Alleloscope: Integrative single cell analysis of allele-specific copy number alterations and chromatin accessibility in cancer

Chi-Yun Wu[1,2], Billy T. Lau[3,4], Heon Seok Kim[3], Anuja Sathe[3], Susan M. Grimes[4], [*]Hanlee P. Ji[3,4], [*]Nancy R. Zhang[1,2]

Institutions

[1]Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA

[2]Department of Statistics, University of Pennsylvania, Philadelphia, PA

[3]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA

[4]Stanford Genome Technology Center, Stanford University, Palo Alto, CA

[*]Corresponding author

Corresponding authors

Hanlee P. Ji

Email: genomics_ji@stanford.edu

Nancy R. Zhang

Email: nzh@wharton.upenn.edu

Abstract

Cancer progression is driven by both somatic copy number aberrations (CNAs) and chromatin remodeling, yet little is known about the interplay between these two classes of events in shaping the clonal diversity of cancers. We present Alleloscope, a method for allele-specific copy number estimation that can be applied to single cell DNA and ATAC sequencing data, either separately or in combination. This approach allows for integrative multi-omic analysis of allele-specific copy number and chromatin accessibility on the same cell. On scDNA-seq data from gastric, colorectal, and breast cancer samples, with extensive validation using matched linked-read sequencing, Alleloscope finds pervasive occurrence of highly complex, multi-allelic copy number aberrations, where cells that carry varying allelic configurations adding to the same total copy number co-evolve within a tumor. The contributions of such allele-specific events to intratumor heterogeneity have been under-reported and under-studied due to the lack of methods for their detection. On scATAC-seq from two basal cell carcinoma samples and a gastric cancer cell line, Alleloscope detects multi-allelic copy number events and copy neutral loss-of-heterozygosity, enabling the dissection of the contributions of chromosomal instability and chromatin remodeling in tumor evolution.

36 **Introduction**

37 Cancer is a disease caused by genetic alterations and epigenetic modifications
38 which, in combination, shape the dysregulated transcriptional programming of tumor
39 cells[1, 2]. These somatic genomic events lead to a diverse cellular population from which
40 clones with advantageous alterations proliferate and eventually metastasize[3]. The
41 comprehensive study of cancer requires the integrative profiling of genetic and epigenetic
42 changes at the resolution of single cells. We combined the analysis of two such genomic
43 dimensions – DNA copy number and chromatin accessibility – through massively parallel
44 single cell sequencing assays.

45 First, consider copy number aberrations (CNAs), through which we have derived much of
46 our current understanding of the relationship between genome instability and tumor
47 evolution[4]. Total copy number profiling, which estimates the sum of the copy numbers of
48 the two homologous chromosomes, is inadequate to characterize some types of cancer
49 genomic aberrations. Such events include the pervasive copy-neutral losses of
50 heterozygosity (LOH)[5-8] and the intriguing "mirrored events"[9, 10], where cells carrying
51 amplification of one haplotype are intermingled with cells carrying amplification of the
52 other haplotype. While the importance of allele-specific copy number profiling has been
53 emphasized in bulk DNA sequencing analysis[5-8, 11], most single-cell studies profile only
54 total copy number due to the difficulty of reliable allele-specific estimation at low per-cell
55 coverage[12-19]. Recently, a major advance was made by Zaccaria and Raphael in the
56 development of CHISEL[10], a method for single-cell allele-specific copy number analysis.
57 By default, CHISEL relies on externally phased haplotypes derived from available
58 reference cohorts for accurate estimation of allele-specific CNAs, and found evidence for

59 the existence of mirrored amplifications in a deeply sequenced breast tumor. This was
60 the first time that allele-specificity of CNA events were reported at single-cell resolution,
61 but the requirements for high sequencing depth and external phasing limit the general
62 applicability of CHISEL. Thus, despite these recent advances, the prevalence of mirrored
63 amplification events and the general complexity of CNA regions at single-cell allele-
64 specific resolution are still largely unexplored.

