allele and haplotype- specific copy numbers for 2075 cells in section E of patient S0. The allele-specific, haplotype-specific, and total copy numbers are reported in **a**, **b**, and **c**, respectively, across all autosomes (grey rectangles in the first row). CHISEL groups 1448 cells into 6 clones (colors in the left-side bar) and classifies the remaining cells as noisy (grey in the left-side bar). The cells are equally ordered in each grid by their copy-number distance; as such, noisy cells are placed next to the closest clone. The few noisy bins located right in the centromeres of some chromosomes have been excluded as they exhibited noisy copy numbers across all cells.



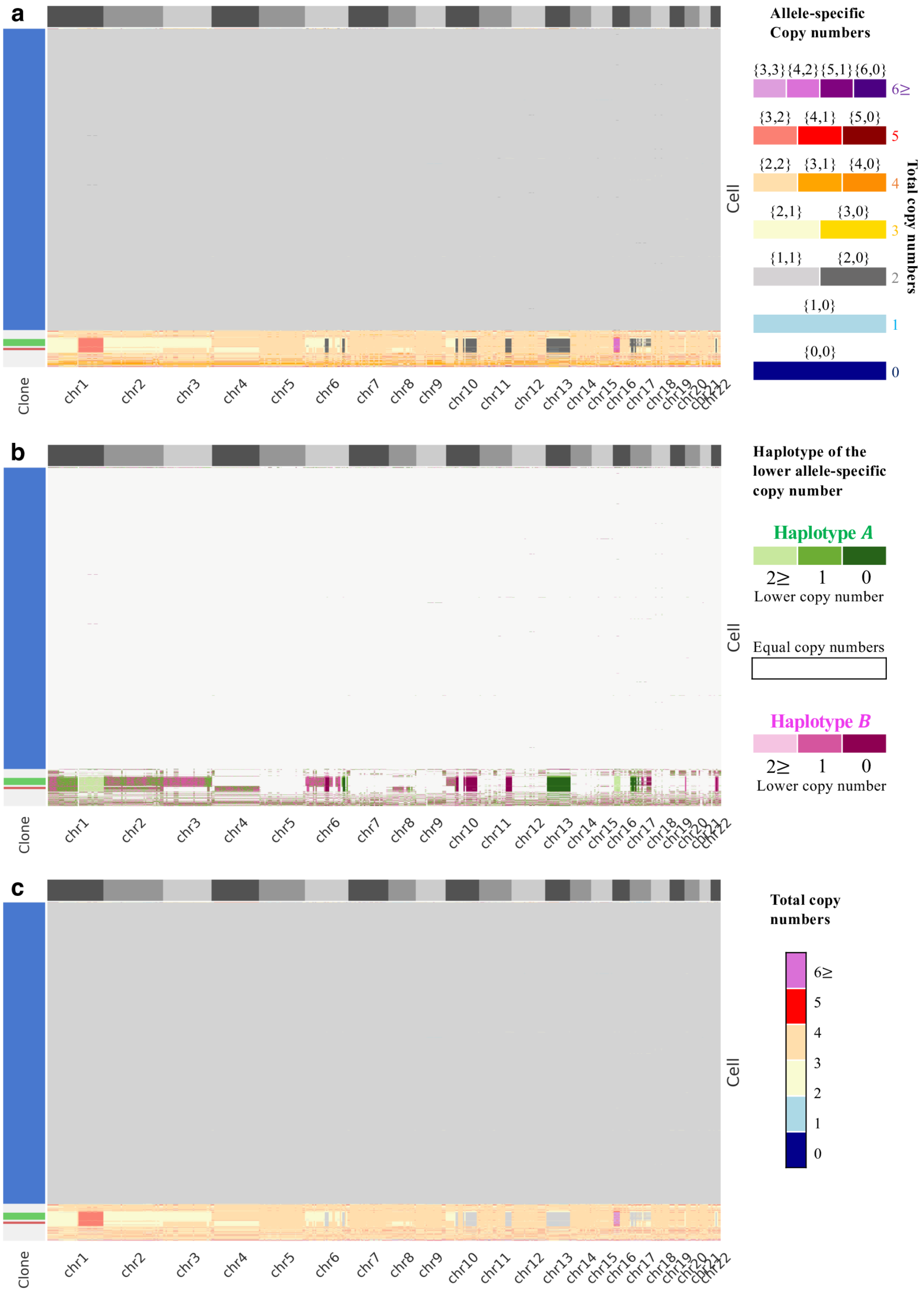
Supplementary Fig. 6: CHISEL's allele- and haplotype-specific copy numbers for 1943 cells in section D of patient S0. The allele-specific, haplotype-specific, and total copy numbers are reported in **a**, **b**, and **c**, respectively, across all autosomes (grey rectangles in the first row). CHISEL groups 1385 cells into 5 clones (colors in the left-side bar) and classifies the remaining cells as noisy (grey in the left-side bar). The cells are equally ordered in each grid by their copy-number distance; as such, noisy cells are placed next to the closest clone. The few noisy bins located right in the centromeres of some chromosomes have been excluded as they exhibited noisy copy numbers across all cells.



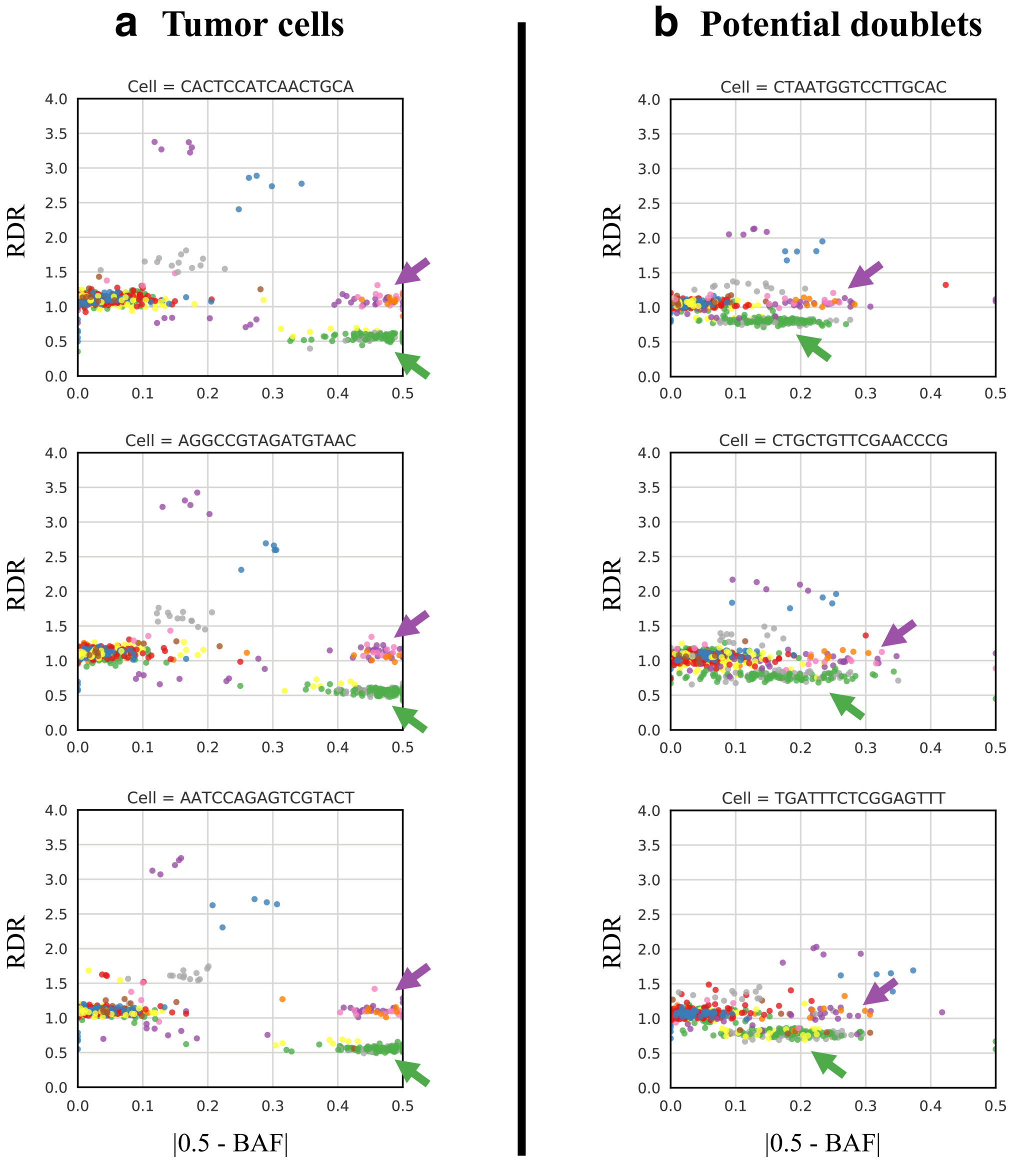
Supplementary Fig. 7: CHISEL's allele- and haplotype-specific copy numbers for 1754 cells in section C of patient S0. The allele-specific, haplotype-specific, and total copy numbers are reported in **a**, **b**, and **c**, respectively, across all autosomes (grey rectangles in the first row). CHISEL groups 1259 cells into 6 clones (colors in the left-side bar) and classifies the remaining cells as noisy (grey in the left-side bar). The cells are equally ordered in each grid by their copy-number distance; as such, noisy cells are placed next to the closest clone. The few noisy bins located right in the centromeres of some chromosomes have been excluded as they exhibited noisy copy numbers across all cells.



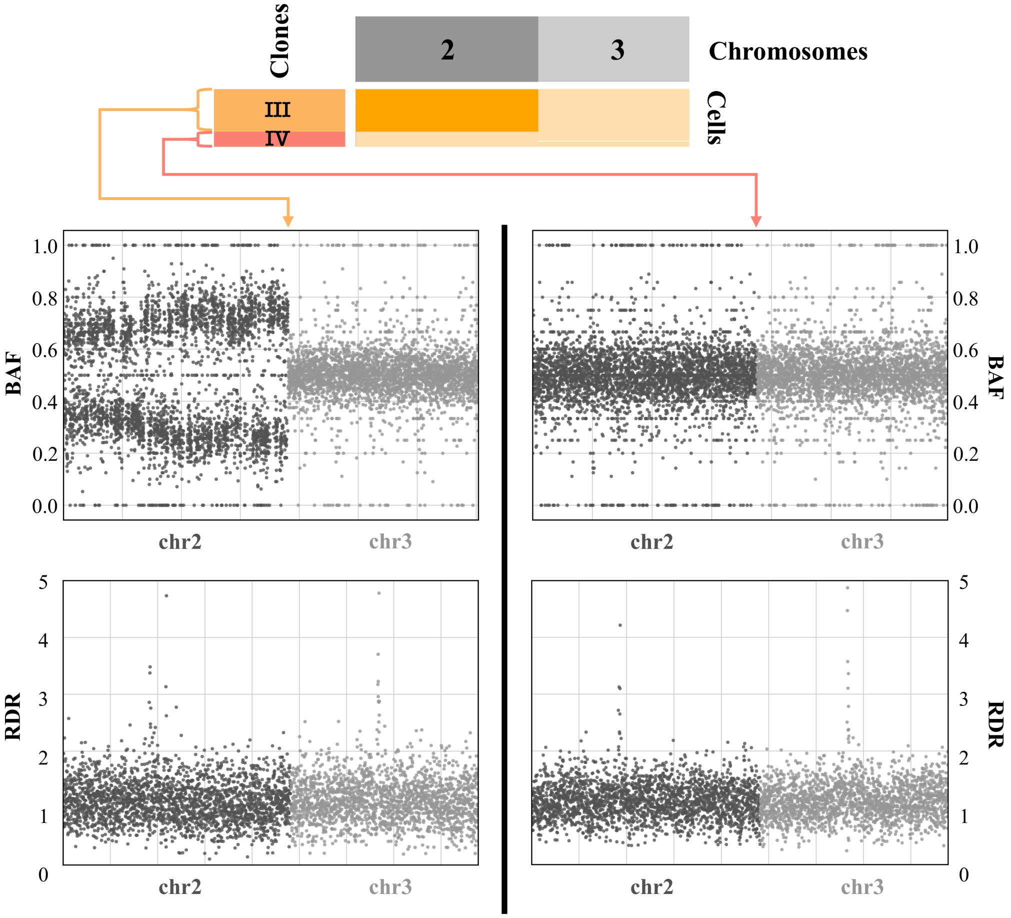
Supplementary Fig. 8: CHISEL's allele- and haplotype-specific copy numbers for 2239 cells in section B of patient S0. The allele-specific, haplotype-specific, and total copy numbers are reported in **a**, **b**, and **c**, respectively, across all autosomes (grey rectangles in the first row). CHISEL groups 1694 cells into 6 clones (colors in the left-side bar) and classifies the remaining cells as noisy (grey in the left-side bar). The cells are equally ordered in each grid by their copy-number distance; as such, noisy cells are placed next to the closest clone. The few noisy bins located right in the centromeres of some chromosomes have been excluded as they exhibited noisy copy numbers across all cells.



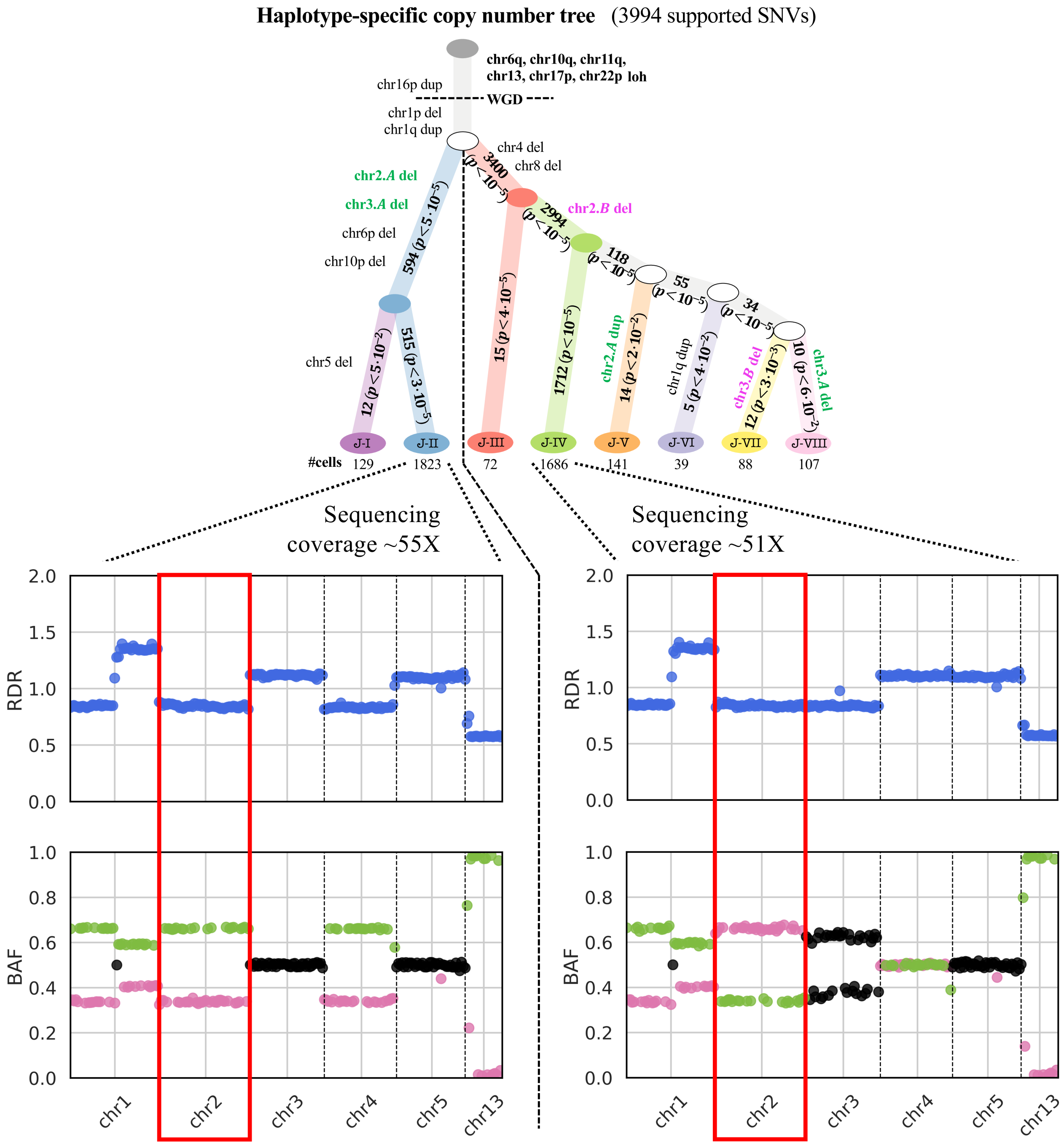
Supplementary Fig. 9: CHISEL's allele- and haplotype-specific copy numbers for 2191 cells in section A of patient S0. The allele-specific, haplotype-specific, and total copy numbers are reported in **a**, **b**, and **c**, respectively, across all autosomes (grey rectangles in the first row). CHISEL groups 2008 cells into 3 clones (colors in the left-side bar) and classifies the remaining cells as noisy (grey in the left-side bar). The cells are equally ordered in each grid by their copy-number distance; as such, noisy cells are placed next to the closest clone. The few noisy bins located right in the centromeres of some chromosomes have been excluded as they exhibited noisy copy numbers across all cells.



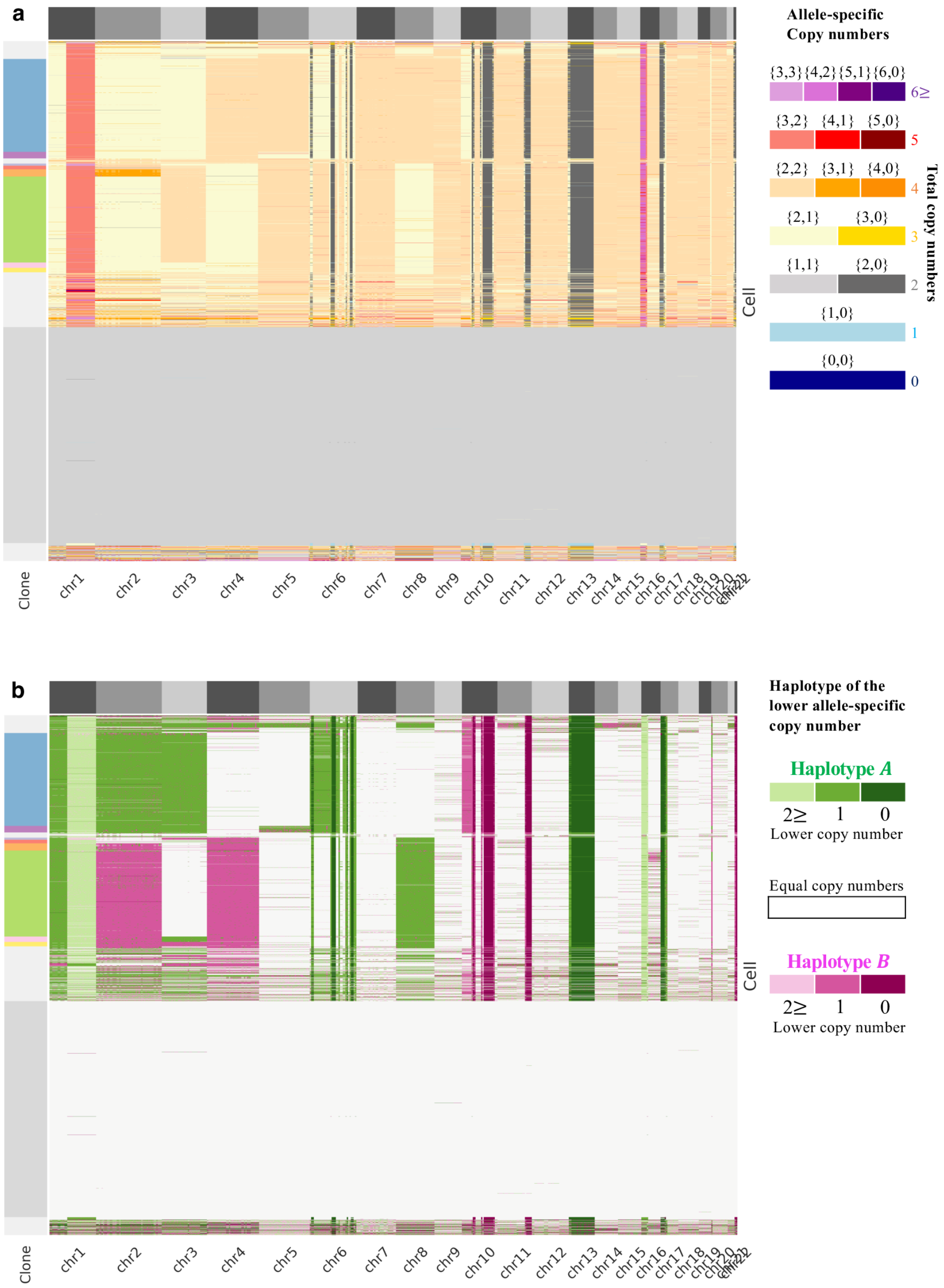
Supplementary Fig. 10: Evidence of doublets in section A1 of patient S1. a, RDRs and BAFs for all 5Mb genomic bins are reported for 3 tumor cells in section A1 of patient S1. Two large clusters of genomic bins (green and purple arrows) have a value of BAF consistent with a LOH event (|0.5 - BAF|≈0.5), i.e. loss of one of the two alleles. **b**, RDRs and BAFs for all 5Mb genomic bins are reported for 3 tumor cells in section A1 of patient S1. The BAFs of the same clusters of bins that are consistent with LOH in **a** have a clear shift away from |0.5 - BAF|≈0.5 towards values that indicate the presence of both the alleles, i.e. |0.5 - BAF|≈0.2 and |0.5 - BAF|≈0.25. All the other clusters also exhibit a similar shift in BAF. These shifts only observed in a limited number of cells are consistent with the occurrence of doublets where a tumor cell and a normal cell have received the same barcode and we thus observe a mixture of sequencing reads belonging to both cells.



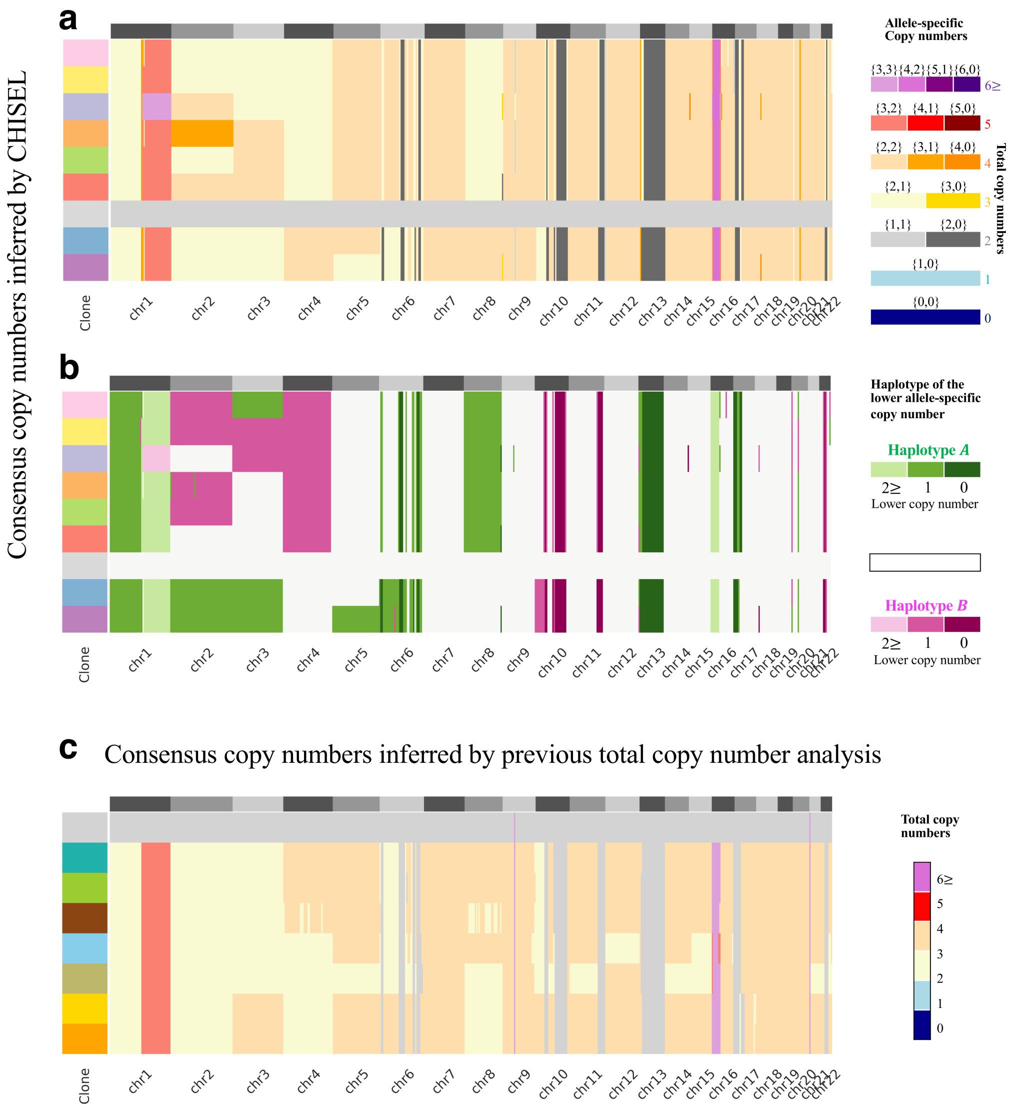
Supplementary Fig. 11: Allele-specific CNAs reveal two distinct clones in section E of patient S0. The clones `III` and `IV` identified by CHISEL in section E of patient S0 and comprising 58 and 20 cells, respectively, have the same total copy number 4 in chromosomes 2 and 3. However, while the two clones have the same allele-specific copy numbers $\{2, 2\}$ (light orange) in chromosome 3, they have different allele-specific copy numbers in chromosome 2 equal to $\{3, 1\}$ (dark orange) and $\{2, 2\}$ (light orange), respectively. RDRs (computed across 100kb genomic regions) and BAFs (computed across 50kb haplotype blocks) across chromosomes 2 and 3 are computed by pooling the sequencing reads from all cells in either clone `III` (left) or clone `IV` (right). The different allele-specific copy numbers of the two clones are well supported by RDRs and BAFs: while all the cells in both clones have similar RDRs across chromosomes 2 and 3 as well as similar BAFs≈0.5 in chromosome 3 consistent with allele-specific copy numbers $\{2, 2\}$, the cells in clone `IV` have BAF≈0.5 also in chromosome 2 but the cells in clone `III` exhibit a clear shift away from BAF=0.5 in chromosome 2 which is consistent with allele-specific copy numbers $\{3, 1\}$.



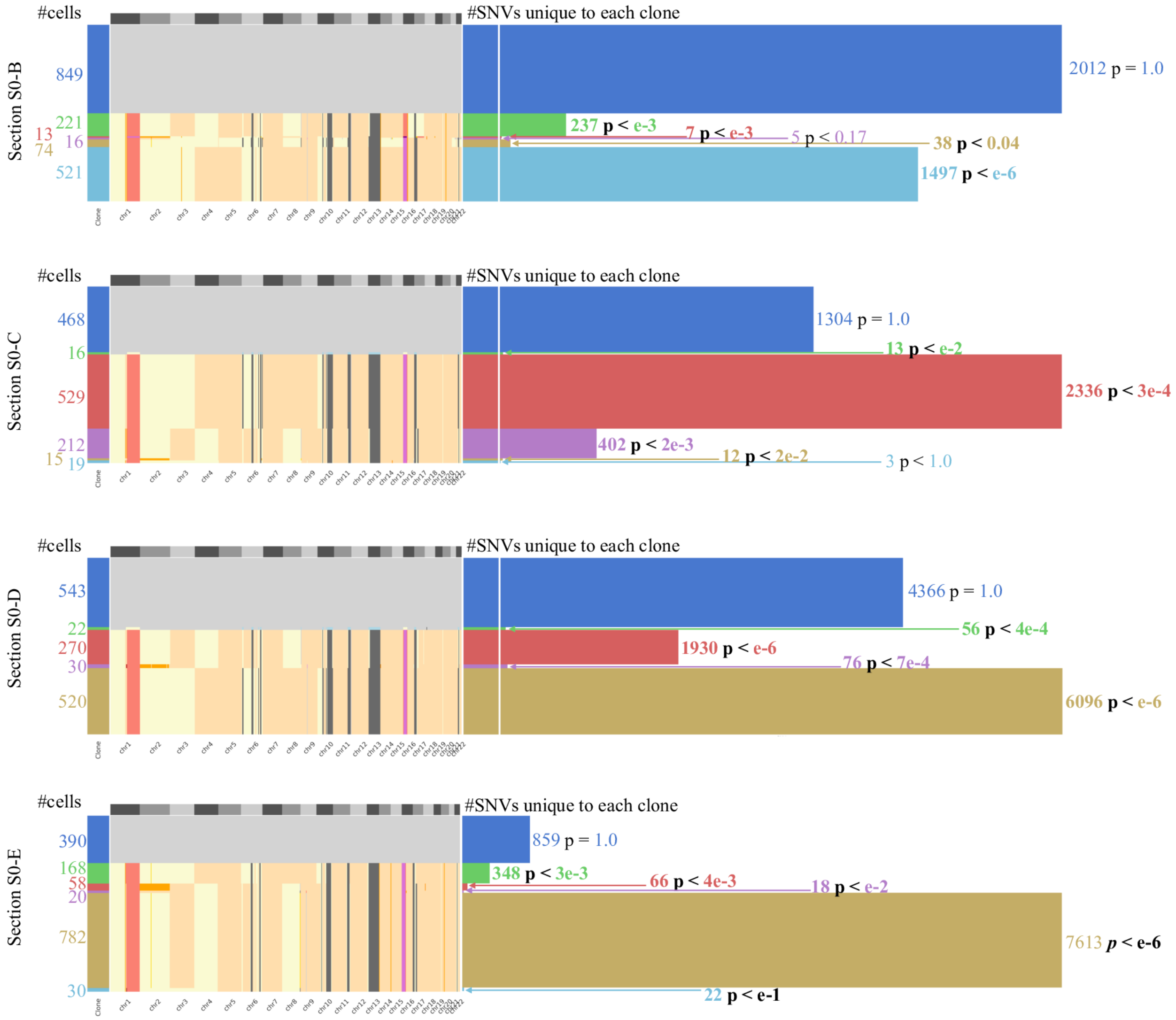
Supplementary Fig. 12: The RDRs and BAFs computed in pseudo-bulk samples support the mirrored-subclonal CNAs identified by CHISEL. (Top) The copy-number tree inferred by CHISEL contains 8 distinct tumor clones (`J-I`, . . . , `J-VIII`) with the same haplotype-specific copy numbers and varying number of cells (indicated as "#cells"). **(Bottom left)** The RDRs and BAFs in 5Mb genomic bins across five chromosomes from a pseudo-bulk sample formed by merging all the sequencing reads from the 1823 cells in clone `J-II` with a resulting sequencing coverage of $\approx 55\times$. The genomic regions are colored according to the values of BAF (green for BAF>0.5, black for BAF≈ 0.5, and pink for BAF<0.5). **(Bottom right)** The RDRs and BAFs in 5Mb genomic bins across five chromosomes from a pseudo-bulk samples formed by merging all the sequencing reads from the 1686 cells in clone `J-IV` with a resulting sequencing coverage of $\approx 51\times$. The genomic regions are colored with the same color as in **(Bottom left)**. Bins on chromosome 2 have the same values of RDR but swap the values of BAF between the two pseudo-bulk samples, supporting CHISEL's inference of a mirrored-subclonal CNA on chromosome 2 that distinguishes these two clones.



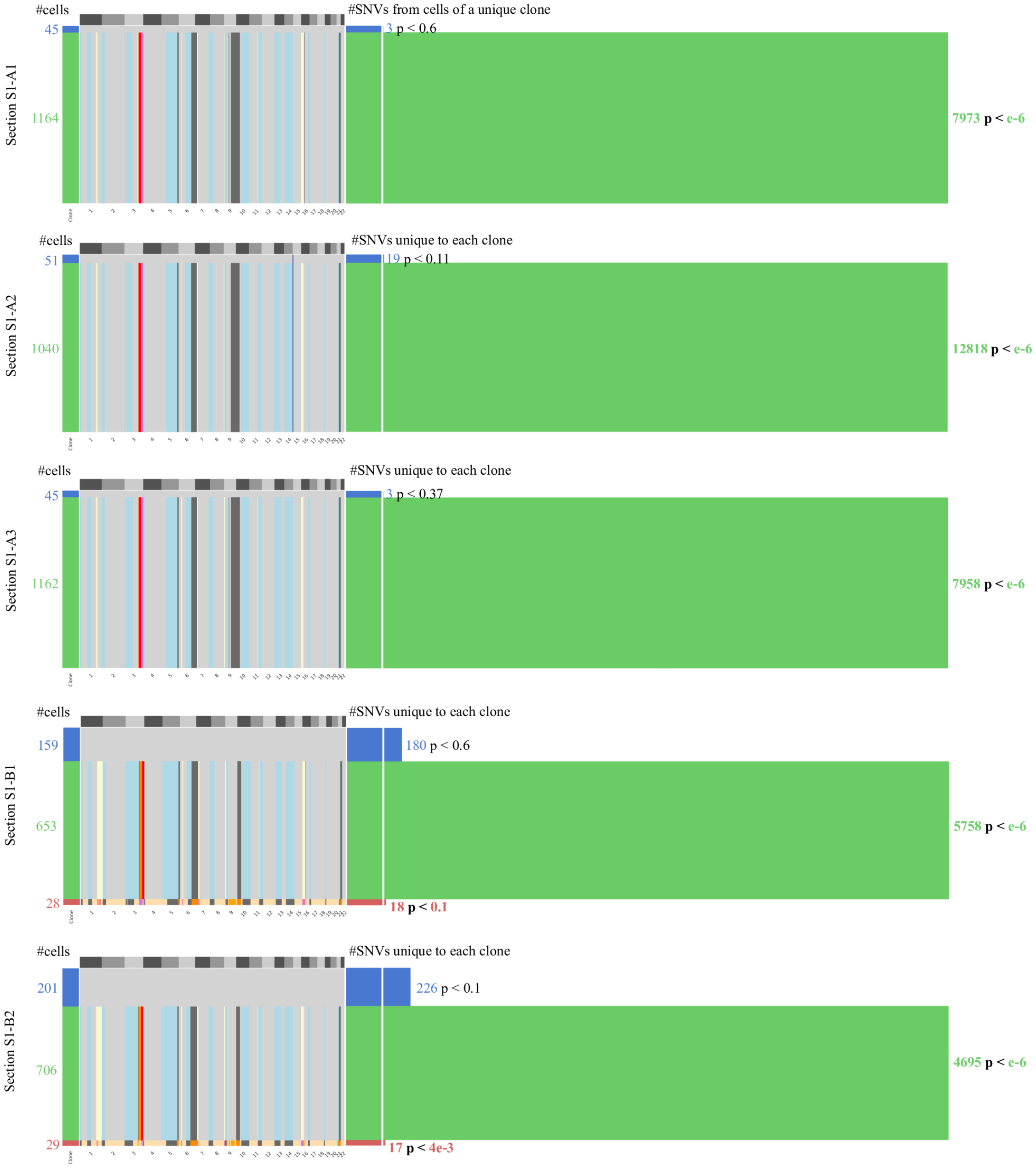
Supplementary Fig. 13: CHISEL infers allele- and haplotype-specific copy numbers for 10 202 **cells in patient S0.** The allele- and haplotype-specific copy numbers are reported in **a** and **b**, respectively, across all autosomes (grey rectangles in the first row). CHISEL groups 8 324 cells into 9 clones (colors in the left-side bar), including eight tumor clones and one diploid clone, and classifies the remaining cells as noisy (grey in the left-side bar). The cells are equally ordered in each grid by their copy-number distance; as such, noisy cells are placed next to the closest clone. The few noisy bins located right in the centromeres of some chromosomes have been excluded as they exhibited noisy copy numbers across all cells.



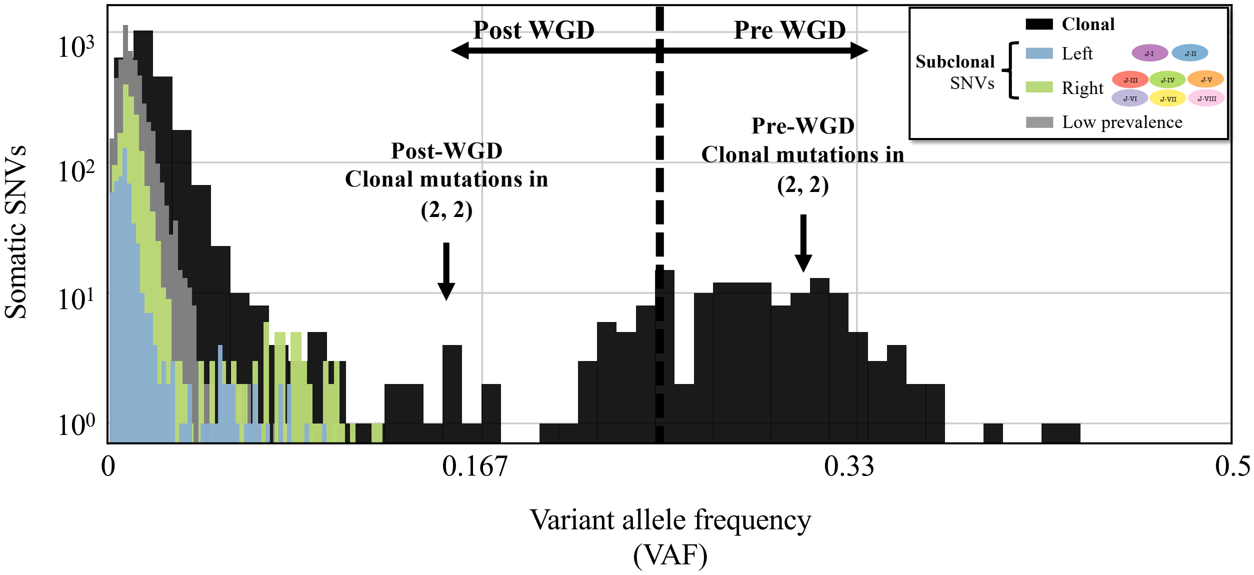
Supplementary Fig. 14: Consensus copy numbers for the clones inferred by CHISEL and previous total copy-number analysis across all cells of patient S0. **a**, Allele-specific copy numbers for the 9 clones identified by CHISEL, including the diploid clone. **b**, Haplotype-specific copy numbers for the 9 clones identified by CHISEL, including the diploid clone, are represented by specifying the haplotype where the lower allele-specific copy number is located. **c**, Total copy numbers for the 8 clones identified by previous total copy-number analysis, including the diploid clone.



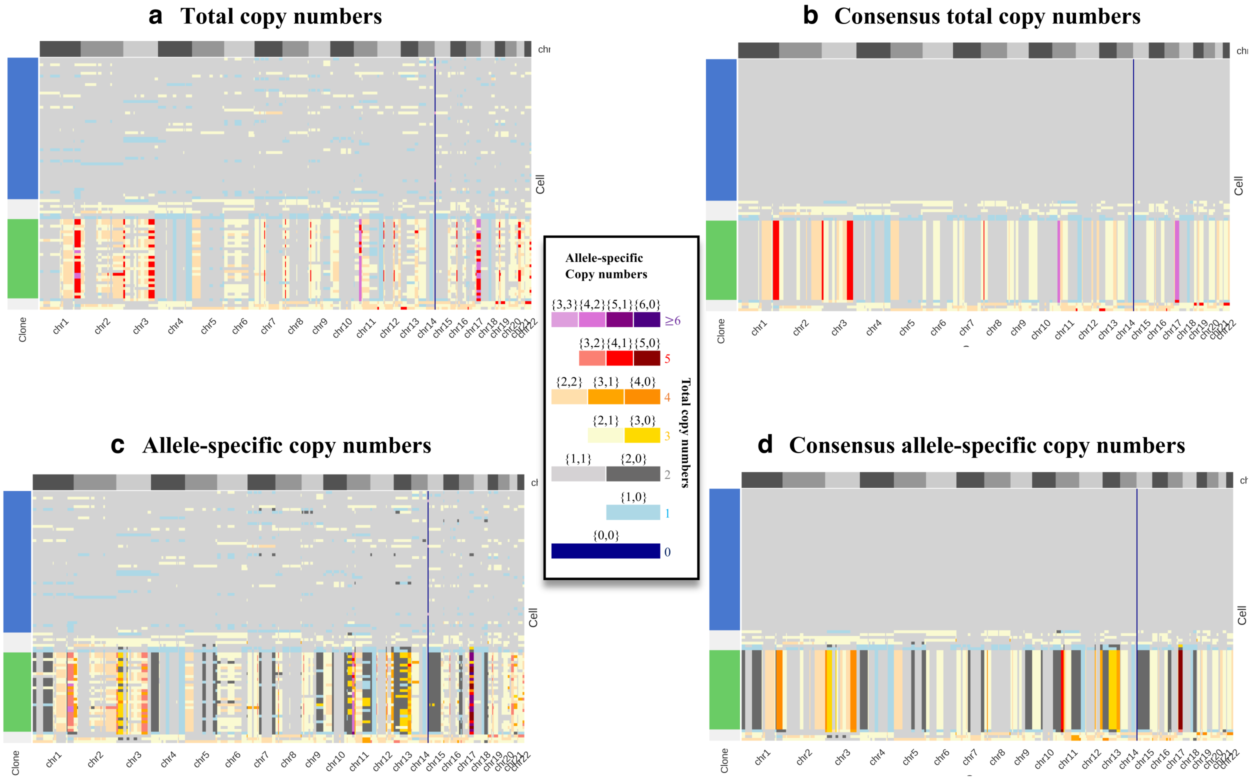
Supplementary Fig. 15: Somatic SNVs support the tumor clones inferred by CHISEL in patient S0. For each tumor section of patient S0 with high tumor purity, CHISEL infers the allele-specific copy numbers (left side) of distinct clones (rows) in 5Mb genomic bins across all autosomes (columns). The height of each row is proportional to the number of cells in the corresponding clone (reported on the left side). The number of unique SNVs to each clone are represented with bars whose height is proportional to the corresponding number of cells and whose width is proportional to the corresponding number of unique SNVs. Each bar is correspondingly labelled with the number of unique SNVs. To compare the number of unique SNVs to those expected by chance for every clone, p-values are computed using a permutation test with $n = 10^5$ randomly sampled subsets of cells of the same size as the considered clone and significant values ($p < 10^{-1}$) are in boldface. Nearly all tumor clones are significantly supported by their number of unique SNVs but the diploid clone; this is expected as the normal cells in the diploid clone generally has less somatic SNVs than tumor cells.