65 Epigenetic modifications are also an important genomic feature of cancer. One type of
66 epigenetic profiling is the measurement of chromatin openness, for which a variety of
67 methods have been developed[20] including transposase-accessible chromatin sequencing
68 (ATAC-seq)[21, 22], which can be performed at the bulk-tissue or single-cell level. Analysis
69 of chromatin structure has shown that epigenetic remodeling modulates the plasticity of
70 cells in cancer[23-27], leads to stem-like properties[28-30] and generates therapeutic
71 resistance[31-34]. Since copy number alterations involve large gains and losses of chromatin,
72 we expect the chromatin accessibility of a region to be influenced by the changes in
73 underlying copy number. Current scATAC-seq studies estimate total copy number profiles
74 by smoothing the read coverage and comparing position-wise against a control cell
75 population[24, 26]. Yet this appropriate control is difficult to identify, as even between normal
76 cells of distinct lineages, broad chromatin remodeling can lead to chromosome-level shifts
77 in coverage (examples of this will be given in Results). There is yet no method for reliable
78 total or allele-specific copy number profiling in scATAC-seq data, and thus, the estimation
79 of CNA and chromatin accessibility are confounded in current analysis pipelines.

80 Addressing these challenges, we present Alleloscope, a method for **allel**e-**s**pecific **cop**y
81 number **e**stimation and multiomic profiling in single cells. Alleloscope does not rely on

external phasing information, and can be applied to low coverage scDNA-seq data or to scATAC-seq data with sample-matched bulk DNA sequencing data. To investigate the single cell landscape of allele-specific CNA, we first apply Alleloscope on scDNAseq data from four gastric cancer samples, four colorectal cancer samples, and two datasets from different regions of a breast cancer sample[10, 12, 35]. For five of the gastrointestinal cancer samples, results are extensively validated by 10x linked-read sequencing which provides accurate phasing information[36-38]. In these datasets, Alleloscope accurately identifies LOH and mirrored-subclonal amplification events, and furthermore, finds pervasive occurrence of highly complex multi-allelic loci where cells that carry varying allelic configurations adding to the same total copy number co-evolve within a tumor. The ubiquity of such events in all three cancer types analyzed reveal that they may be an important overlooked source of intratumor genetic heterogeneity.

Having characterized the complexity of allele-specific CNA events at single cell resolution, we turn to scATAC-seq data from two basal cell carcinoma samples with paired bulk whole exome sequencing data[26] and a complex polyclonal gastric cancer cell line that we also analyzed by scDNA-seq. In these samples, we evaluate the accuracy of Alleloscope in genotyping and clone assignment for scATAC-seq data and demonstrate its application to the integrative analysis of CNA and chromatin accessibility.

Results

Overview of Alleloscope model and algorithm

First, we briefly overview Alleloscope's method for allele-specific copy number estimation (Fig. 1). Clone assignment and integration with peak signals in scATAC-seq data will be

104 described later. Alleloscope relies on two types of data features: coverage, derived from
105 all reads that map to a given region, and allelic imbalance, derived from allele-informative
106 reads that cover heterozygous loci in the region. We start with some essential definitions.
107 For a given single nucleotide polymorphism (SNP) site, we refer to its mean coverage
108 across cells as *bulk coverage* and its mean variant allele frequency (VAF = ratio of
109 alternative allele read count to total read count) across cells as its *bulk VAF*. Between the
110 two parental haplotypes, we define the term "major haplotype" as the haplotype with
111 higher mean count across cells. Note that a haplotype may be the "major haplotype" of a
112 sample, but have lesser copy number than the other haplotype within some cells. For
113 each cell i, in any given CNA region r, we define two key parameters: (1) the major
114 haplotype proportion (θ_{ir}), defined as the count of the copies of the major haplotype
115 divided by the total copy number for the region in cell i and (2) total copy fold change
116 (ρ_{ir}), defined as the ratio of the total copy number of the region in cell i relative to that in
117 normal cells.

118 The algorithm starts by segmenting the genome into regions of homogeneous allele-
119 specific copy number using both the bulk coverage and bulk VAF profiles (Fig.1 Step 2).
120 This can be achieved by multiple existing algorithms, which may be combined to increase
121 detection sensitivity, see Methods for details. In our analyses of scATAC-seq data, the
122 segmentation was performed on the matched bulk or single cell DNA-seq data, which
123 ensures that the CNA regions passed to the characterization step are supported by
124 evidence from DNA and thus are not simply due to the broad chromatin remodeling that
125 occur in cancer.