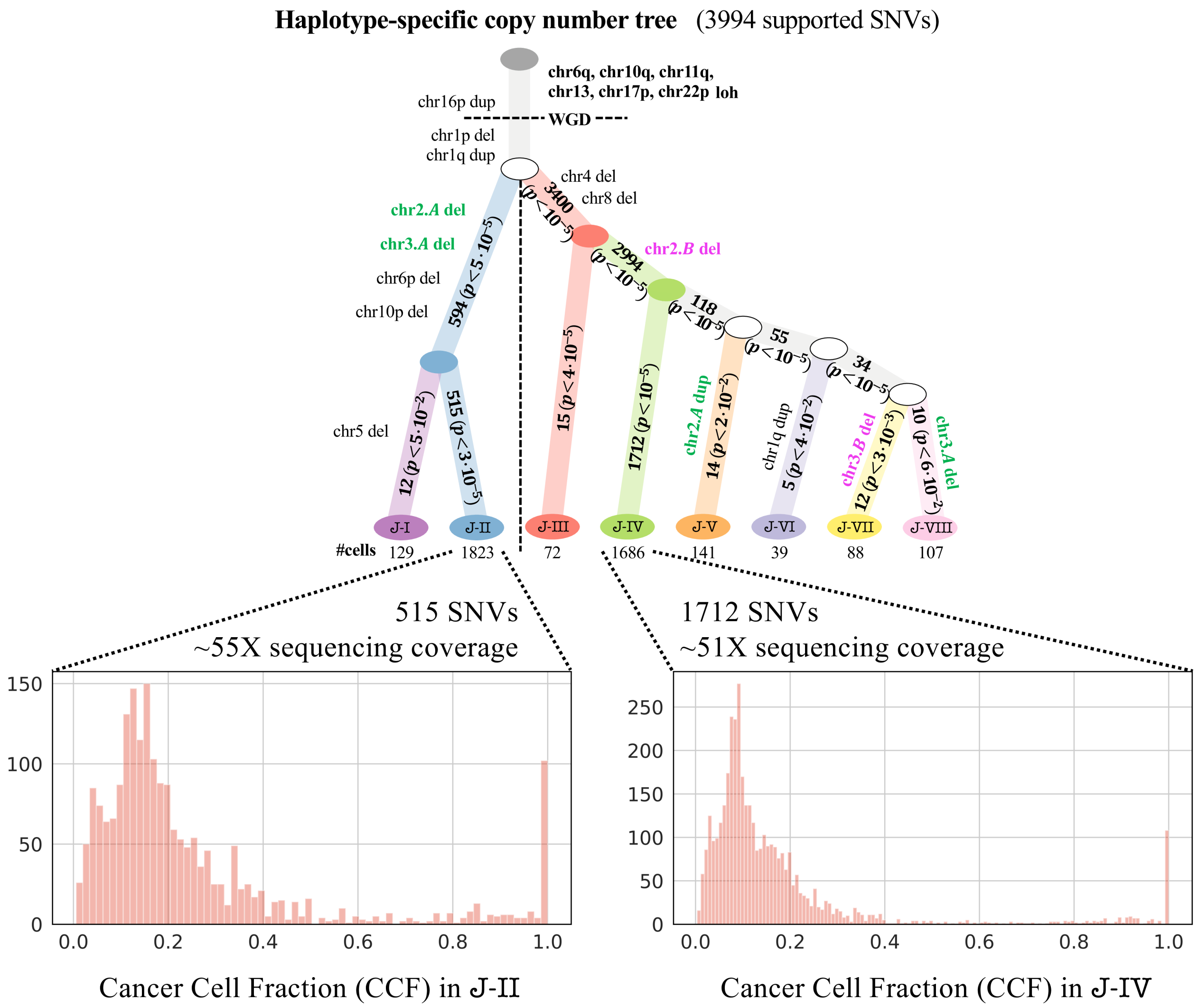
Supplementary Fig. 16: Somatic SNVs support the tumor clones inferred by CHISEL in patient S1. For each single-cell dataset of patient S1, CHISEL infers the allele-specific copy numbers (left side) of distinct clones (rows) in 5Mb bins across all autosomes (columns). The height of each row is proportional to the number of cells in the corresponding clone (reported on the left side). The number of unique SNVs to each clone are represented with bars whose height is proportional to the corresponding number of cells and whose width is proportional to the corresponding number of unique SNVs. Each bar is correspondingly labelled with the number of unique SNVs. To compare the number of unique SNVs to those expected by chance for every clone, p-values are computed using a permutation test with $n = 10^5$ randomly sampled subsets of cells of the same size as the considered clone and significant values ($p < 10^{-1}$) are in boldface. Nearly all tumor clones are significantly supported by their number of unique SNVs but the diploid clone; this is expected as the normal cells in the diploid clone generally has less somatic SNVs than tumor cells. Datasets S1-A1, S1-A2, and S1-A3 as well as datasets S1-B1 and S1-B2 corresponds to technical replicates of the same tumor sample, and their similarities support the corresponding results.



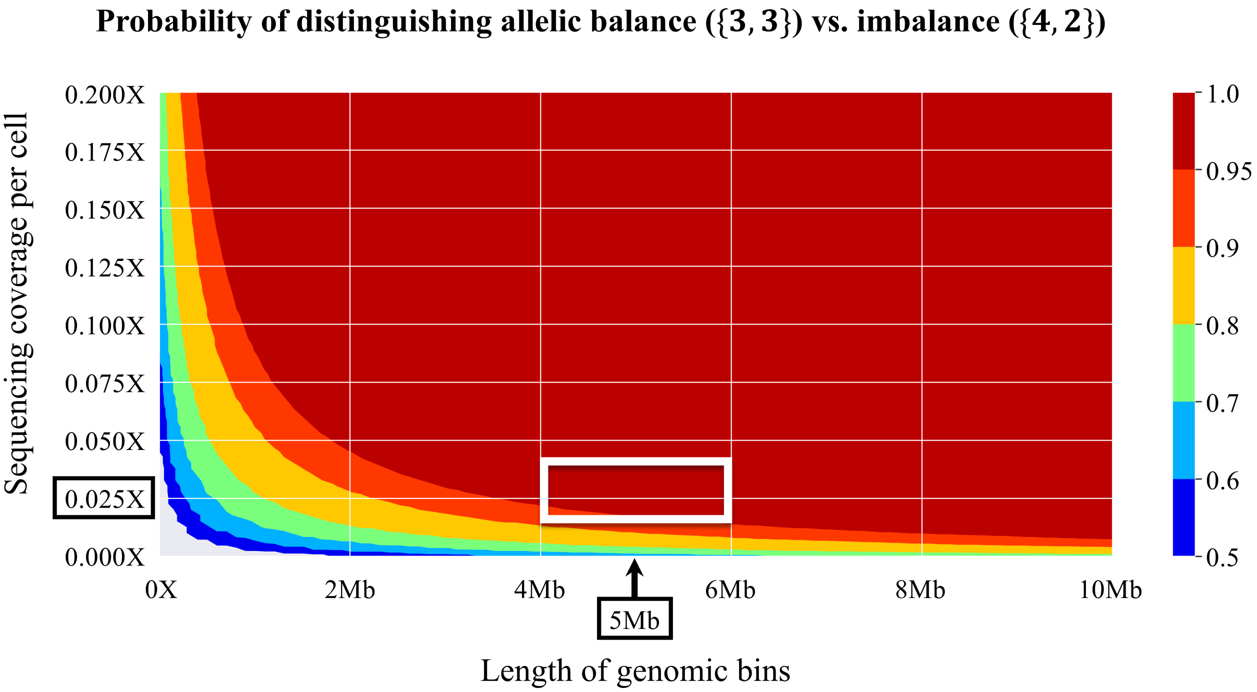
Supplementary Fig. 17: Observed VAF of somatic SNVs is consistent with CHISEL tree. The variant-allele frequency (VAF) is computed across all cells of patient S0 as the fraction of variant sequencing reads covering each of the 10 551 SNVs present in the tumor clones identified by CHISEL. Moreover, the SNVs are partitioned by CHISEL into 4 classes: clonal SNVs (black) that are present in all tumor clones, subclonal SNVs that are unique to either the left (green) or right (blue) branch of the CHISEL tree, and low prevalence SNVs which are identified as potential false positives. Since tumor purity is $\approx$50% across all cells of patient S0, the expected VAF for clonal SNVs not affected by further CNAs but the WGD (i.e. with allele-specific copy numbers $\{2, 2\}$) is $\approx$0.33 when occurred before WGD (Pre WGD) because 2 copies are mutated over 4 in $\approx$50% of tumor cells and the remaining $\approx$50% are normal cells with no copies mutated over 2. The expected VAF for clonal SNVs not affected by further CNAs but the WGD is $\approx$0.167 when occurred after WGD (Post WGD) because 1 copy is mutated over 4 in $\approx$50% of tumor cells and the remaining $\approx$50% are normal cells with no copies mutated over 2. In accordance with CHISEL's classification of SNVs, only clonal SNVs have expected values of VAF for mutations occurred before WGD. In addition, all the subclonal SNVs have values of VAF expected after WGD or lower, in accordance with the two branches of the CHISEL tree that separate after the occurrence of the clonal WGD. Note that some clonal SNVs have values of VAF $\approx$0.2 due to CNAs; in fact, both SNVs occurring after WGD and before WGD have expected VAF of $\approx$0.2 in genomic regions affected by a deletion after WGD (i.e. with allele-specific copy numbers $\{2, 1\}$) and deleting one mutated copy in case the SNV occurred before WGD, because in both cases 1 copy is mutated over 3 in the 50% of tumor cells and the remaining $\approx$50% are normal cells with no copies mutated over 2. Also, the low prevalence SNVs have very low values of VAF, underscoring their classification.



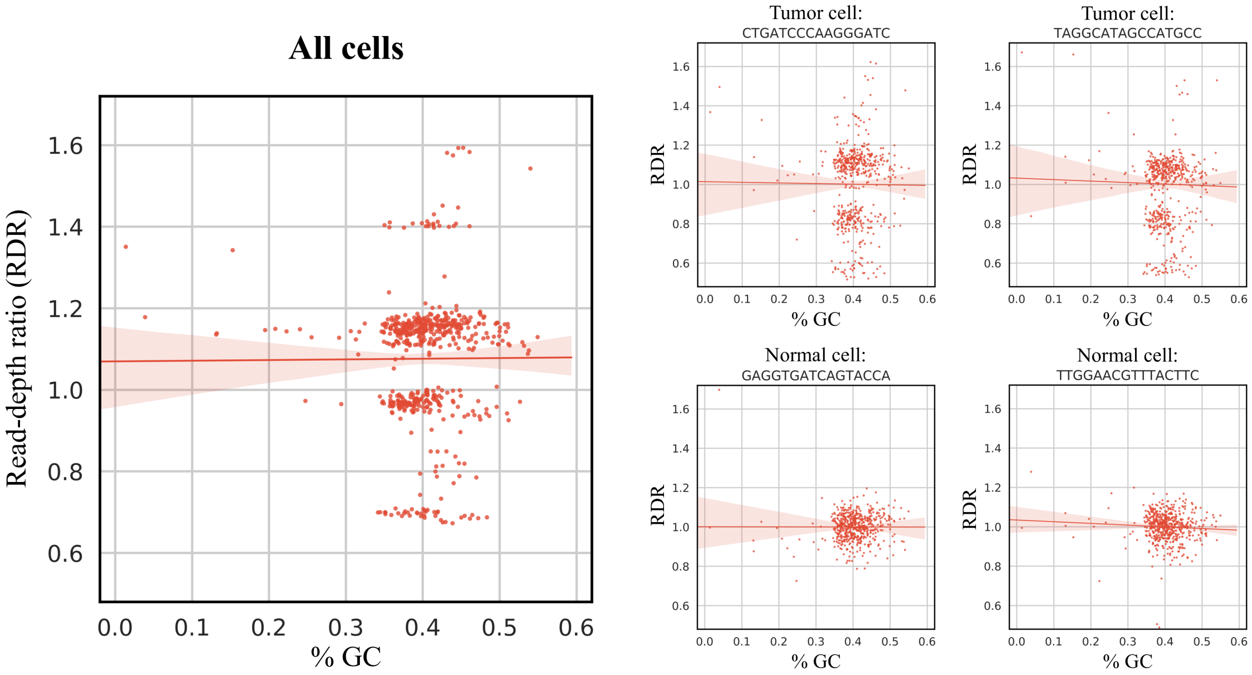
Supplementary Fig. 18: CHISEL identifies allele-specific CNAs in DOP-PCR sequencing data of 89 breast cancer cells. **a**, Total copy numbers inferred by CHISEL in DOP-PCR sequencing data of 89 single cells from breast cancer patient P2 in Kim et al. (2018)[1]. CHISEL identified two distinct clones, in agreement with the published analysis. **b**, Consensus total copy numbers inferred by CHISEL using all the cells that were identified in the same clone. **c**, Allele-specific copy numbers inferred by CHISEL from the sequencing data of the same 89 cells. **d**, Consensus allele-specific copy numbers inferred by CHISEL using all the cells that were identified in the same clone. The tumor clone (green) is characterized by extensive allele-specific CNAs, including copy-neutral LOHs (dark grey genomic regions) and others (e.g. yellow regions on chromosomes 3 and 13 with allele-specific copy numbers $\{3, 0\}$).



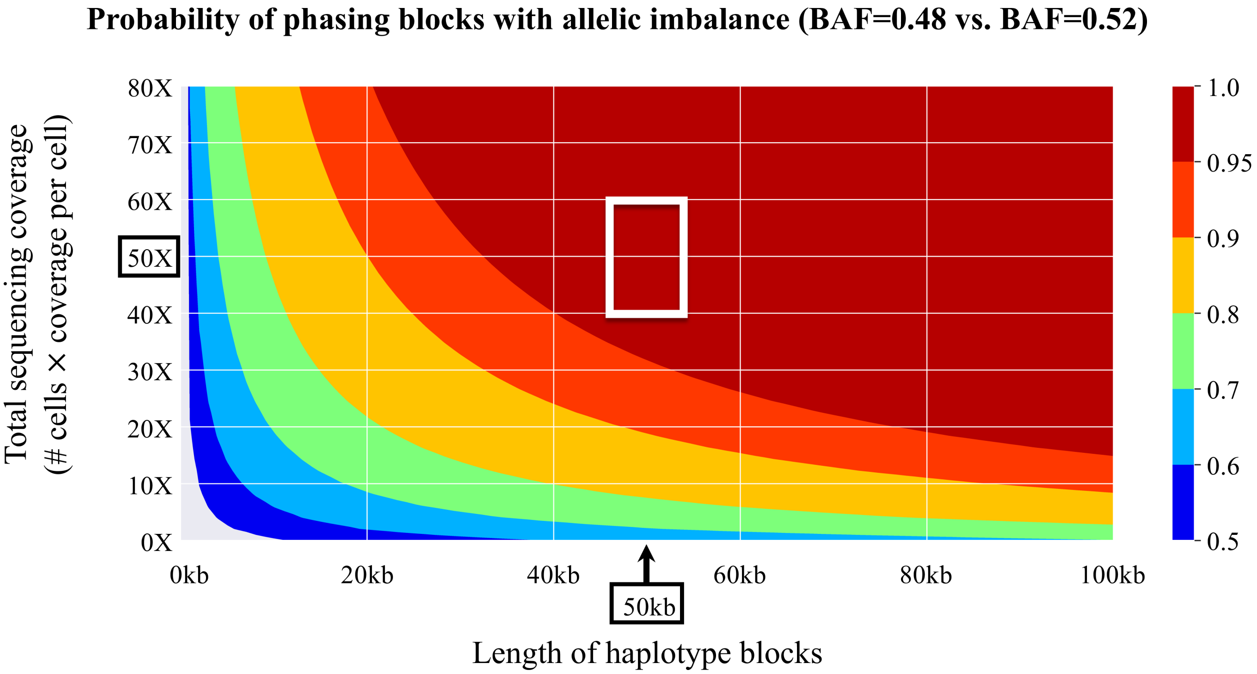
Supplementary Fig. 19: CCF analysis of somatic SNVs in two high-coverage pseudo bulk-tumor samples reveal additional intra-tumor heterogeneity. **(Top)** The copy-number tree inferred by CHISEL contains 8 distinct tumor clones (`J-I`, …, `J-VIII`) with the same haplotype-specific copy numbers and varying number of cells (indicated as "#cells"). **(Bottom left)** The cancer-cell fractions (CCFs) of 515 somatic single-nucleotide mutations (SNVs) unique to clone `J-II` are computed using an existing method[2] from a pseudo-bulk sample formed by pooling all reads from single cells in clone `J-II`, resulting in a sequencing coverage $\approx 55\times$. **(Bottom right)** The CCFs of 1712 SNVs unique to clone `J-IV` are computed from a pseudo bulk-tumor sample formed by pooling all reads from single cells in clone `J-IV`, resulting in a sequencing coverage $\approx 51\times$. **(Bottom)** Many SNVs have CCF$=1$ indicating presence in all cells of either clone `J-II` or `J-IV`, but there are also a large number of SNVs with CCF<1 indicating groups of cells with distinct complements of SNVs.



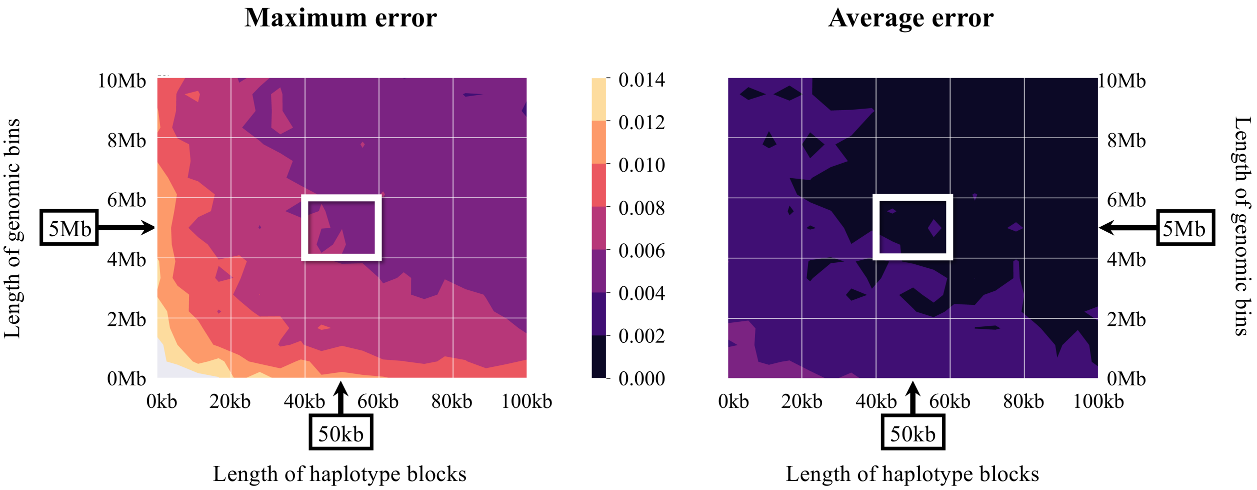
Supplementary Fig. 20: Probability of identifying allelic imbalance in a single cell as a function of bin length and sequencing coverage per cell. The probability of distinguishing bins with different allele-specific copy-numbers of $\{3, 3\}$ and $\{4, 2\}$ in single cells for various bin lengths and sequencing coverages per cell is computed analytically. 5Mb genomic bins give a probability >95% (white rectangle) when the sequencing coverage per cell is 0.025×.



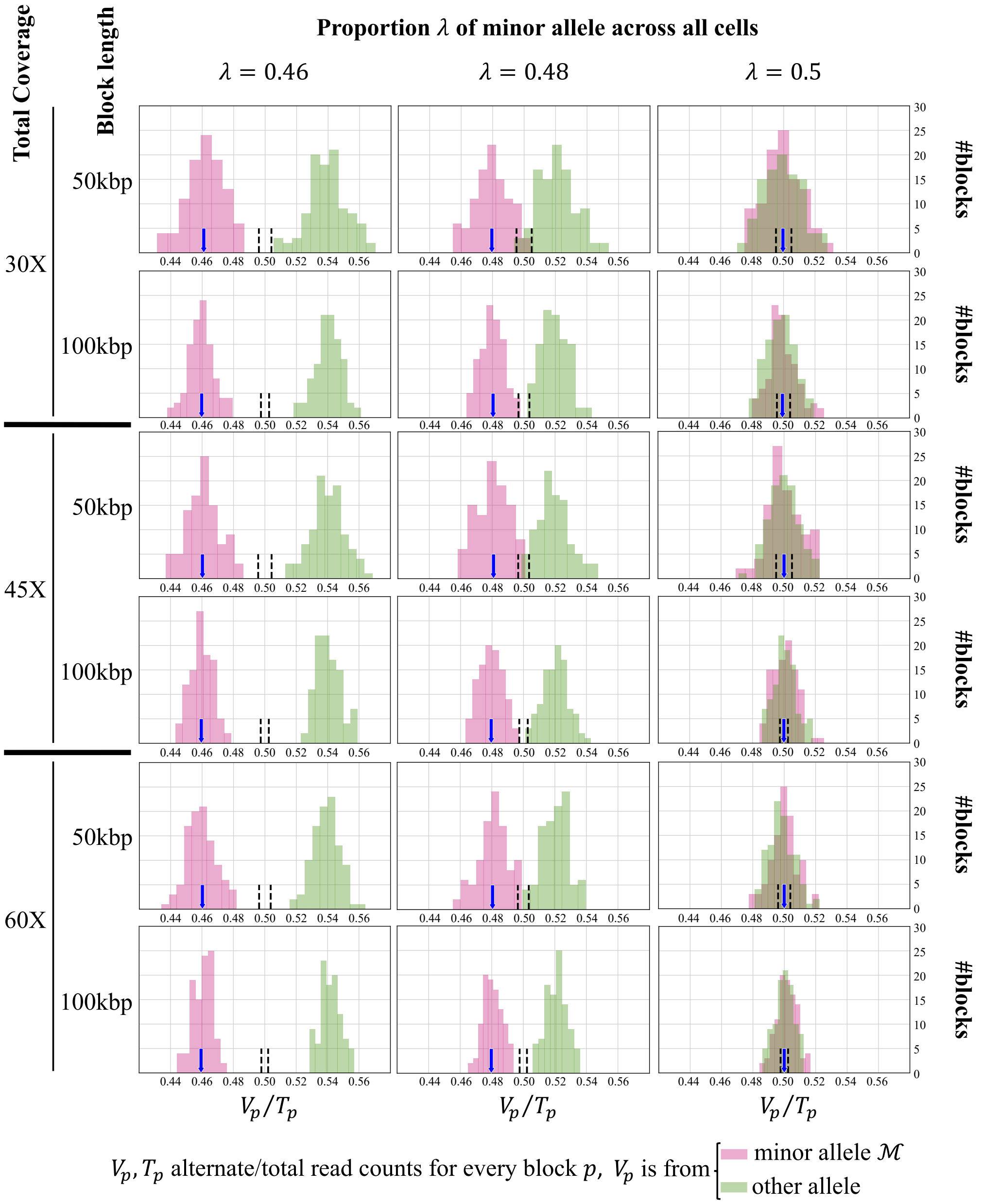
Supplementary Fig. 21: GC bias in RDRs estimated by CHISEL with a matched-normal sample. Read-depth ratios (RDRs) as a function of GC content in 5Mb genomic bins (points) across all 1448 cells from section E of breast cancer patient S0 (left) and across four individual cells, including two tumor cells (top right) and two diploid cells (bottom right). A linear regression (red line) and 95% confidence interval (shaded area) indicate that RDRs show little apparent GC bias.



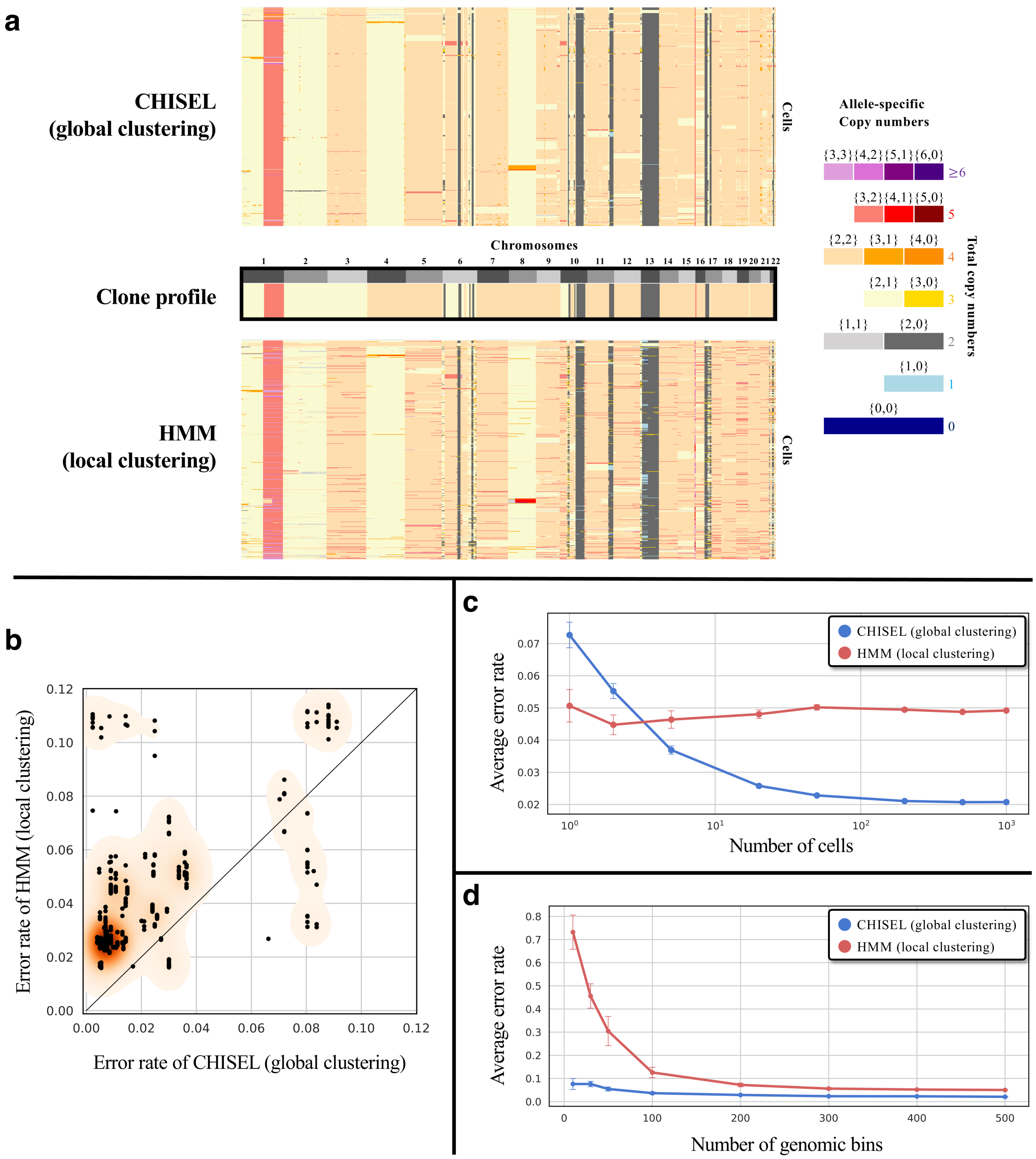
Supplementary Fig. 22: Probability of correctly phasing haplotype blocks within a 5Mb genomic bin with allelic imbalance (BAF=0.48 vs. BAF=0.52) as a function of block length and total sequencing coverage. The probability of correctly phasing haplotype blocks across cells for various haplotype-block lengths and total sequencing coverages (i.e. the product of number of cells and sequencing coverage per cell) is computed analytically. 50kb haplotype blocks give a probability >95% (white rectangle) when the total sequencing coverage across all cells is 50×.



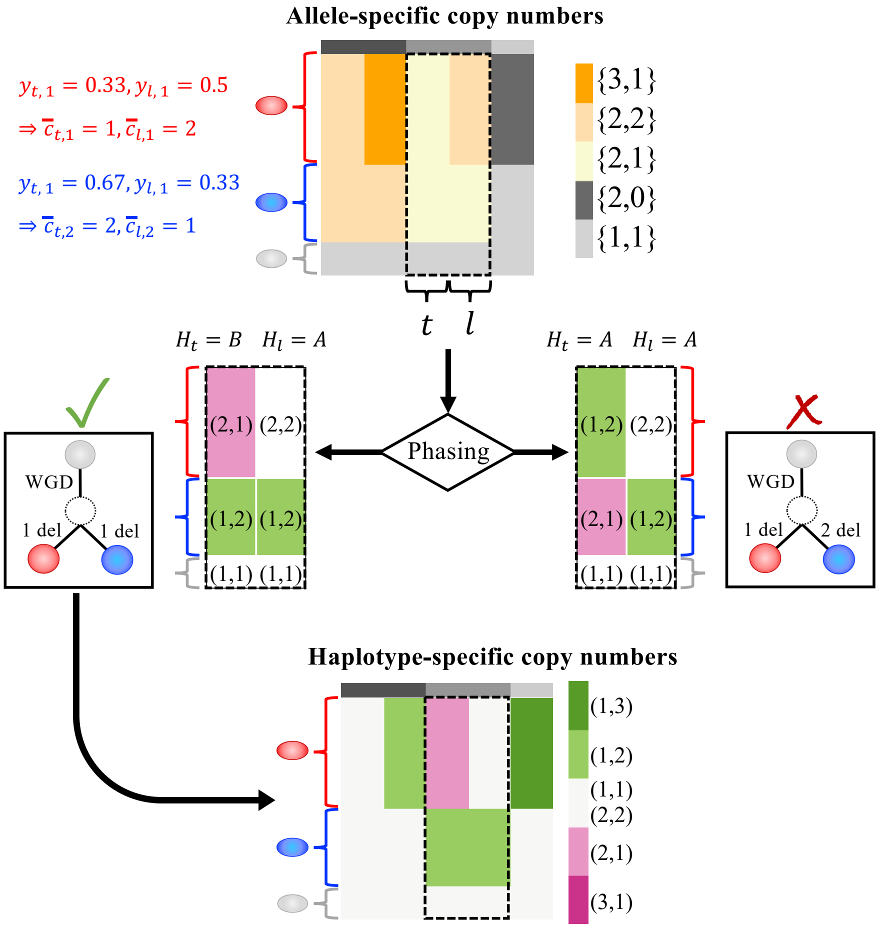
Supplementary Fig. 23: Error in the estimation of minor-allele proportion as function of the length of haplotype blocks and the length of genomic bins. The maximum error (left) and average error (right) is computed over 100 random samples. For each combination of lengths of haplotype blocks and genomic bins, the error is calculated for 2000 single cells with a sequencing coverage 0.03× per cell. The maximum and average errors are low (<0.005 and <0.002, respectively) using 50kb of haplotype blocks and 5Mb of genomic bins (white rectangles).



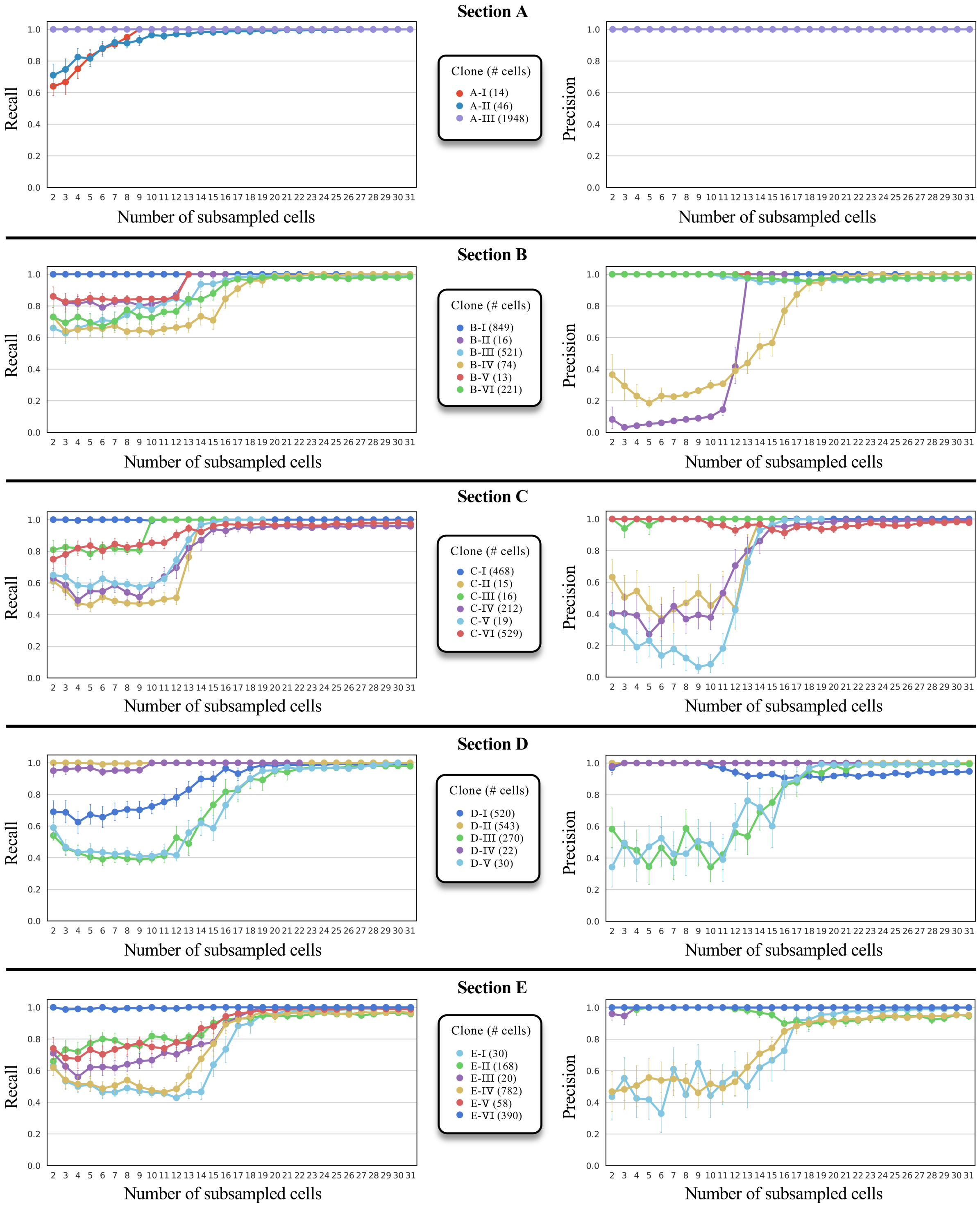
Supplementary Fig. 24: CHISEL accurately infers the minor-allele proportion of a genomic bin across cells. Alternate and total read counts V_p, T_p across cells are simulated for 200 haplotype blocks p assuming that the alternate reads belongs to the minor/major allele (magenta/green color) in 50%/50% of the blocks and varying the block length (50kb or 100kb), the total sequencing coverage (across all cells with 30X, 45X, and 60X), and the minor-allele proportion λ (0.46, 0.48, and 0.5). CHISEL accurately estimates the value Y (blue arrow) of λ in every case. Assuming allelic balance (i.e. $\lambda = 0.5$), a 95% confidence interval (CI) for Y (black dashed lines) is computed with a bootstrapping approach ($n = 400$). The estimated Y is correctly within CI only when $\lambda = 0.5$. Thus, CHISEL can accurately infer the haplotypes for all cases but those with $\lambda = 0.5$ by consistently assigning the lower/major counts to the same haplotype, since $V_p/T_p \leqslant 0.5$ for almost every block p whose count V_p is from the minor allele (i.e. magenta).



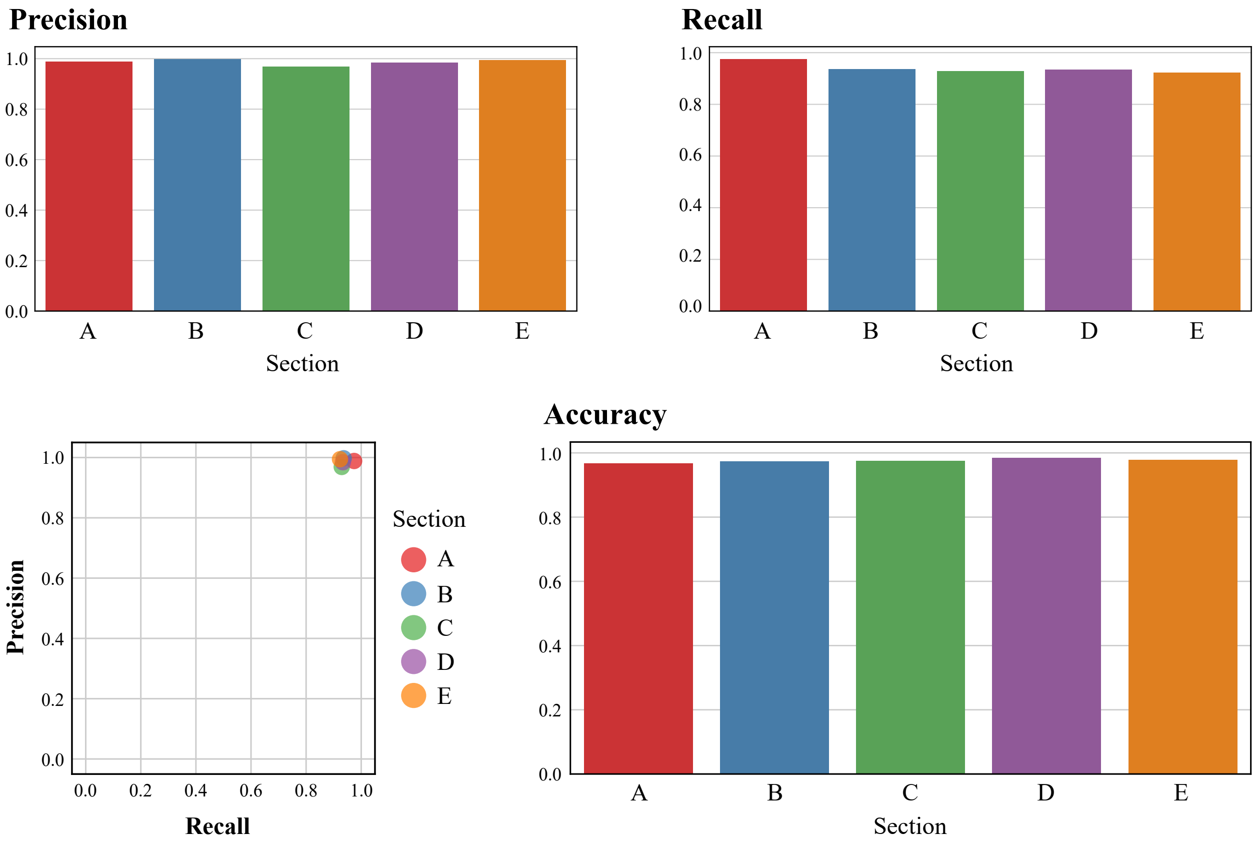
Supplementary Fig. 25: CHISEL's global clustering outperforms local clustering from an HMM in estimating allele-specific copy numbers. **a**, Allele-specific copy numbers inferred by CHISEL (top) or an HMM (bottom) on a random sample of 500 cells from the breast cancer patient S0 that were identified as belonging to the same clone, whose consensus profile is shown (middle). **b**, The error rate for CHISEL 's global clustering (x-axis) and the HMM (y-axis) for each 5Mb genomic bin (point) along the entire genome. The error rate of the HMM is higher (above diagonal) than the error rate of CHISEL for 502/539 genomic bins (shaded area represents a Gaussian kernel-density estimation). **c**, The average error rates of CHISEL (blue) and the HMM (red) across the genome as a function of the number of subsampled cells. Each bar indicates the standard error of the mean over 200 subsampled datasets. **d**, The average error rates of CHISEL (blue) and the HMM (red) across the genome as a function of the number of genomic bins. Each bar indicates the standard error of the mean over 200 subsampled datasets.


Allele-specific copy numbers
$y_{t,1} = 0.33, y_{l,1} = 0.5$
$\Rightarrow \bar{c}_{t,1} = 1, \bar{c}_{l,1} = 2$
$y_{t,1} = 0.67, y_{l,1} = 0.33$
$\Rightarrow \bar{c}_{t,2} = 2, \bar{c}_{l,2} = 1$
{3,1}
{2,2}
{2,1}
{2,0}
{1,1}
t
l
$H_t = B$ $H_l = A$
$H_t = A$ $H_l = A$
Phasing
(2,1) (2,2)
(1,2) (1,2)
(1,1) (1,1)
(1,2) (2,2)
(2,1) (1,2)
(1,1) (1,1)
WGD
1 del
1 del
WGD
1 del
2 del
Haplotype-specific copy numbers
(1,3)
(1,2)
(1,1)
(2,2)
(2,1)
(3,1)

Supplementary Fig. 26: CHISEL infers haplotype-specific copy numbers by modelling CNA evolution. Allele-specific copy numbers (top) are inferred for n cells (rows) in 5 distinct genomic regions (columns) across three chromosomes (grey rectangles in the first row). Cells are grouped into three distinct subpopulations (red, blue, and gray) with different allele-specific copy numbers. In the two regions t, l of the second chromosome, red cells have allele-specific copy numbers $\{2, 1\}$ and $\{2, 2\}$, blue cells have $\{2, 1\}$ in both regions, and gray cells have $\{1, 1\}$ in both regions. Since red cells have BAFs $y_{t,1} = 0.33$ and $y_{l,1} = 0.5$, the minor allele $\mathcal{M}_t$ has copy numbers $\bar{c}_{t,1} = 1$ and $\bar{c}_{l,1} = 2$ in these cells. Since blue cells have BAFs $y_{t,2} = 0.67$ and $y_{l,2} = 0.33$, the minor allele $\mathcal{M}_t$ has copy numbers $\bar{c}_{t,2} = 2$ and $\bar{c}_{l,2} = 1$ in these cells. Thus, there are 2 distinct ways of phasing the allele-specific copy numbers of t, l into haplotype-specific copy numbers. (Right) The first phasing places $\mathcal{M}_t$ and $\mathcal{M}_l$ on the same haplotype $\mathcal{A}$ (i.e. $H_t = A$ and $H_l = A$), resulting in t and l having different haplotype-specific copy numbers in the blue cells, i.e. $(2, 1)$ and $(1, 2)$, respectively. When a WGD occurs, this phasing results in an evolutionary scenario composed of three deletions: one deletion affects t in the red cells and the other two deletions affect t and l in the blue cells on two distinct haplotypes. (Left) The second phasing places $\mathcal{M}_t$ and $\mathcal{M}_l$ on the two different haplotypes $\mathcal{B}$ and $\mathcal{A}$, respectively (i.e. $H_t = B$ and $H_l = A$), resulting in t and l having the same haplotype-specific copy numbers $(1, 2)$ in the blue cells. When a WGD occurs, this phasing results in an evolutionary scenario composed of two deletions: one deletion affecting t in the red cells, as before, and the other affecting both t and l in the blue cells on the same haplotype. Under a principle of parsimony, CHISEL chooses the left evolutionary scenario with the minimum number of CNAs and infers the corresponding haplotype-specific copy numbers (bottom). Interestingly, the first and less likely scenario is chosen when assuming that the minor alleles $\mathcal{M}_t$ and $\mathcal{M}_l$ (i.e. those with less copies across all cells) are located on the same haplotype (i.e. $H_t = A$ and $H_l = A$). This happens because $\mathcal{M}_t, \mathcal{M}_t$ are defined across all cells, therefore $\mathcal{M}_t$ is the allele with fewer copies in the red cells since there are more red then blue cells, while $\mathcal{M}_l$ is the allele with fewer copies in the blue cells.



Supplementary Fig. 27: CHISEL accurately recovers clones containing as few as 10-20 cells. CHISEL was run on datasets obtained by subsampling cells from each clone in each tumor section. Recall (y-axis on left-side plots) and precision (y-axis on right-side plots) were recorded over $n = 100$ independent samples (dots indicate the mean and bars indicate the standard deviations) for each subsampled clone (different color in each distinct section) and for varying numbers of subsampled cells (x-axis).



Supplementary Fig. 28: Accurate identification of diploid cells from the counts of aligned sequencing reads. Precision, recall, and accuracy are computed in every tumor section of patient S0 for the diploid cells identified by the method designed to form a pseudo matched-normal sample. This method identifies diploid cells only using the counts of aligned sequencing reads and it has been applied with a threshold of 90% on the fraction of the genome with potential total copy number equal to 2. The true diploid cells have been defined as those inferred with total copy numbers 2 by previous total copy-number analysis and with allele-specific copy numbers $\{1, 1\}$ by CHISEL.

Supplementary Results

1 Comparison of the total copy numbers inferred by CHISEL with previous analysis

We compared the total copy numbers inferred by CHISEL for all cells in section E of patient S0 with those reported in previous total copy-number analysis[3]. More specifically, previous analysis used Cell Ranger DNA[4] to infer total copy numbers and showed that its results were consistent[5] with Ginkgo[6], another state-of-the-art method for inferring total copy numbers from single-cell DNA sequencing data. We observed that CHISEL and Cell Ranger DNA infer similar tumor purity, genome ploidy, and total copy numbers for the ≈90% of cell across all sections of patient S0 (Supplementary Fig. 2a-c). However, CHISEL infers more accurate total copy numbers than Cell Ranger DNA for the remaining ≈10% of cells (Supplementary Fig. 3) and with less outlying and noisy CNAs (Supplementary Fig. 4). The higher accuracy of the copy numbers inferred by CHISEL is directly related to the novel key features of CHISEL described in Methods. First, the more accurate total copy numbers inferred by CHISEL reflect the advantages of the constrained and probabilistic approach that jointly infers the scale factor γ and the allele-specific copy numbers by integrating BAFs. In contrast, the different genome ploidies inferred by Cell Ranger DNA across different subpopulations of cells reflect the issues in inferring the ploidy and the related total copy numbers among a large number of feasible solutions without considering BAFs. Second, the fewer outlying and noisy CNAs inferred by CHISEL reflect the advantages of the global clustering of RDRs and BAFs across all genomic bins, which leverages the shared evolutionary process among the cells. In contrast, the isolated and small CNAs inferred by Cell Ranger DNA reflect the issues of the standard local clustering in identifying genomic bins with the same copy numbers through local segmentation of genomic bins.

2 Analysis of DOP-PCR single-cell DNA sequencing data with CHISEL

We applied CHISEL to a DOP-PCR[7] single-cell sequencing dataset comprising 89 cells from breast cancer patient P2 in Kim et al. (2018)[1] with an average coverage of ≈0.24× per cell. We performed this analysis to investigate the performance of CHISEL on a single-cell dataset with a very different trade-off between the number of sequenced cells (i.e. 89 vs. ≈2 000) and sequencing coverage per cell (i.e. ≈0.24× vs. <0.03×) compared to the 10X Genomics CNV datasets analyzed in this work. Since a matched-normal sample was not available, we used the method that we have specifically introduced to form a pseudo matched-normal sample by extracting diploid cells. Due to the very low number of cells, the resulting pseudo matched-normal sample had a relatively low sequencing coverage of ≈9×, which limited the number of heterozygous germline SNPs that we could identify and phase. This issue resulted in noisy and high variable signals from DNA sequencing data, in addition to the high noise that has been already noted in the previous analysis[1]. As such, we applied CHISEL using default parameters but using a modification of the BIC model that more stringently penalizes solutions with more free parameters, using an additional factor $\gamma = 6$ similar to previous studies[8,9] in order to deal with the high-variability and noise of this dataset.

CHISEL identified the presence of a diploid clone and a tumor clone whose total copy numbers were largely in agreement with the published analysis[1] of the same 89 cells (Supplementary Fig. 18a-b). However, CHISEL also computed allele-specific CNAs (Supplementary Fig. 18c-d) revealing large regions of 11 chromosomes with copy-neutral loss of heterozygosity (LOH) which were erroneously classified as normal diploid regions in the published analysis[1]. For example, CHISEL found a copy-neutral LOH on chromosome 13 which affects *BRCA2* and *RB1*, genes implicated in breast cancer[10].

3 Comparison of global and local clustering with varying numbers of cells and bins

We compared the performance of the global clustering method used by CHISEL with an Hidden Markov Model (HMM) local method on a large number of subsampled datasets with varying number of cells or varying number of bins. In particular, we used TITAN[11], a method that uses an HMM to infer CNAs from jointly RDRs and BAFs. TITAN is commonly applied to bulk tumor samples, and to perform an appropriate comparison we applied TITAN to single cells in single genome mode (i.e. limiting to the presence of a single genome and fixing the tumor purity to be 100%) and also fixing the tumor ploidy to the same value inferred by CHISEL. Note that a slight variation of the same HMM method has been previously used to infer total copy numbers from RDRs in low-coverage single-cell sequencing data[12,13]. We generated ≈2 000 datasets by subsampling either a varying number of cells between 1 and 1 000 or a varying number of genomic bins between 1 and 500. We applied both the global clustering of CHISEL and the local HMM method of TITAN to each of these datasets. Since the two methods inferred very similar allele-specific copy numbers (i.e. in most of the cells the differences affected <10% of the genome) and indicated the presence of the same clones, we used the allele-specific copy numbers of these clones as ground truth. We compute the error rate for every genomic bin as the fraction of cells with different allele-specific copy numbers in the bin.

We observed that CHISEL's global clustering identifies allele-specific copy numbers with a substantially lower error rate than the HMM method (Supplementary Fig. 25a). Specifically, the HMM has an error rate that is 2-fold higher than CHISEL's global clustering method in >93% of genomic regions (Supplementary Fig. 25b). The error rate of the HMM is constant with increasing cell number, as the HMM does not use information from multiple cells. In contrast, the error rate of CHISEL's global clustering decreases steadily with increasing number of cells, and is >2-fold lower than the HMM with >5 cells (Supplementary Fig. 25c). These results confirm that the global clustering is able to leverage the shared information across multiple cells to accurately infer allele-specific copy numbers from the RDRs and BAFs of single cells. In addition, the error in CHISEL's global clustering is consistently lower than the error of the HMM when the number of bins is varied; in contrast, the HMM's error increases dramatically as the number of bins in decreased (Supplementary Fig. 25d).

4 Forming a pseudo matched-normal sample from read counts of single cells

We integrated in CHISEL a method to identify diploid cells only from the numbers of sequencing reads aligned to genomic bins. The CHISEL's analysis of allele- and haplotype-specific copy numbers uses a matched-normal sample for two specific steps in the computation of RDRs and BAFs: (1) normalization of read counts to correct for mapping and other biases; (2) identification of heterozygous germline SNPs for obtaining haplotype blocks through reference-based phasing algorithms. As a matched-normal sample may be unavailable in certain cases, this integrated method can be used to identify diploid cells and we can thus pool the corresponding sequencing reads to form a pseudo matched-normal sample; the details of the methods are reported in Supplementary Note 10. To assess the accuracy in the identification of diploid cells, we applied this method on all the tumor sections of patient S0 by only considering the number of aligned sequencing reads. Since this integrated method is independent from the allele- and haplotype-specific analysis of CHISEL, we use the diploid clone identified by CHISEL and the one identified in previous total copy number analysis as the ground truth. Therefore, we observed that the integrated method of CHISEL enables the accurate identification of diploid cells only from the counts of aligned sequencing reads (Supplementary Fig. 28).

Supplementary Methods

1 Computation of read-depth ratio

CHISEL partitions the reference genome into fixed-size bins; note that breakpoints are thus inferred only at the bin level when two neighboring bins are inferred with different haplotype-specific copy numbers. As such, CHISEL infers the read-depth ratio (RDR) $x_{t,i}$ of every bin t in cell i from the corresponding DNA sequencing data. Suppose we sequence $T_{t,i}$ reads aligned to t from i and $\breve{T}_t$ reads aligned to t from a matched-normal sample. Similarly to existing CNA methods for bulk sequencing[9, 14–19], we normalize $T_{t,i}$ using $\breve{T}_t$ to correct for mapping biases, for example due to GC content and low mappability, and to account for potential differences in the length of the bins as well as for the presence of germline copy-number variations in the normal genome. Moreover, we observe that the total number of reads substantially vary across different cells and, thus, we further normalize the read counts by the total number of reads $R_i, \breve{R}$ obtained from i and from the matched-normal sample, respectively. As such, we compute the RDR $x_{t,i}$ as follows

$$x_{t,i} = \frac{T_{t,i}}{\breve{T}_t} \frac{\breve{R}}{R_i} \tag{5}$$

On the 10 single-cell sequencing datasets, we observed that such estimated RDRs do not exhibit GC bias (Supplementary Fig. 21). CHISEL provides to the user the option of further applying a correction for GC bias using a LOWESS function as in existing single-cell methods[6, 12], especially when a matched-normal sample is not available.

2 Selecting the lengths of genomic bins and haplotype blocks

The selection of appropriate lengths for genomic bins and haplotype blocks depends on the sequencing coverage of individual cells and the total sequencing coverage across all cells. We use a probabilistic model to select these lengths, which we describe below. For the 10X Chromium single-cell datasets analyzed in this paper, we show that genomic bins of length 5Mb and haplotype blocks of length 50kb give high probabilities of identifying allelic imbalance and correcting phasing haplotype blocks within a bin. The probabilistic model is available in the CHISEL software and can be used to choose the lengths of genomic bins and haplotype blocks in datasets with different sequencing coverages.

We select the length of the genomic bins used in CHISEL so that the estimated BAFs have a high probability of distinguishing allele-specific CNAs in individual cells with the same total copy numbers but different allele-specific copy numbers; e.g. $\{3, 3\}$ vs. $\{4, 2\}$. Thus, the choice of the length of genomic bins is related to the sequencing coverage per cell, since CHISEL estimates the BAF $y_{t,i}$ of a genomic bin t in each cell i by using the sequencing reads that cover all the SNPs within bin t in cell i. We calculate this probability as follows. First, given the sequencing coverage per cell, we estimate the expected number of reads from a single cell that cover heterozygous germline SNPs in a genomic bin under the assumption that the $\approx$1.6 million heterozygous germline SNPs that can be reliably phased with current reference-based phasing methods[20, 21] are uniformly distributed along the genome. Next, using a binomial model similar to Eq. (7), we compute the probability $\mathrm{P}(V_{t,i} > V_{l,i})$ of distinguishing the minor-allele read counts $V_{t,i}$,$V_{l,i}$ from two bins t, l with allele-specific copy numbers $\{3, 3\}$ and $\{4, 2\}$. We chose $\{3, 3\}$ and $\{4, 2\}$ because these are the most difficult allele-specific copy numbers to distinguish (i.e. smallest difference in allelic proportions) for the most frequent values of total copy numbers (i.e. $\leqslant$6)[22]. We found that with average coverage of 0.025× per cell (typical coverage for 10X Genomics CNV solution) 5Mb bins provide BAF estimates that allow CHISEL to accurately distinguish allele-specific CNAs in individual cells, while smaller bin sizes (<1Mb) are reasonable if sequencing coverage per cell is 5-fold higher (Supplementary Fig. 20).

We select the length of haplotype blocks used in CHISEL so that there is a high probability of correctly phasing haplotype blocks within a bin. Thus, the choice of the length of haplotype blocks is related to the total sequencing coverage across all cells, since CHISEL phases haplotype blocks according to the difference in allelic proportions across all cells. Thus, the length of haplotype blocks is related to the number of reads covering germline SNPs in each block that are required to distinguish small differences in proportions of the two alleles (e.g. 0.48 vs. 0.52). CHISEL phases the haplotype blocks using the idea that the lower (respectively higher) read counts belong to the same allele when there is allelic imbalance (first step of CHISEL). Thus, we estimate the total number of reads in each block as above and use a binomial model for the numbers V', V'' of reads that belong to the allele with proportion p and the other allele with proportion $1-p$, respectively. Assuming $p > 1-p$ without loss of generality, we analytically compute the probability $\mathrm{P}(V' > V'')$ for different total sequencing coverages across all cells and various block lengths. We found with a total sequencing coverage of 50× across cells (the coverage for the datasets from the 10X Genomics CNV Solution that we analyzed in the manuscript) haplotype blocks of length of 50kb have a probability >95% of correct phasing when the proportions of the two haplotypes are 0.48 and 0.52 (Supplementary Fig. 22).

An additional criterion affecting the choice of lengths of genomic bins and haplotype blocks is the accuracy in the estimation of the minor-allele proportion across all cells (the first step of CHISEL). Thus, we estimated the error in the CHISEL's estimation of the minor-allele proportion for different lengths of genomic bins and haplotype blocks. Specifically, we simulated read counts from the two alleles using a binomial model with allelic proportion 0.5 (since this is the most difficult value to estimate) with number of trials equal to the expected number of reads on real data for each choice of lengths of genomic bins and haplotype blocks. We used CHISEL's EM algorithm to estimate the allelic proportion for these simulated read counts and computed the error in the estimated allelic proportion for each simulation, as well as the maximum and average error observed over 100 simulations. We found that 5Mb bins and 50kb haplotype blocks yielded low maximum and average errors for the sequencing coverages of the datasets in this work (Supplementary Fig. 23).

3 Expectation-maximization algorithm to estimate the minor-allele proportion across cells

We designed an expectation-maximization (EM) algorithm[23] to compute the maximum-likelihood estimation (MLE) Y_t of the proportion λ_t of copies of the minor allele $\mathcal{M}_t$ for a bin t across all cells. According to the definitions in Methods, the minor allele $\mathcal{M}_t$ is the allele of t with fewer copies across all cells such that $0.0 \leqslant \lambda_t \leqslant 0.5$ (and $0.0 \leqslant Y_t \leqslant 0.5$) and $\bar{c}_{t,i}$ is the copy number of $\mathcal{M}_t$ in cell i. Therefore, λ_t is the fraction of $\bar{c}_{t,i}$ across every cell i, i.e. $\lambda_t = \frac{\sum_i \bar{c}_{t,i}}{\sum_i c_{t,i}}$ where $c_{t,i}$ is the total copy number of t in i. First we describe the model and problem formulation, and next we describe the EM algorithm with the two composing steps.

We observe the number of sequencing reads covering k blocks of a bin t. Each block is composed of two sequences of nucleotides at consecutive SNPs called the reference and alternate sequences, with each sequence located on a different allele of t. As such, we observe in each block p the total number T_p of reads covering block p across all cells and the corresponding number V_p of reads only covering the alternate sequence of p across all cells. More specifically, these read counts are obtained by pooling the corresponding sequencing reads from all cells with $T_p = \sum_i T_{p,i}$ and $V_p = \sum_i V_{p,i}$ since the reference and alternate sequences of p are defined uniquely across all cells. We do not know whether the alternate sequence of p is located on either $\mathcal{M}_t$ or the other allele of t and we represent the two possibilities with the phase h_p as follows

$$h_p = \begin{cases} 1 & \text{if the alternate sequence of } p \text{ belongs to } \mathcal{M}_t \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Assuming reads are sequenced uniformly, every read sequenced from t and from all the cells has a probability of covering the alternate sequence of p equal to λ_t if $h_p = 1$ and equal to $1 - \lambda_t$ if $h_p = 0$. Therefore, we model the alternate-specific number V_p of reads for every block p as the following binomial distribution

$$V_p \sim \begin{cases} \mathrm{Binom}(T_p, \lambda_t) & \text{if } h_p = 1 \\ \mathrm{Binom}(T_p, 1 - \lambda_t) & \text{if } h_p = 0 \end{cases} \tag{7}$$

When observing the total number $T_p = \tau_p$ and the alternate-specific number $V_p = \nu_p$ of reads, we aim to infer the MLE Y_t of λ_t. For simplicity, we denote all the observations $(T_1 = \tau_1, \ldots, T_k = \tau_k)$ by $\mathbf{T}$ and all the observations $(V_1 = \nu_1, \ldots, V_k = \nu_k)$ by $\mathbf{V}$. We thus aim to solve the following problem.

Problem 1. *Given the observed total and alternate-specific numbers $\mathbf{T}, \mathbf{V}$ of reads for all blocks in the same bin t, find the MLE Y_t of λ_t, i.e.*

$$Y_t = \operatorname*{argmax}_{\overline{Y}_t} \Pr(\mathbf{V} \mid \lambda_t = \overline{Y}_t, \mathbf{T}) \tag{8}$$

To solve Problem 1, we designed an EM algorithm which corresponds to an iterative algorithm: every iteration r aims to find an estimation $Y_t^{(r+1)}$ of λ_t by computing a lower bound of the likelihood in Eq. (8) from a previous estimation $Y_t^{(r)}$ and the complete likelihood $\Pr(\mathbf{V}, \mathbf{h} \mid \lambda_t = Y_t, \mathbf{T})$, where the phases $\mathbf{h} = (h_1, \ldots, h_k)$ are the latent variables. In fact, we can easily compute the complete likelihood $\Pr(V_p = \nu_p, h_p \mid \lambda_t = Y_t, T_p = \tau_p)$ of every block p from the model in Eq. (7) as follows

$$\begin{gathered} \Pr(V_p = \nu_p, h_p \mid \lambda_t = Y_t, T_p = \tau_p) = \\ \left(\Pr(h_p = 1)\binom{\tau_p}{\nu_p} Y_t^{\nu_p} (1 - Y_t)^{\tau_p - \nu_p}\right)^{h_p} \left(\Pr(h_p = 0)\binom{\tau_p}{\nu_p} (1 - Y_t)^{\nu_p} Y_t^{\tau_p - \nu_p}\right)^{1-h_p} \end{gathered} \tag{9}$$

As such, the iterations start with a random value $Y_t^{(0)}$ and end at convergence. In particular, every iteration is composed of two steps, the E-step and the M-step, and we use a large number (> 400) of random restarts to deal with local optima, such that Y_t is chosen has the estimation with highest likelihood among all restarts. In the following two subsections we describe the E-step and the M-step of the algorithm.

3.1 Expectation in the E-step

The E-step aims to compute a function $g_r(Y_t \mid Y_t^{(r)})$ as the expected value of the complete log-likelihood with respect to the phases $\mathbf{h}$, given the current value $Y_t^{(r)}$ and the observations $\mathbf{V}, \mathbf{T}$, i.e

$$g_r(Y_t \mid Y_t^{(r)}) = \mathbb{E}_{\mathbf{h}|Y_t^{(r)},\mathbf{V},\mathbf{T}}[\log \Pr(\mathbf{V}, \mathbf{h} \mid \lambda_t = Y_t, \mathbf{T})] \tag{10}$$

The resulting function $g_t(Y_t \mid Y_t^{(r)})$ thus provides a lower bound for the objective function in Eq. (8)[23].

We first compute a closed formula for the complete log-likelihood $\log \Pr(\mathbf{V}, \mathbf{h} \mid \lambda_t = Y_t, \mathbf{T})$. In particular, we observe that the alternate-specific numbers $V_1, \ldots, V_k$ of reads are stochastically independent by the model in Eq. (7) when Y_t is given. As such, assuming that $\Pr(h_p = 1) = \Pr(h_p = 0) = 0.5$ for every block p, we have the following

$$\begin{aligned}
&\log \Pr(\mathbf{V}, \mathbf{h} \mid \lambda_t = Y_t, \mathbf{T}) = \\
&\log \prod_p \Pr(V_p = \nu_p, h_p \mid \lambda_t = Y_t, \mathbf{T}) = \\
&\sum_p \log \left(\left(0.5 \binom{\tau_p}{\nu_p} Y_t^{\nu_p} (1 - Y_t)^{\tau_p - \nu_p} \right)^{h_p} \left(0.5 \binom{\tau_p}{\nu_p} (1 - Y_t)^{\nu_p} Y_t^{\tau_p - \nu_p} \right)^{1 - h_p} \right) = \\
&\sum_p h_p \left(\log 0.5 + \log \binom{\tau_p}{\nu_p} + \nu_p \log Y_t + (\tau_p - \nu_p) \log(1 - Y_t) \right) \\
&+ \sum_p (1 - h_p) \left(\log 0.5 + \log \binom{\tau_p}{\nu_p} + \nu_p \log(1 - Y_t) + (\tau_p - \nu_p) \log Y_t \right)
\end{aligned} \tag{11}$$

Next, we derive the function $g_r(Y_t \mid Y_t^{(r)})$ as the expected value of Eq. (11) with respect to the phases $\mathbf{h}$, given the current value $Y_t^{(r)}$ and the observations $\mathbf{V}, \mathbf{T}$, i.e.

$$\begin{aligned}
&g_r(Y_t \mid Y_t^{(r)}) = \\
&\mathbb{E}_{\mathbf{h}|Y_t^{(r)},\mathbf{V},\mathbf{T}}[\log \Pr(\mathbf{V}, \mathbf{h} \mid \lambda_t = Y_t, \mathbf{T})] = \\
&\mathbb{E}_{\mathbf{h}|Y_t^{(r)},\mathbf{V},\mathbf{T}} \left[\begin{array}{l} \sum_p h_p \left(\log 0.5 + \log \binom{\tau_p}{\nu_p} + \nu_p \log Y_t + (\tau_p - \nu_p) \log(1 - Y_t) \right) \\ + \sum_p (1 - h_p) \left(\log 0.5 + \log \binom{\tau_p}{\nu_p} + \nu_p \log(1 - Y_t) + (\tau_p - \nu_p) \log Y_t \right) \end{array} \right] = \\
&\sum_p \mathbb{E}_{\mathbf{h}|Y_t^{(r)},\mathbf{V},\mathbf{T}}[h_p] \left(\log 0.5 + \log \binom{\tau_p}{\nu_p} + \nu_p \log Y_t + (\tau_p - \nu_p) \log(1 - Y_t) \right) \\
&+ \sum_p (1 - \mathbb{E}_{\mathbf{h}|Y_t^{(r)},\mathbf{V},\mathbf{T}}[h_p]) \left(\log 0.5 + \log \binom{\tau_p}{\nu_p} + \nu_p \log(1 - Y_t) + (\tau_p - \nu_p) \log Y_t \right)
\end{aligned} \tag{12}$$

To conclude the derivation of function $g_r(Y_t \mid Y_t^{(r)})$, we last need to compute the expected value of the phase h_p for every block p. We do this by applying the Bayes's theorem given the current value $Y_t^{(r)}$ and the observations $\mathbf{V}, \mathbf{T}$

as follows

$$\begin{aligned}
&\mathbb{E}_{\mathbf{h}|Y_t^{(r)},\mathbf{V},\mathbf{T}}[h_p] = \\
&\Pr(h_p = 1 \mid V_p = \nu_p, Y_t^{(r)}) = \\
&\frac{\Pr(V_p = \nu_p \mid Y_t^{(r)}, h_p = 1)0.5}{\Pr(V_p = \nu_p \mid Y_t^{(r)}, h_p = 0)0.5 + \Pr(V_p = \nu_p \mid Y_t^{(r)}, h_p = 1)0.5} = \\
&\frac{\binom{\tau_p}{\nu_p}(Y_t^{(r)})^{\nu_p}(1-Y_t^{(r)})^{\tau_p-\nu_p}}{\binom{\tau_p}{\nu_p}(1-Y_t^{(r)})^{\nu_p}(Y_t^{(r)})^{\tau_p-\nu_p} + \binom{\tau_p}{\nu_p}(Y_t^{(r)})^{\nu_p}(1-Y_t^{(r)})^{\tau_p-\nu_p}} = \\
&\frac{(Y_t^{(r)})^{\nu_p}(1-Y_t^{(r)})^{\tau_p-\nu_p}}{(1-Y_t^{(r)})^{\nu_p}(Y_t^{(r)})^{\tau_p-\nu_p} + (Y_t^{(r)})^{\nu_p}(1-Y_t^{(r)})^{\tau_p-\nu_p}}
\end{aligned} \tag{13}$$

To avoid numerical precision loss in the computation, we specifically compute Eq. (13) in the following equivalent function

$$\begin{aligned}
&\frac{(Y_t^{(r)})^{\nu_p}(1-Y_t^{(r)})^{\tau_p-\nu_p}}{(1-Y_t^{(r)})^{\nu_p}(Y_t^{(r)})^{\tau_p-\nu_p} + (Y_t^{(r)})^{\nu_p}(1-Y_t^{(r)})^{\tau_p-\nu_p}} = \\
&\frac{1}{1 + (Y_t^{(r)})^{\tau_p-2\nu_p}(1-Y_t^{(r)})^{-(\tau_p-2\nu_p)}} = \\
&\frac{1}{1 + e^{(\tau_p-2\nu_p)\ln Y_t^{(r)} - (\tau_p-2\nu_p)\ln(1-Y_t^{(r)})}}
\end{aligned} \tag{14}$$

3.2 Maximization in the M-step

The M-step aims to find the new value $Y_t^{(r+1)}$ of Y_t which maximizes the function $g_r(Y_t \mid Y_t^{(r)})$ given the previous $Y_t^{(r)}$, i.e.

$$Y_t^{(r+1)} = \operatorname*{argmax}_{Y_t} g_r(Y_t \mid Y_t^{(r)}) \tag{15}$$

The function $g_r(Y_t \mid Y_t^{(r)})$ is defined by combining Eq. (12) and Eq. (13), and for simplicity we denote by z_p the value of Eq. (13) computed as in Eq. (14) for every block p. As such, we aim to find $Y_t^{(r+1)}$ as the value of Y_t which maximizes the following function

$$\begin{aligned}
f(Y_t) =& \sum_p z_p \left(\log 0.5 + \log\binom{\tau_p}{\nu_p} + \nu_p \log Y_t + (\tau_p - \nu_p)\log(1-Y_t)\right) \\
&+ \sum_p (1-z_p)\left(\log 0.5 + \log\binom{\tau_p}{\nu_p} + \nu_p \log(1-Y_t) + (\tau_p - \nu_p)\log Y_t\right)
\end{aligned} \tag{16}$$

To do this, we first compute the derivative of the function $f(Y_t)$ in Eq. (16) and, next, we find the value $Y_t^{(r+1)}$ for which the derivative is equal to zero.

First, we compute the derivative of the function $f(Y_t)$ in Eq. (16) as follows

$$\begin{aligned}\frac{df}{dY_t} =& \sum_p z_p \left(\frac{\nu_p}{Y_t} - \frac{\tau_p - \nu_p}{1 - Y_t} \right) + \sum_p (1 - z_p) \left(-\frac{\nu_p}{1 - Y_t} + \frac{\tau_p - \nu_p}{Y_t} \right) \\ =& \sum_p \frac{z_p \nu_p}{Y_t} - \frac{z_p \tau_p - z_p \nu_p}{1 - Y_t} - \frac{\nu_p}{1 - Y_t} + \frac{\tau_p - \nu_p}{Y_t} + \frac{z_p \nu_p}{1 - Y_t} - \frac{z_p \tau_p - z_p \nu_p}{Y_t} \\ =& \sum_p \frac{2 z_p \nu_p + \tau_p - \nu_p - z_p \tau_p}{Y_t} - \sum_p \frac{z_p \tau_p + \nu_p - 2 z_p \nu_p}{1 - Y_t} \\ =& \frac{\sum_p 2 z_p \nu_p + \tau_p - \nu_p - z_p \tau_p}{Y_t} - \frac{\sum_p z_p \tau_p + \nu_p - 2 z_p \nu_p}{1 - Y_t} \\ =& \frac{\alpha}{Y_t} - \frac{\gamma}{1 - Y_t}\end{aligned} \tag{17}$$

where we define $\alpha = \sum_p 2z_p\nu_p + \tau_p - \nu_p - z_p\tau_p$ and $\gamma = \sum_p z_p\tau_p + \nu_p - 2z_p\nu_p$, for simplicity.

We now compute the value of $Y_t^{(r+1)}$ which maximizes the function $f(Y_t)$ in Eq. (16) and thus $\frac{df(Y_t^{(r+1)})}{dY_t} = 0$, where the derivative of $f(Y_t)$ is the one derived in Eq. (17). We assume that $Y_t \neq 0$ and $Y_t \neq 1$ as those cases can be easily identified when we observe $\nu_p = 0$ or $\nu_p = \tau_p$ for every block p in t. As such we compute the value $Y_t^{(r+1)}$ which satisfies the following equation

$$\begin{aligned}&\frac{\alpha}{Y_t^{(r+1)}} - \frac{\gamma}{1 - Y_t^{(r+1)}} = 0 \\ &\frac{\alpha - \alpha Y_t^{(r+1)} + \gamma Y_t^{(r+1)}}{Y_t^{(r+1)}(1 - Y_t^{(r+1)})} = 0 \\ &\alpha - (\alpha + \gamma) Y_t^{(r+1)} = 0 \\ &Y_t^{(r+1)} = \frac{\alpha}{\alpha + \gamma}\end{aligned} \tag{18}$$

We substitute α, γ with their actual values, and we obtain the following unique value for $Y_t^{(r+1)}$

$$\begin{aligned}Y_t^{(r+1)} &= \frac{\alpha}{\alpha + \gamma} \\ Y_t^{(r+1)} &= \frac{\sum_p 2 z_p \nu_p + \tau_p - \nu_p - z_p \tau_p}{\sum_p 2 z_p \nu_p + \tau_p - \nu_p - z_p \tau_p + z_p \tau_p + \nu_p - 2 z_p \nu_p} \\ Y_t^{(r+1)} &= \frac{\sum_p \tau_p (1 - z_p) + \nu_p (2 z_p - 1)}{\sum_p \tau_p}\end{aligned} \tag{19}$$

4 Phasing haplotype blocks

We seek phases $h_1, \dots, h_k$ of all k blocks within a bin t such that the corresponding BAF $y_{t,i}$ computed as

$$y_{t,i} = \frac{\sum_{p=1}^{k} h_p V_{p,i} + (1 - h_p)(T_{p,i} - V_{p,i})}{\sum_{p=1}^{k} T_{p,i}}. \tag{20}$$

provides an accurate estimation of $\frac{\bar{c}_{t,i}}{c_{t,i}}$ (see Eq. (1) in Methods). Remember that $\bar{c}_{t,i}$ is the copy number of the minor allele $\mathcal{M}_t$ in cell i and is thus equal to one of the two corresponding allele-specific copy numbers, i.e. $\bar{c}_{t,i} = \hat{c}_{t,i}$ or $\bar{c}_{t,i} = \check{c}_{t,i}$. In particular, we phase the blocks by using the proportion $Y_t \in [0, 0.5]$ of $\mathcal{M}_t$ estimated in Supplementary Methods 3. As in the previous sections, we observe the total number T_p of reads covering block p across all cells and the corresponding number V_p of reads only covering the alternate sequence of p across all cells. As such, we compute the phases $h_1, \ldots, h_k$ differently when $Y_t \not\approx 0.5$ or $Y_t \approx 0.5$, which correspondingly indicate the presence of allelic imbalance or balance. The power to distinguish these two cases with the proportion Y_t inferred by the EM algorithm depends on the total read counts $T_1, \ldots, T_k$. While most of the bins contain sufficiently high read counts to distinguish the two values accurately, there are few exceptions in genomic regions with a lower density of SNPs. As such, we choose for each bin independently a threshold to confidently distinguish the two values. More specifically, we compute a 95% confidence interval by bootstrapping and we observe from simulations the ability to accurately distinguish the two cases when considering the read counts along each block and across all cells (Supplementary Fig. 24).

Allelic imbalance across cells When $Y_t \not\approx 0.5$, we aim to infer the maximum-likelihood estimate $\hat{h}_p$ of the phase h_p of every block p given the observed total numbers $\mathbf{T} = (T_1 = \tau_1, \ldots, T_k = \tau_k)$ of reads, the observed alternate-specific numbers $\mathbf{V} = (V_1 = \nu_1, \ldots, V_k = \nu_k)$ of reads, and the minor-allele proportion $Y_t \in [0, 0.5]$. For simplicity we denote by $\mathbf{h} = (h_1, \ldots, h_k)$ the phases of all k blocks. In the following theorem, we show that the phases obtained by consistently assigning the lower (respectively higher) read counts of all the k blocks to the same allele provide a maximum likelihood estimate for the phases $\mathbf{h}$.

Theorem 1. *Given the total and alternate-specific numbers* $\mathbf{T}, \mathbf{V}$ *of reads, the values* $\hat{\mathbf{h}}$ *defined as follows*

$$\forall p \text{, either} \qquad \hat{h}_p = \begin{cases} 1 & \text{if } 2X_p \leqslant N_p \\ 0 & \text{otherwise} \end{cases} \qquad \text{or} \qquad \hat{h}_p = \begin{cases} 1 & \text{if } 2X_p > N_p \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

provide a maximum likelihood estimation of the phases $\mathbf{h}$ *such that*

$$\log \Pr(\mathbf{V} \mid \hat{\mathbf{h}}, \mathbf{T}, Y_t) \geqslant \operatorname*{argmax}_{\mathbf{h}} \log \Pr(\mathbf{V} \mid \mathbf{h}, \mathbf{T}, Y_t) \tag{22}$$

for any value of the minor-allele proportion $Y_t \in [0, 0.5]$.

Proof. By contradiction, we assume there exist phases $\tilde{\mathbf{h}}$ different from the one described in Theorem 1, i.e. $\tilde{\mathbf{h}} \neq \hat{\mathbf{h}}$ and $\tilde{\mathbf{h}} \neq \underline{1} - \hat{\mathbf{h}}$ with $\underline{1} = (1, \ldots, 1)$, such that $\Pr(\mathbf{V} \mid \tilde{\mathbf{h}}, \mathbf{T}, Y_t) \geqslant \Pr(\mathbf{V} \mid \mathbf{h}, \mathbf{T}, Y_t)$ for every possible phases $\mathbf{h}$. As we assume that read counts across the blocks of the same bin are stochastically independent, we have the following

$$\begin{aligned} &\Pr(\mathbf{V} \mid \tilde{\mathbf{h}}, \mathbf{T}, Y_t) \geqslant \Pr(\mathbf{V} \mid \mathbf{h}, \mathbf{T}, Y_t) = \\ &\sum_p \Pr(V_p = \nu_p \mid \tilde{h}_p, T_p = \tau_p, Y_t) \geqslant \sum_p \Pr(V_p = \nu_p \mid h_p, T_p = \tau_p, Y_t) \end{aligned} \tag{23}$$

Since $\tilde{\mathbf{h}} \neq \hat{\mathbf{h}} \wedge \tilde{\mathbf{h}} \neq \underline{1} - \hat{\mathbf{h}}$ and due to the assumption in Eq. (23), we know there exists at least one block o with $\hat{h}_o \neq \tilde{h}_o$ where $\Pr(V_o = \nu_o \mid \tilde{h}_o, T_o = \tau_o, Y_t) > \Pr(V_o = \nu_o \mid \hat{h}_o, T_o = \tau_o, Y_t)$. This inequality thus leads to the following implications according to the definition in Eq. (21) and the model in Eq. (7).

$$\begin{aligned} &\Pr(V_o = \nu_o \mid \tilde{h}_o, T_o = \tau_o, Y_t) > \Pr(V_o = \nu_o \mid \hat{h}_o, T_o = \tau_o, Y_t) \Rightarrow \\ &\begin{cases} (1 - Y_t)^{\nu_o} Y_t^{\tau_o - \nu_o} > Y_t^{\nu_o} (1 - Y_t)^{\tau_o - \nu_o} & \text{if } 2\nu_o \leqslant \tau_o \\ Y_t^{\nu_o} (1 - Y_t)^{\tau_o - \nu_o} > (1 - Y_t)^{\nu_o} Y_t^{\tau_o - \nu_o} & \text{otherwise} \end{cases} \Rightarrow Y_t > 0.5 \end{aligned} \tag{24}$$

Since $\tilde{\mathbf{h}} \neq \hat{\mathbf{h}} \wedge \tilde{\mathbf{h}} \neq \underline{1} - \hat{\mathbf{h}}$, we know there also exists at least another block q with $\hat{h}_q = \tilde{h}_q$ and, therefore, $\Pr(V_q = \nu_q \mid \tilde{h}_q, T_q = \tau_q, Y_t) = \Pr(V_q = \nu_q \mid \hat{h}_q, T_q = \tau_q, Y_t)$. However, we also know that $\Pr(V_q = \nu_q \mid \tilde{h}_q, T_q = \tau_q, Y_t) \geqslant \Pr(V_q = \nu_q \mid 1 - \tilde{h}_q, T_q = \tau_q, Y_t)$ because the phases $\tilde{\mathbf{h}}$ provide the maximum value of the log likelihood across all phases from the assumption in Eq. (23). As such, the following implications hold according to the definition in Eq. (21) and the model in Eq. (7).

$$\begin{gathered}\Pr(V_q = \nu_q \mid \hat{h}_q, T_q = \tau_q, Y_t) \geqslant \Pr(V_q = \nu_q \mid 1 - \hat{h}_q, T_q = \tau_q, Y_t) \Rightarrow \\ \begin{cases} Y_t^{\nu_o}(1 - Y_t)^{\tau_o - \nu_o} \geqslant (1 - Y_t)^{\nu_o} Y_t^{\tau_o - \nu_o} & \text{if } 2\nu_o \leqslant \tau_o \\ (1 - Y_t)^{\nu_o} Y_t^{\tau_o - \nu_o} \geqslant Y_t^{\nu_o}(1 - Y_t)^{\tau_o - \nu_o} & \text{otherwise} \end{cases} \Rightarrow Y_t \leqslant 0.5\end{gathered} \tag{25}$$

Eq. (24) and Eq. (25) are in contradiction and this proves Theorem 1. □

The haplotype phases $\hat{\mathbf{h}}$ obtained from Theorem 1 are typically used to estimate BAF in previous bulk-tumor studies because they provide an unbiased estimator for Y_t in any genomic region t with allelic imbalance, i.e. $Y_t \not\approx 0.5$[24–26]. Intuitively, $Y_t \not\approx 0.5$ indicates the presence of allelic imbalance in t and nucleotides with lower read counts belong to the same haplotype[27]; this approach is especially accurate when considering high read counts, as those we generally observe in this case along a whole block and across all cells. In particular, we show that Theorem 1 can accurately infer the phases $\mathbf{h}$ by simulating the expected read counts from the haplotype blocks across a bin, varying the values of Y_t and considering different experimental settings that influence the observed read counts (Supplementary Fig. 24). The capability of accurately identifying the true haplotypes from the read counts thus result in the BAF $y_{t,i}$ computed in Eq. (20) to be an accurate estimate of the corresponding proportion $\frac{\bar{c}_{t,i}}{c_{t,i}}$ of the minor-allele copies for every bin t and cell i.

Allelic balance across cells When $Y_t \approx 0.5$, the bin t is allelic balanced across all cells and the haplotype phases $\hat{\mathbf{h}}$ from Theorem 1 provide a biased estimator of Y_t as well as of the phases $\mathbf{h}$[28]. In fact the expected numbers of reads for the two sequences of every block p are approximately the same and, therefore, we cannot accurately infer the phases $\mathbf{h}$ from the observed read counts. However, we show that we can obtain different phases in this case such that the corresponding BAF $y_{t,i}$ provides an unbiased estimation of $\frac{\bar{c}_{t,i}}{c_{t,i}}$ for every cell i. In particular, we do this under a clearly-stated assumption: we assume that $Y_t \approx 0.5$ implies that $\frac{\bar{c}_{t,i}}{c_{t,i}} = \frac{\hat{c}_{t,i}}{c_{t,i}} = \frac{\check{c}_{t,i}}{c_{t,i}} = 0.5$ in every cell i. We consider this to be a reasonable assumption because there are only two possible violations, which are rare. The first violation corresponds to the rare case where all tumor cells are partitioned in two (or its multiples) almost-exact halves with precisely inverted haplotype-specific copy numbers for t (e.g. $(2, 1)$ for 50% of the cells and $(1, 2)$ for the other 50%). Moreover, when this rare case occurs, the two halves likely have other CNAs that distinguish the two distinct clones and CHISEL can be applied to each clone separately in a second-pass approach to identify the missing CNAs. The second violation occurs when there is only an extremely low number of tumor cells with CNAs in t, which we can ignore as there are always few noisy cells.

Under the previous assumption, we observe that the alternate-specific number $V_{p,i}$ of reads from every block p in cell i is independent from the haplotype phase h_p when $\frac{\bar{c}_{t,i}}{c_{t,i}} = 0.5$. In fact, we can model $V_{p,i}$ similarly to the alternate-specific number V_p of reads across all cells in Eq. (20) as follows:

$$V_{p,i} \sim \begin{cases} \mathrm{Binom}(T_{p,i}, \frac{\overline{c}_{t,i}}{c_{t,i}}) & \text{if } h_p = 1 \\ \mathrm{Binom}(T_{p,i}, 1 - \frac{\overline{c}_{t,i}}{c_{t,i}}) & \text{if } h_p = 0 \end{cases} \tag{26}$$

When $\frac{\overline{c}_{t,i}}{c_{t,i}} = 0.5$, the corresponding model in Eq. (26) yields $V_{p,i} \sim \mathrm{Binom}(\nu_p, 0.5)$. Therefore, the BAF $y_{t,i}$ computed in Eq. (20) with randomly-chosen phases provide an unbiased estimator $\frac{\overline{c}_{t,i}}{c_{t,i}}$, as stated in the following theorem.

Theorem 2. *When $\frac{\overline{c}_{t,i}}{c_{t,i}} = 0.5$, the BAF $y_{t,i}$ computed in Eq. (20) is an unbiased estimator of $\frac{\overline{c}_{t,i}}{c_{t,i}}$ when the phase h_p of each block p is uniformly selected at random according to $\Pr(h_p = 1) = \Pr(h_p = 0) = 0.5$.*

Proof. According to Eq. (20), the phases $h_1, \ldots, h_k$ provide the following value of the BAF $y_{t,i}$

$$y_{t,i} = \frac{\sum_{p=1}^{k} V_{p,i} h_p + (\tau_{p,i} - V_{p,1}) h_p}{\sum_{p=1}^{k} \tau_{p,i}} \tag{27}$$

when we observe the total numbers $T_{1,i} = \tau_{1,i}, \ldots, T_{1,k}, = \tau_{1,k}$ of reads. Since $\frac{\overline{c}_{t,i}}{c_{t,i}} = 0.5$, the alternate-specific number $V_{p,i}$ of reads is distributed according to the model in Eq. (26) as follows

$$V_p \sim \mathrm{Binom}(\tau_p, 0.5) \tag{28}$$

because $\overline{c}_{t,i} = \widehat{c}_{t,i} = \check{c}_{t,i}$. As such, the alternate-specific number $V_{p,i}$ of reads and the phase h_p are stochastically independent for every block p by Eq. (28). Thus, the expected value of BAF $y_{t,i}$ is equal to the following

$$\mathbb{E}[y_{t,i}] = \frac{\sum_{p=1}^{k} \mathbb{E}[V_{p,i}]\, \mathbb{E}[h_p] + (\tau_{p,i} - \mathbb{E}[V_{p,i}])\, \mathbb{E}[h_p]}{\sum_{p=1}^{k} \tau_{p,i}} \tag{29}$$

The expected value of V_p is $\mathbb{E}[V_p] = \tau_p 0.5$ by the model in Eq. (28) and the expected value of the phase h_p is $\mathbb{E}[h_p] = 1 \Pr(h_s = 1) + 0 \Pr(h_s = 0) = 0.5$. As such, the expected value of $y_{t,i}$ is equal to the following

$$\mathbb{E}[y_{t,i}] = \frac{\sum_p \tau_p 0.5^2 + (\tau_p - \tau_p 0.5) 0.5}{\sum_p \tau_p} = 0.5 \tag{30}$$

Since $\mathbb{E}[y_{t,i}] - \frac{\overline{c}_{t,i}}{c_{t,i}} = 0$, $y_{t,i}$ is an unbiased estimator of $\frac{\overline{c}_{t,i}}{c_{t,i}}$ when this proportion is equal to 0.5. □

Theorem 2 guarantees that the BAF $y_{t,i}$ computed as in Eq. (20) using phases chosen uniformly at random is an unbiased estimator of $\frac{\overline{c}_{t,i}}{c_{t,i}}$.

5 Inferring scale factor and allele-specific copy numbers

In each cell, CHISEL infers the scale factor γ and the allele-specific copy numbers $\{\widehat{c}_t, \check{c}_t\}$ of every bin t in a two-stage procedure: first, CHISEL identifies the potential candidates of the scale factor γ by integrating the BAFs and, second, CHISEL chooses γ among the candidates and the corresponding maximum likelihood allele-specific copy numbers using the standard Bayesian information criterion (BIC). In this section, we describe the details of this procedure in three subsections. In Section 5.1, we prove that the cell-specific scale factor γ can be computed by identifying the total copy number of a single genomic region in each cell. In Section 5.2, we use the BAFs to identify all the potential candidates of γ by selecting the largest balanced cluster and its potential allele-specific copy numbers. In Section 5.3, CHISEL chooses γ among the candidates and the corresponding maximum-likelihood allele-specific copy numbers of every bin using BIC.

5.1 Identification of scale factor from a given total copy number

Suppose that a total of R sequencing reads of length E have been sequenced uniformly from the genome of the cell, and $\breve{R}$ reads from a matched-normal sample. In this work (Supplementary Methods 1) as well as in previous bulk-tumor[7,11,16,18,24,27–34] and single-cell[4–6,12,13,35–38] analysis, the RDR x_t of bin t is defined as the ratio between the total number T_t of reads observed in t from the cell and the corresponding number $\breve{T}_t$ of reads observed from a matched-normal sample. To account for varying total number of reads across different cells, we further normalize x_t by the R and $\breve{R}$, and we thus obtain the following

$$x_t = \frac{T_t}{\breve{T}_t}\frac{\breve{R}}{R}. \tag{31}$$

According to the Lander-Waterman equation[39], the average number of reads obtained from a single copy of a genomic position in the cell is

$$\frac{ER}{L} \tag{32}$$

where L is the genome length of the cell, corresponding to the sum of the total copy numbers of every genomic positions, i.e. $L = \sum_t \ell_t c_t$ when ℓ_t is the number of genomic positions in every bin t. The read count T_t corresponds to the total number of reads obtained from all the copies of the genomic positions in t and is hence equal to

$$T_t = \frac{ER}{L}\ell_t c_t \tag{33}$$

by Eq. (32) because the cell contains c_t copies of the ℓ_t genomic positions in t. Similarly, the read count $\breve{T}_t$ from the matched-normal sample, which is assumed to only contain normal diploid cells, is equal to the following

$$\breve{T}_t = \frac{E\breve{R}}{\breve{L}}\ell_t 2 \tag{34}$$

because normal diploid cell have 2 copies of every genomic region and where $\breve{L}$ is the normal genome length, i.e. $\breve{L} = \sum_t \ell_t 2$. By combining the definitions in Eq. (33) and Eq. (34) with Eq. (35), we thus model the RDR x_t as follows

$$x_t = \frac{\breve{L}R}{L\breve{R}2}c_t \tag{35}$$

Observe that $L, \breve{L}, R, \breve{R}$ are constant values across all the bins of the cell. Therefore, we define a cell-specific constant scale factor $\gamma = \breve{L}R/(L\breve{R}2)$ and we obtain from Eq. (35) the corresponding direct proportionality between x_t and c_t as follows

$$x_t = \gamma c_t \tag{36}$$

Following Eq. (36), we know that the scale factor γ determines the direct proportionality between the total copy number c_t and the RDR x_y for each bin t. Note that the scale factor γ is the same across all bins because γ only depends on cell-specific constant values. Therefore, knowing the total copy number c_t of a single bin t is sufficient to compute the corresponding γ. The following theorem results from this simple observation, which is a direct extension of the equivalent theorem previously introduced in HATCHet[33].

Theorem 3. *Given the RDR x_t and the total copy number c_t of a bin t, the scale factor has a unique value which is computed as follows:*

$$\gamma = \frac{c_t}{x_t} \tag{37}$$

5.2 Constrained identification of candidates for the scale factor

We use the BAFs to identify the candidate values of γ under the assumption that the genome of every cell contains a reasonable number of *balanced* bins, i.e. bins with equal copy numbers $\hat{c}_t = \check{c}_t$ of both alleles. This assumption follows from the observation that, in a cell, bins unaffected by CNAs have allele-specific copy numbers $\{1, 1\}$ without WGD, $\{2, 2\}$ with one WGD, and so on. CHISEL thus identifies these bins as the largest cluster whose BAF is approximately equal to 0.5, among those previously inferred in the second step of CHISEL. More specifically, we estimate the BAF of each cluster s using the EM algorithm from Supplementary Methods 3 applied to all the bins within s and where the read counts of each bin are computed from the phases obtained in Supplementary Methods 4. To deal with potential over-clustering, we also merge in each cell all the clusters with very similar values of RDRs and BAFs and we obtain a set S containing all the identified balanced genomic bins. We thus apply Theorem 3 to infer all the potential values of γ by averaging the RDR of every bin in S and by considering the set Θ containing all the possible values for the allele-specific copy numbers of balanced bins, i.e. $\Theta = \{1, 2, \ldots\}$ according to the observation from previous assumption. Note that the total copy number of the balanced bins is equal to 2θ when $\theta \in \Theta$ is the value of the allele-specific copy numbers of the bins in S. Therefore, we compute the set Γ comprising the potential candidates for γ as follows

$$\Gamma = \{\gamma = \frac{2\theta}{\frac{1}{|S|}\sum_{t \in S} x_t} : \theta \in \Theta\} \tag{38}$$

5.3 Selecting scale factor and allele-specific copy numbers

CHISEL finally uses BIC to choose γ among the candidates in Γ and the corresponding maximum-likelihood pair $\{\hat{c}_t, \check{c}_t\}$ of allele-specific copy numbers for every bin t. Note that every bin t in a cluster s has the same allele-specific copy numbers, i.e. $|\{\{\hat{c}_t, \check{c}_t\} : t \in s\}| = 1$. Therefore, when denoting by $\{\hat{C}_s, \check{C}_s\}$ the pair of allele-specific copy numbers of every bin in cluster s, we model the RDR x_t and the mirrored BAF $\overline{y}_t$ of bin $t \in s$ as observations from two normal distributions

$$x_t \sim \mathcal{N}\left(\frac{\hat{C}_s + \check{C}_s}{\gamma}, \sigma_x\right) \qquad \text{and} \qquad \overline{y}_t \sim \mathcal{N}\left(\frac{\min\{\hat{C}_s, \check{C}_s\}}{\hat{C}_s + \check{C}_s}, \sigma_y\right) \tag{39}$$

where the sample variances σ_x, σ_y are estimated from the clusters inferred in the second step of CHISEL and the corresponding mean. For every candidate value of $\gamma \in \Gamma$, we find the maximum likelihood estimates for $\{\hat{C}_s, \check{C}_s\}$

using an exhaustive search, which is feasible as the number of candidate values of γ (e.g. 3 when considering the occurrence of at most 2 WGDs) and the number of distinct pairs $\{\widehat{C}_s, \widecheck{C}_s\}$ of allele-specific copy numbers for a cluster s are relatively small. More specifically, we identify for every $\gamma \in \Gamma$ the allele-specific copy numbers $\{\widehat{C}_s, \widecheck{C}_s\}$ of every cluster s which maximize the following log likelihood

$$\mathcal{L}(\gamma) = \sum_s \sum_{t \in s} \ln \Pr(x_t \mid \widehat{C}_s, \widecheck{C}_s, \gamma, \sigma_x) + \ln \Pr(\overline{y}_t \mid \widehat{C}_s, \widecheck{C}_s, \gamma, \sigma_y) \tag{40}$$

where every cluster is stochastically independent given γ, according to our model in Eq. (39). The higher the value of γ is, the higher the number $\kappa(\gamma)$ of possible combinations of allele-specific copy numbers, i.e.

$$\kappa(\gamma) = \sum_{c=0}^{C} \left(\left\lfloor \frac{c}{2} \right\rfloor + 1 \right) \tag{41}$$

where $C = \lceil \gamma \max_t x_t \rceil$ is the maximum total copy number. As such, we choose the candidate value of γ with minimum BIC to balance between higher likelihood and lower model complexity to avoid overfitting. In fact, higher values of γ always have higher likelihood but also higher model complexity, as they induce more combinations of allele-specific copy numbers according to Eq. (41). More specifically, we choose γ as the candidate that minimizes the corresponding BIC and we obtain the following

$$\gamma = \operatorname*{argmin}_{\gamma'} \ln(m)\kappa(\gamma') - 2\mathcal{L}(\gamma') \tag{42}$$

6 Dynamic programming algorithm to infer haplotype-specific copy numbers

CHISEL phases the minor allele $\mathcal{M}_t$ of every bin t to minimize the number of CNAs needed to explain the resulting haplotype-specific copy numbers $(a_{t,i}, b_{t,i})$ of every bin t in cell i. To do this, CHISEL identifies the phase $H_t \in \{\mathcal{A}, \mathcal{B}\}$ of the minor allele $\mathcal{M}_t$ such that $H_t = \mathcal{A}$ if the minor allele $\mathcal{M}_t$ is located on haplotype $\mathcal{A}$, and $H_t = \mathcal{B}$ otherwise. When the phase H_t of bin t is known, we can indeed obtain the haplotype-specific copy numbers $(a_{t,i}, b_{t,i})$ in every cell i as in the following

$$(a_{t,i}, b_{t,i}) = \begin{cases} (\overline{c}_{t,i}, c_{t,i} - \overline{c}_{t,i}) & H_t = \mathcal{A} \\ (c_{t,i} - \overline{c}_{t,i}, \overline{c}_{t,i}) & H_t = \mathcal{B} \end{cases} \tag{43}$$

given the minor-allele copy number $\overline{c}_{t,i}$, which corresponds to one of the two allele-specific copy numbers $\{\widehat{c}_{t,i}, \widecheck{c}_{t,i}\}$, and the total copy number $c_{t,i} = \widehat{c}_{t,i} + \widecheck{c}_{t,i}$. In fact, $\overline{c}_{t,i}$ can be easily obtained by combining the estimated BAF $y_{t,i}$ and the inferred pair $\{\widehat{c}_{t,i}, \widecheck{c}_{t,i}\}$ of allele-specific copy numbers when assuming that $\widehat{c}_{t,i} \geqslant \widecheck{c}_{t,i}$ w.l.o.g. as follows

$$\overline{c}_{t,i} = \begin{cases} \widehat{c}_{t,i} & y_{t,i} \geqslant 0.5 \\ \widecheck{c}_{t,i} & \text{otherwise} \end{cases} \tag{44}$$

CHISEL infers the phase H_t of $\mathcal{M}_t$ for every bin t to minimize the number of CNAs required to explain the haplotype-specific copy numbers $(\mathbf{a}_i, \mathbf{b}_i)$ of every cell i by using the model of interval events that model CNAs as events that either increase or decrease the copy numbers of neighboring genomic regions on the same haplotype[40–42].

Under a principle of parsimony, we thus aim to minimize the total number of interval events across all cells. Given the phases $H_{t-1}, H_t \in \{\mathcal{A}, \mathcal{B}\}$ for any pair of consecutive bins $t-1, t$, we compute the corresponding number $d(t, H_{t-1}, H_t)$ of interval events as follows

$$d(t, H_{t-1}, H_t) = \sum_i |a_{t-1,i} - a_{t,i}| + |b_{t-1,i} - b_{t,i}| \tag{45}$$

where the haplotype-specific copy numbers $(a_{t-1,i}, b_{t-1,i})$ and $(a_{t,i}, b_{t,i})$ are computed as in Eq. (43) given the allele-specific copy numbers $\bar{c}_{t-1,i}, \bar{c}_{t,i}$ of the minor alleles $\mathcal{M}_{t-1}, \mathcal{M}_t$. Therefore, we aim to solve the following problem.

Problem 2. *Given the minor-allele and total copy numbers $\bar{c}_{t,i}, c_{t,i}$ of every bin t in every cell i, find the phases $H_1^*, \ldots, H_m^* \in \{\mathcal{A}, \mathcal{B}\}^m$ such that*

$$H_1^*, \ldots, H_m^* = \operatorname*{argmin}_{H_1, \ldots, H_m} \sum_{t=2}^{m} d(t, H_{t-1}, H_t) \tag{46}$$

To solve Problem 2, we design a dynamic programming algorithm (DP) since the objective can be computed recursively. In fact, the minimum number $D(l, H_l) = \min_{H_1, \ldots, H_{l-1}} \sum_{t=2}^{l} d(t, H_{t-1}, H_t)$ of interval events for the first l bins when the phase H_l of the the minor allele $\mathcal{M}_l$ for the last bin l is given can be computed as follows

$$D(l, H_l) = \min \begin{cases} D(l-1, \mathcal{A}) + d(l, \mathcal{A}, H_l) \\ D(l-1, \mathcal{B}) + d(l, \mathcal{B}, H_l) \end{cases} \tag{47}$$

As such, DP iteratively computes $D(l, H_l)$ for every bin l. In the following lemma, we prove the correctness of DP.

Lemma 4. *Given the minor-allele copy numbers $\bar{c}_{1,i}, \ldots, \bar{c}_{l,i}$ and total copy numbers $c_{1,i}, \ldots, c_{l,i}$ of the first l bins in every cell i, the following statements hold:*

1. *if $D(l, H_l) = \lambda$, there exists phases $H_1, \ldots, H_l$ such that $\sum_{t=2}^{l} d(t, H_{t-1}, H_t) \leqslant \lambda$;*

2. *if there exists phases $H_1, \ldots, H_l$ with $\sum_{t=2}^{l} d(t, H_{t-1}, H_t) = \lambda$, it follows $D(l, H_l) \leqslant \lambda$.*

Proof. We prove both the statements of the lemma by induction on the m bins. The statements obviously hold for $t = 2$ because $\min\{d(2, \mathcal{A}, H_2), d(2, \mathcal{B}, H_2)\} \leqslant d(2, H_1, H_2)$ for any $H_1 \in \{\mathcal{A}, \mathcal{B}\}$ when H_2 is known. Therefore, we assume by induction that each statement holds for $l-1$ and we prove that it also holds for l.

1. We assume by induction that, if $D(l-1, H_{l-1}) = \lambda'$, there exists phases $H_1, \ldots, H_{l-1}$ with $\sum_{t=2}^{l-1} d(t, H_{t-1}, H_t) \leqslant \lambda'$. By the definition in Eq. (47), there exists a phase $H_{l-1} \in \{\mathcal{A}, \mathcal{B}\}$ for bin $t-1$ such that $D(l-1, H_{l-1}) = D(l-2, H_{l-2}) + d(l-1, H_{l-2}, H_{l-1})$. If $D(l-1, H_{l-1}) = \lambda'$, the inductive assumption thus determines the existence of phases $H_1, \ldots, H_{l-1}$ such that $\sum_{t=2}^{l-1} d(t, H_{t-1}, H_t) \leqslant \lambda'$. Adding the phase H_t of the minor allele $\mathcal{M}_t$ thus results in having phases $H_1, \ldots, H_{l-1}, H_l$ for the first l bins with $\sum_{t=2}^{l-1} d(t, H_{t-1}, H_t) + d(l, H_{l-1}, H_l) \leqslant \lambda' + d(l, H_{l-1}, H_l)$. If $D(l, H_l) = \lambda$, the number of events for the phases $H_1, \ldots, H_{l-1}, H_l$ is equivalent to the following $\sum_{t=2}^{l} d(t, H_{t-1}, H_t) \leqslant \lambda$, since $D(l-1, H_{l-1}) = \lambda'$ and $D(l, H_l) = D(l-1, H_{l-1}) + d(l, H_{l-1}, H_l)$. This proves the first statement.

2. We assume by induction that, if there exists phases $H_1, \ldots, H_{l-1}$ with $\sum_{t=2}^{l-1} d(t, H_{t-1}, H_t) = \lambda'$, it follows $D(l-1, H_{l-1}) \leqslant \lambda'$. We consider any combination $(H_1, \ldots, H_{l-1}, H_l) \in \{\mathcal{A}, \mathcal{B}\}^l$ of phases for the minor alleles of the first l bins such that $\sum_{t=2}^{l} d(t, H_{t-1}, H_t) \leqslant \lambda$. When $\sum_{t=2}^{l-1} d(t, H_{t-1}, H_t) = \lambda'$ for the first $l-1$ bins only, it follows that $D(l-1, H_{l-1}) \leqslant \lambda'$ by the inductive assumption and we equivalently obtain that all the phases $H_1, \ldots, H_{l-1}, H_l$ thus result in having the following number of interval events $\sum_{t=2}^{l} d(t, H_{t-1}, H_t) = \lambda' + d(l, H_{l-1}, H_l)$ with $\sum_{t=2}^{l} d(t, H_{t-1}, H_1) \geqslant D(l-1, H_{l-1}) + d(l, H_{l-1}, H_l)$. By definition in Eq. (47), we know that $D(l, H_l) \leqslant D(l-1, H_{l-1}) + d(l, H_{l-1}, H_l)$ and therefore $D(l, H_l) \leqslant \lambda$ because $\sum_{t=2}^{l} d(t, H_{t-1}, H_t) = \lambda$ by definition. This proves the second statement.

□

Lemma 4 proves that DP solves Problem 2 because the first statement guarantees the existence of a solution for any value of $D(l, H_l)$ and the second statement proves the optimality of the corresponding solution. As such, DP infers the phases $H_1, \ldots, H_m$ of the minor alleles $\mathcal{M}_1, \ldots, \mathcal{M}_m$ that minimize the number of interval events by iteratively computing all the values of $D(l, H_l)$ and by using the standard backtracking strategy for dynamic programming. The resulting running time is thus linear as it corresponds to $\mathcal{O}(m)$, and we obtain the following theorem, which determines that Problem 2 is solvable in linear time and it does belong to the complexity class P.

Theorem 5. *DP solves Problem 2 in linear time.*

7 Inferring clones from haplotype-specific copy numbers

CHISEL infers distinct subpopulations of cells, or *clones*, with the same complement of CNAs. While the presence of clones is expected from the cancer evolutionary process[43], we do not directly observe these clones and their identification is complicated by two main factors. First, we do not know the number N of clones. Second, the inferred copy numbers of each cell may be affected by errors in the measurements (e.g. due to low number of sequenced reads) or may be characterized by spurious aberrations (e.g. due to the different cell-cycle states). One thus needs to cluster the cells into the corresponding clones and to separate the noisy cells. Current methods do not directly perform this inference or they do it based on total copy numbers[4–6, 12, 13, 35–38]. However, distinct clones with different haplotype-specific copy numbers may have the same total copy numbers and even the same allele-specific copy numbers.

CHISEL identifies N clones by first clustering cells with sufficiently similar haplotype-specific copy numbers and then selecting clusters that correspond to clones. In particular, we define the *copy-number distance* $d(i, j)$ between two cells i, j as the fraction of the genome with different haplotype-specific copy numbers, i.e.

$$d(i, j) = \frac{1}{m} \sum_{t} \min\{|a_{t,i} - a_{t,j}| + |b_{t,i} - b_{t,j}|, 1\} \tag{48}$$

Since the inferred copy numbers are affected by errors in the measurements, we expect that any two cells in the same clone may have different haplotype-specific copy numbers in a maximum fraction ε of the genome (e.g. 4–7%). As such, CHISEL aims to group cells into a minimum number of clusters such that each cluster $I \subseteq \{1, \ldots, n\}$ only contains cells with copy-number distance below ε, i.e. $d(i, j) \leqslant \varepsilon$ for each pair of cells $i, j \in I$. To do this, CHISEL uses a standard algorithm for hierarchical clustering[44,45]. Next, CHISEL selects the clusters that correspond to clones by assuming that each clone contains at least a minimum number ι of cells, i.e. I is a clone if $|I| \geqslant \iota$. This is a reasonable assumption as subpopulations containing a small number of cells are more likely due to the presence of noisy cells. For visualization purposes, CHISEL also sorts all the cells to minimize the distance between neighbors. At last, CHISEL obtains a consensus copy-number profile for each inferred clone. Such profiles can be thus used to correct errors in the inferred haplotype-specific copy numbers.

8 Estimating the minimum size of detectable clones

We investigated the minimum size of a clone that CHISEL can accurately detect using the following approach. First, we generated in-silico datasets by subsampling cells from each clone identified by CHISEL. Each in-silico dataset is obtained by selecting a random subset of cells from a specific clone and all of the remaining cells. Second, we apply CHISEL to each such dataset and we measure precision and recall between the subsampled clone and the clone inferred by CHISEL that best matches the subsampled clone.

We applied the subsampling approach to each of the 5 single-cell datasets from the breast cancer patient S0. Specifically, for each clone and each tumor section, we generated a dataset containing a random subset of 2-31 cells of the clone and all the remaining cells. Note that this subsampling procedure preserves all other features of the single-cell dataset, including rates and sizes of the CNAs in different clones as well as errors and biases in the DNA sequencing signals. We ran CHISEL on each subsampled dataset and quantified CHISEL's ability to detect the small subpopulation of subsampled cells in terms of precision and recall. We found that CHISEL accurately recovers clones containing as few as 10-20 cells in every case (Supplementary Fig. 27). Since the minimum size of detectable clones may vary by dataset we added this subsampling method to the CHISEL software.

9 Selecting cells from barcodes

The standard approach for 10X Genomics Single-cell SNV Solution identifies individual cells by choosing the subset of barcodes that effectively correspond to the targeted cells[4]. This selection is necessary because the potential presence of spurious barcodes that do not characterize reads sequenced from the genome of an individual cell[5]. To do this, we applied the same standard approach which only selects the barcodes that are associated to a sufficiently large number of sequencing reads. More specifically, the threshold of 10^5 reads (i.e. corresponding to a sequencing coverage of $> 0.003\times$ per cell) on the minimum number of sequencing reads has been empirically computed in previous analysis[3]. Using the same threshold (which is user adjustable), CHISEL selects the same cells (with few exceptions) as in previous analysis[3].

10 Identifying diploid cells from read counts

CHISEL requires a matched-normal sample composed of normal diploid cells from the same patient for the estimation of RDRs and BAFs. More specifically, the computation of RDRs requires the read counts from the matched-normal sample for standard normalization and the computation of BAFs requires a matched-normal sample to identify germline heterozygous SNPs. However, a matched-normal sample may not be available. In this case, we thus propose to generate a pseudo matched-normal sample by identifying normal diploid cells and merge the corresponding sequencing reads. As such, we propose a method to identify normal diploid cells simply from the observed number $\tau_{t,i}$ of sequencing reads aligned to every bin t and cell i.

In the absence of errors and germline copy-number variations (CNV) in the normal genome, we know that the observed read counts are directly proportional to the corresponding total copy numbers (Supplementary Methods 1). In particular, $c_{t,i} = \gamma_i \tau_{t,i}$ for every bin t in a cell i with an unknown scaling factor $\gamma_i \in \mathbb{R}$ (Supplementary Methods 5). Moreover, we expect the total copy number of every bin to be equal to 2, i.e. $c_{t,i} = 2$, in any normal cell i under the previous assumption. When the error- and CNV-free assumption does not hold, we expect the read counts to be noisy, i.e. $\gamma_i \tau_{t,i} \approx 2$, and to be different than 2 in some cases, i.e. $\gamma_i \tau_{t,i} \neq 2$. We assume to know the maximum fraction ξ of the genome with total copy numbers different than 2 in normal diploid cells and we identify whether a cell i is a normal diploid cell in two steps. First, we aim to find γ_i using a method similar to the one applied by current single-cell methods for inferring total copy numbers[5,6,35–37]. As such, we aim to minimize the error between the expected total copy number 2 and the inferred total copy number $\lceil \gamma_i \tau_{t,i} \rfloor$ across all bins, i.e.

$$\gamma_i = \underset{\gamma}{\operatorname{argmin}} \sum_t |2 - \lceil \gamma_i \tau_{t,i} \rfloor| \tag{49}$$

Specifically, we solve this problem by using a local search across a large subset of potential values of γ_i, starting from the value estimated by averaging the read counts across all genome. Next, we classify the cell i as a normal diploid cell if and only if the fraction of the genome with an estimated total copy number equal to 2 is lower than ξ, i.e. i is normal if and only if

$$\frac{1}{m} \sum_t |2 - \lceil \gamma_i \tau_{t,i} \rfloor| \leqslant \xi \tag{50}$$

11 Identifying somatic single-nucleotide variants in single cells

We examined somatic single-nucleotide variants (SNVs) as an orthogonal signal for supporting the results obtained from copy-number analysis because these mutations were not used in either the copy-number inference or tree reconstruction. However, the identification of SNVs in individual cells is impractical from low-coverage DNA sequencing data and we thus propose a two-step approach. First, we pooled sequencing reads from all cells into a pseudo-bulk sample and we identified SNVs across all cells by using a standard method for bulk-tumor sequencing data, which is Varscan 2[46,47]. Second, we assigned each SNV to those cells with a variant read and we identified SNVs that are present in a subpopulation of cells, or clone, by considering whether any of the corresponding cells has a variant read. Unfortunately, this approach is complicated by two issues. On the one hand, the matched-normal sample may contain some number of aneuploid cells, for example the matched-normal sample used for patient S0 comprises $\approx 8\%$ of aneuploid tumor cells. We observed that the presence of tumor cells in the matched-normal sample result in several SNVs having a lower probability to be somatic mutations. On the other hand, some clones identified by CHISEL and in previous total-copy number analysis contain a relatively small number of cells; for example, some clones contain

<100 cells. The SNVs in these clones are thus covered by a number of sequencing reads lower than expected, for example we approximately expect only 2–4 reads covering each SNV in clones with ≈100 cells.

We dealt with the mentioned issues in the identification of somatic SNVs by introducing some additional filters and properly relaxing some of the parameters of Varscan 2. First, we lowered the minimum coverage to select potential SNV loci from 6 to 2 sequencing reads (according to the minimum expected number of covering reads in relatively-small clones) and, to reduce potential errors and false positives, we only considered SNVs with at least 2 variant reads in two distinct cells. Next, we selected only loci whose total number of covering reads across all cells is reasonable according to GATK best practices[48]. Since the low number of sequencing reads from relatively-small clones may result in genomic regions that are not covered by any sequencing read in these clones, we have also restricted the analysis to loci that are covered by sequencing reads in all clones. Last, we increased the threshold on the p-value score computed by Varscan 2 to choose somatic mutations. We did this to deal with the observed low frequencies of the SNVs that result in higher scores and are due to two main factors: (1) the low tumor purity, i.e. fraction of tumor cells, which is estimated to be between 60–83% for the tumor sections of patient S0 and is only ≈50% across all cells of patient S0; and (2) the identified occurrence of a whole-genome duplication (WGD) which is associated to a lower number of copies harboring the SNVs, for example a SNV occurred after WGD affects only 1 copy over 4 in genomic regions with no further CNAs. Moreover, the higher threshold allows to deal with the higher scores resulting from the presence of variant reads in the matched-normal sample due to the admixture of aneuploid tumor cells. As such, we empirically computed the threshold (i.e. 0.19–0.24) from the results of Varscan 2 in order to include the highest-quality SNVs that are covered by at least 2 variant reads in distinct cells.

12 Analysis of variant-allele frequency for somatic single-nucleotide variants

We examined the relationship between the variant-allele frequency (VAF) of each somatic single-nucleotide variant (SNV) and the clonal status of the SNV induced by the CHISEL tree for the 10 551 SNVs identified in the tumor clones. We calculated the VAF of each SNV using the standard definition as the fraction of variant reads over the total number of reads covering the SNV locus. When assuming that reads are sequenced uniformly, the VAF of a SNV e located in a bin t is an estimator of the fraction of copies of t harboring e across all cells. More specifically, the VAF of e is an estimator of the following fraction[2,49]

$$\frac{\sum_{i=1}^{n} \dot{c}_{e,t,i}}{\sum_{i=1}^{n} c_{t,i}} \tag{51}$$

where the *mutated copy number* $\dot{c}_{e,t,i}$ represents the number of copies of t harboring e in cell i.

We also defined a *restricted VAF* for a SNV with respect to a subpopulation of cells by restricting to sequencing reads with barcodes matching the cells in the subpopulation. In particular, we computed a *left-restricted VAF* and a *right-restricted VAF* by restricting to the sequencing reads from cells belonging to the left (clones `J-I` and `J-II`) and right (clones `J-III`, …, `J-VIII`) branches of the CHISEL tree. The restricted VAF of e is an estimator of the fraction of copies of the bin t harboring e similarly to Eq. (51) but restricted to the cells in the corresponding subpopulation. This approach enables to *simultaneously* compare the restricted VAFs of the same SNV in the left and right branches, providing further support for the estimated fraction of mutated copies in both branches.

In the next subsections, we first describe the expected values of VAF (Section 12.1) and restricted VAF (Section 12.2) for a SNV in the CHISEL tree. In particular, we do this by considering the occurrence of a whole-genome duplication (WGD) in the trunk of the tumor phylogeny, as the one identified in patient S0, and by distinguishing between *clonal* SNVs that are present in *all tumor* cells and *subclonal* SNVs that are only present in a *subset* of the tumor cells. At last (Section 12.3), we classify the 10 551 SNVs identified in the tumor clones according to the clonal status induced by the CHISEL tree.

12.1 Expected value of VAF

The value of VAF for a somatic SNV e in a genomic bin t depends on three factors: (1) tumor purity μ which corresponds to the proportion of tumor cells; (2) allele-specific copy numbers $\{\widehat{c}_{t,i}, \check{c}_{t,i}\}$ with total copy number $c_{t,i} = \widehat{c}_{t,i} + \check{c}_{t,i}$ in every cell i; (3) the mutated copy number $\dot{c}_{e,t,i}$ in every cell i. Tumor purity μ is a key factor when considering all cells in patient S0 because none of the copies of t in normal diploid cells harbor the SNV e and, therefore, $\dot{c}_{e,t,i} = 0, c_{t,i} = 2$ for every normal cell i. Specifically, we observe $\mu \approx 0.5$ across all cells in patient S0.

We first focus on a clonal SNV e located in a genomic bin t, which is thus present in all tumor cells. We start by assuming that t is affected by the WGD (as all the bins) but no further CNAs; therefore we know that t has allele-specific copy numbers $\{\widehat{c}_{t,i}, \check{c}_{t,i}\} = \{2, 2\}$ and total copy number $c_{t,i} = 4$ in every tumor cell i. Since the mutated copy number $\dot{c}_{e,t,i}$ only depends in this case on the occurrence of e either before or after the WGD, we can estimate the value of VAF by Eq. (51) in the two distinct cases. When the SNV e occurs before the WGD, we expect that $\dot{c}_{e,t,i} = 2$ in every tumor cell i because the mutated copy number is doubled as well as all the copies of t; the expected value of the corresponding VAF is thus ≈ 0.33 because the expected value of the fraction in Eq. (51) is $\approx \frac{0.5 \cdot 2}{0.5 \cdot 2 + 0.5 \cdot 4}$. When the SNV e occurs after the WGD, $\dot{c}_{e,t,i} = 1$ in every tumor cell i because e affects only a single copy over the 4 total copies resulting from the WGD and the expected value of the corresponding VAF is thus ≈ 0.167 because the expected value of the fraction in Eq. (51) is $\approx \frac{0.5}{0.5 \cdot 2 + 0.5 \cdot 4}$. These expected values are correct for all genomic bins that are not affected by further CNAs besides the WGD, which are the majority across the genome of patient S0. However, the expected value of VAF is affected by CNAs. In fact, CNAs result in genomic regions with different allele-specific copy numbers, as observed in both patients S0 and S1. As such, a genomic bin t with total copy number $c_{t,i} > 4$ in all or some cells may result in an expected value of VAF lower than ≈ 0.167 as well as a bin t with total copy number $c_{t,i} < 4$ may have an expected value of VAF higher than ≈ 0.33. Note that a complete investigation of all the possible combinations of allele-specific and mutated copy numbers is beyond the scope of this study.

We next focus on a subclonal SNV e located in a genomic bin t, which is thus present only in a subset of the tumor cells. When considering a genomic bin t not affected by further CNAs besides the WGD with allele-specific copy numbers $\{\widehat{c}_{t,i}, \check{c}_{t,i}\} = \{2, 2\}$ and total copy number $c_{t,i} = 4$ in every tumor cell i, the expected value of VAF of a subclonal SNV e is $\leqslant 0.167$ because any subclonal SNV must have occurred after WGD and $\dot{c}_{e,t,i} > 0$ only in a subpopulation of the tumor cells. The expected value of VAF for subclonal SNVs is affected by potential CNAs similarly to the previous case of clonal SNVs. Moreover, we expect that subclonal SNVs with higher VAF have occurred earlier in the tumor evolution than subclonal SNVs with lower VAF, since these SNVs are generally present in more cells.

12.2 Expected value of restricted VAF.

The value of restricted VAF of a somatic SNV e in a genomic bin t depends on only two factors, differently than the previous VAF: (1) allele-specific copy numbers $\{\widehat{c}_{t,i}, \widecheck{c}_{t,i}\}$ with total copy number $c_{t,i} = \widehat{c}_{t,i} + \widecheck{c}_{t,i}$ in every cell i of the restricted subpopulation of cells; (2) the mutated copy number $\dot{c}_{e,t,i}$ in every cell i of the restricted subpopulation of cells. We first focus on a clonal SNV that is present in all tumor cells and, consequently, is present in all cells of the restricted subpopulation. Similarly to the previous case of VAF, we can estimate the expected value of restricted VAF for a SNV e located in a genomic bin t not affected by further CNAs besides the WGD, with allele-specific copy numbers $\{\widehat{c}_{t,i}, \widecheck{c}_{t,i}\} = \{2, 2\}$ and total copy number $c_{t,i} = 4$ in every tumor cell i. We do this according to the occurrence of e either before or after the WGD identified in patient S0. When the clonal SNV e occurs before or after WGD, we expect that $\dot{c}_{e,t,i} = 2$ or $\dot{c}_{e,t,i} = 1$, respectively, in every cell within the restricted subpopulation and the expected value of the corresponding restricted VAF is thus ≈ 0.5 or ≈ 0.25. Note that these values are different than the corresponding values of VAF because here we restricted to a subpopulation of tumor cells with no admixture of normal cells. Moreover, since clonal SNVs that are not affected by further CNAs besides the WGD are equally present in all tumor cells, we expect the *same* values of restricted VAF for the same SNV e when restricting to either the left or right branch of the CHISEL tree.

The expected value of VAF for clonal SNVs is affected by CNAs that result in genomic regions with different allele-specific copy numbers, as in the case of VAF. In this study we specifically focused on the genomic bins of chromosome 2 where CHISEL identified a mirrored-subclonal CNA distinguishing the two branches: the clones of the left branch in the CHISEL tree have haplotype-specific copy numbers $(1, 2)$, while all clones but a small one of the right branch have haplotype-specific copy numbers $(2, 1)$. As the allele-specific copy numbers are $\{2, 1\}$ in both branches, we can estimate the expected values of the restricted VAFs for a clonal SNV e occurring before the WGD, distinguishing when e is located on the haplotype of either the deleted or retained allele. When e is located on the haplotype of the retained allele, we expect $\dot{c}_{e,t,i} = 2$ and $c_{t,i} = 3$ because the mutated copy number has been doubled by the WGD but the deletion did not affect the copies harboring e; therefore, the expected value of the corresponding VAF is ≈ 0.66. When e is located on the haplotype of the deleted allele, we expect $\dot{c}_{e,t,i} = 1$ and $c_{t,i} = 3$ because the mutated copy number has been doubled by the WGD and the mirrored-subclonal CNA deletes one of these copies; therefore, the expected value of the corresponding VAF is ≈ 0.33. We observed the presence of SNVs with restricted VAF equal to ≈ 0.33 in the left branch and equal to ≈ 0.66 in the right branch, or vice versa. These values are compatible with SNVs located on different haplotypes and the different values of the restricted VAFs in the two branches confirm the mirrored deletions of two distinct haplotypes.

Last, we investigate the expected values of the restricted VAF for subclonal SNVs which are not present in all tumor cells and are unique to cells in either the left or right branch of the CHISEL tree. Following the same previous observations on VAF, we expect the values of restricted VAF to be $\leqslant 0.25$ for subclonal SNVs in regions not affected by further CNAs besides the WGD because any subclonal SNV must have occurred after the WGD. However, these values may be different when affected by CNAs. For example, in the specific case of chromosome 2 with allele-specific-copy numbers $\{2, 1\}$ in each branch, we expect the values of restricted VAF to be $\leqslant 0.33$ because subclonal SNVs occurred after the WGD and affected only one copy of t, resulting in $\dot{c}_{e,t,i} = 1$ and $c_{t,i} = 3$. As mentioned before, a complete investigation of all the possible combinations of allele-specific and mutated copy numbers is beyond the scope of this study, however we can separate subclonal SNVs that occurred early or late in the evolution of the cells in the restricted subpopulation, i.e. in either the left or right branch of the CHISEL tree. For example, in the specific case of chromosome 2, we can distinguish *early* SNVs occurred in the evolutionary history of each branch as we expect that these SNV have restricted VAF≈ 0.33 because they are present in nearly all cells in the corresponding branch. In contrast, only subpopulation of cells within the branch harbor later SNVs, resulting in restricted VAF$\leqslant 0.33$.

12.3 Classification of SNVs based on a copy-number tree

We classified the SNVs according to the corresponding clonal status induced by the CHISEL tree. More specifically, we separated all the 10 551 SNVs identified in the tumor clones into *clonal* SNVs, which are present in all tumor clones, and *subclonal* SNVs, which are *unique* to either the left or right branch of the CHISEL tree. Unfortunately, the classification is complicated by two main issues. First, the rate of false positive identified by Varscan 2 is higher than expected due to the low-frequency of several mutations and due to the relaxed filters (Supplementary Methods 11). Second, we do not observe all the SNVs that are present in each clone for two reasons: (1) small clones with a low number of cells also have a low number of sequencing reads; (2) the variant frequencies of the SNVs are low due to the occurrence of WGD as the fraction of copies harboring a mutation is generally lower, e.g. a SNV occurred after WGD affects only 1 copy over 4 in genomic regions with no further CNAs besides the WGD. Therefore, we distinguished true and false SNVs by identifying SNVs with high or low prevalence, respectively, across the clones of the corresponding branch. In particular, we selected high prevalence SNVs that are present in at least a minimum number of clones which is chosen proportionally to the total number of clones in the corresponding branch: 3 for clonal SNVs, 2 for SNVs unique to the right branch of the CHISEL tree, and 1 for the SNVs unique to the left branch of the CHISEL tree. We thus identified 2 798 clonal SNVs present in both branches, 1 632 SNVs unique to the right branch, and 594 SNVs unique to the left branch. The remaining low-prevalence SNVs have both low VAFs (Supplementary Fig. 17) and low restricted VAFs in both branches, underscoring the low confidence in these mutation calls.

Supplementary References

[1] Kim, C. *et al.* Chemoresistance evolution in triple-negative breast cancer delineated by single-cell sequencing. *Cell* **173**, 879–893 (2018).

[2] Dentro, S. C., Wedge, D. C. & Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor perspectives in medicine* **7**, a026625 (2017).

[3] 10X Genomics. Assessing tumor heterogeneity with single cell cnv. `https://www.10xgenomics.com/solutions/single-cell-cnv`.

[4] 10X Genomics. What is cell ranger dna? `https://support.10xgenomics.com/single-cell-dna/software/pipelines/latest/what-is-cell-ranger-dna`.

[5] Andor, N. *et al.* Joint single cell dna-seq and rna-seq of gastric cancer reveals subclonal signatures of genomic instability and gene expression. *bioRxiv* `https://doi.org/10.1101/445932` (2018).

[6] Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nature methods* **12**, 1058 (2015).

[7] Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

[8] Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proceedings of the National Academy of Sciences* **108**, E1128–E1136 (2011).

[9] Oesper, L., Mahmoody, A. & Raphael, B. J. Theta: inferring intra-tumor heterogeneity from high-throughput dna sequencing data. *Genome biology* **14**, R80 (2013).

[10] Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47 (2016).

[11] Ha, G. *et al.* Titan: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research* **24**, 1881–1893 (2014).

[12] Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nature methods* **14**, 167 (2017).

[13] Laks, E. *et al.* Clonal decomposition and dna replication states defined by scaled single-cell genome sequencing. *Cell* **179**, 1207–1221 (2019).

[14] Sathirapongsasuti, J. F. *et al.* Exome sequencing-based copy-number variation and loss of heterozygosity detection: Exomecnv. *Bioinformatics* **27**, 2648–2654 (2011).

[15] Li, Y. & Xie, X. Deconvolving tumor purity and ploidy by integrating copy number alterations and loss of heterozygosity. *Bioinformatics* **30**, 2121–2129 (2014).

[16] Fischer, A., Vázquez-García, I., Illingworth, C. J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell reports* **7**, 1740–1752 (2014).

[17] Amarasinghe, K. C. *et al.* Inferring copy number and genotype in tumour exome data. *BMC genomics* **15**, 732 (2014).

[18] Chen, H., Bell, J. M., Zavala, N. A., Ji, H. P. & Zhang, N. R. Allele-specific copy number profiling by next-generation dna sequencing. *Nucleic acids research* **43**, e23–e23 (2014).

[19] Li, Y. & Xie, X. Mixclone: a mixture model for inferring tumor subclonal populations. *BMC genomics* **16**, S1 (2015).

[20] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

[21] Choi, Y., Chan, A. P., Kirkness, E., Telenti, A. & Schork, N. J. Comparison of phasing strategies for whole human genomes. *PLOS Genetics* **14**, e1007308 (2018).

[22] Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134 (2013).

[23] Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nature biotechnology* **26**, 897 (2008).

[24] Staaf, J. *et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome snp arrays. *Genome Biology* **9**, R136 (2008).

[25] Nilsen, G. *et al.* Copynumber: efficient algorithms for single-and multi-track copy number segmentation. *BMC genomics* **13**, 591 (2012).

[26] Wang, L. *et al.* Novel somatic and germline mutations in intracranial germ cell tumours. *Nature* **511**, 241 (2014).

[27] Carter, S. L., Meyerson, M. & Getz, G. Accurate estimation of homologue-specific dna concentration-ratios in cancer samples allows long-range haplotyping. *Nature Precedings* `https://doi.org/10.1038/npre.2011.6494.1` (2011).

[28] Cheng, Y. *et al.* Quantification of multiple tumor clones using gene array and sequencing data. *The annals of applied statistics* **11**, 967 (2017).

[29] Greenman, C. D. *et al.* Picnic: an algorithm to predict absolute allelic copy number variation with microarray cancer data. *Biostatistics* **11**, 164–175 (2009).

[30] Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences* **107**, 16910–16915 (2010).

[31] Carter, S. L. *et al.* Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology* **30**, 413 (2012).

[32] Shen, R. & Seshan, V. E. Facets: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput dna sequencing. *Nucleic acids research* **44**, e131–e131 (2016).

[33] Zaccaria, S. & Raphael, B. J. Accurate quantification of copy-number aberrations and whole-genome duplications in multi-sample tumor sequencing data. *bioRxiv* `https://doi.org/10.1101/496174` (2018).

[34] McPherson, A. W. *et al.* ReMixT: clone-specific genomic structure estimation in cancer. *Genome biology* **18**, 140 (2017).

[35] Bakker, B. *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome biology* **17**, 115 (2016).

[36] Wang, X., Chen, H. & Zhang, N. R. Dna copy number profiling using single-cell sequencing. *Briefings in bioinformatics* **19**, 731–736 (2017).

[37] Dong, X., Zhang, L., Hao, X., Wang, T. & Vijg, J. SCCNV: a software tool for identifying copy number variation from single-cell whole-genome sequencing. *bioRxiv* `https://doi.org/10.1101/535807` (2019).

[38] Wang, R., Lin, D.-Y. & Jiang, Y. SCOPE: a normalization and copy number estimation method for single-cell dna sequencing. *bioRxiv* `https://doi.org/10.1101/594267` (2019).

[39] Lander, E. S. & Waterman, M. S. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**, 231–239 (1988).

[40] Schwarz, R. F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLOS Computational Biology* **10**, 1–11 (2014).

[41] El-Kebir, M. *et al.* Complexity and algorithms for copy-number evolution problems. *Algorithms for Molecular Biology* **12**, 13 (2017).

[42] Zaccaria, S., El-Kebir, M., Klau, G. W. & Raphael, B. J. Phylogenetic copy-number factorization of multiple tumor samples. *Journal of Computational Biology* **25**, 689–708 (2018).

[43] Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).

[44] Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17**, 261–272 (2020).

[45] Jain, A. K., Dubes, R. C. *et al.* *Algorithms for clustering data*, vol. 6 (Prentice hall Englewood Cliffs, NJ, 1988).

[46] Koboldt, D. C. *et al.* Varscan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).

[47] Koboldt, D. C. *et al.* Varscan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568–576 (2012).

[48] DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation dna sequencing data. *Nature genetics* **43**, 491 (2011).

[49] Roth, A. *et al.* Pyclone: statistical inference of clonal population structure in cancer. *Nature methods* **11**, 396 (2014).