126 Now consider each putative CNA region, indexed by r. An expectation-maximization (EM)
127 based algorithm is used to iteratively phase each SNP and estimate the major haplotype
128 proportion (θ_{ir}) for each cell (Fig. 1, step 3). For each SNP j lying in the region, let $I_j \in$
129 $\{0,1\}$ be the indicator of whether the reference allele of SNP j is carried by the major
130 haplotype (please see the toy example in Fig. 1). An initial estimate $\hat{I}_j^{(0)}$ is first derived
131 from the bulk VAF profile. Then, in iteration t, Alleloscope computes $\hat{\theta}_{ir}^{(t)}$ by pooling
132 counts across sites within the region r, weighted by the current phasing $\hat{I}_j^{(t)}$, then updates
133 the estimate of I_j based on $\hat{\theta}_{ir}^{(t)}$ by pooling counts across cells. The estimates of θ_{ir} and
134 I_j usually converge within a few iterations as described in the Methods. If matched
135 scDNA-seq data are available for a sample sequenced by scATAC-seq, I_j values can be
136 estimated from scDNA-seq and then used to compute θ_{ir} for each cell in the scATAC-
137 seq data, enabling more information sharing between the two data types.

138 The estimated major haplotype proportions ($\hat{\theta}_{ir}$'s), along with a preliminarily normalized
139 coverage statistic ($\tilde{\rho}_{ir}$'s, see Methods for details), are then used to identify a set of normal
140 cells and diploid regions (Fig. 1, Step 4). This information is then used to estimate an
141 improved relative coverage fold-change ($\hat{\rho}_{ir}$) for each cell within each CNA region. If cell
142 i's true allele-specific copy numbers are homogeneous within the given region r, then its
143 true value of (θ_{ir}, ρ_{ir}) should belong to a set of canonical points displayed in Step 5 of
144 Fig. 1. Thus, the estimated values $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ are clustered across cells and associated with
145 one of the canonical values to yield the cell-level haplotype profiles for the CNA region.
146 These cell- and region-specific haplotype profiles serve as the basis for clone assignment
147 and subsequent integration with peak signals in scATAC-seq data. (Fig. 4b).

148 As we will show next in benchmark experiments, the above method allows accurate CNA
149 state characterization in a wide range of scenarios. However, there is one scenario that
150 is extremely difficult: balanced mirrored subclones at small clonal frequencies. To
151 increase sensitivity for such events, Alleloscope allows an additional refinement step that
152 starts with the $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ estimates and uses thresholding on $\hat{\rho}_{ir}$ to improve the estimation
153 of $\hat{\theta}_{ir}$. Please see Supplementary Methods for details.

154 **Whole genome haplotypes validate Alleloscope in scDNA-seq allele-specific copy**
155 **number estimation**

156 First, we explore the single cell landscape of allele-specific CNAs using scDNA-seq data.
157 The scDNA-seq samples included in this study are given in Supplementary Table 1. To
158 benchmark the phasing and allele-specific copy number state estimation accuracy, we
159 performed matched linked-read whole-genome sequencing data on five gastrointestinal
160 tumor samples representing different levels of chromosomal instability: P5931, P6335,
161 P6198, P5915, and P6461. Linked-read sequencing, in which one derives reads from
162 individual high molecular weight DNA molecules, provides variants that can be phased
163 into extended haplotypes covering megabases in length[36-38]. To evaluate the accuracy of
164 phasing, we compared the haplotypes estimated by Alleloscope to the haplotypes
165 obtained from linked-read WGS. Additionally, we used the linked-read WGS haplotype
166 with Alleloscope to derive gold-standard allele-specific copy number state assignments
167 for each cell, which are used to assess the impact of phasing errors on accuracy in allele-
168 specific copy number state estimation (Fig. 2a). This allows us to compute the sensitivity
169 (recall rate of CNA regions detected in gold-standard) and specificity of a method, as
170 defined in Supplementary Methods.

171 We first use the gastric cancer sample P5931 to demonstrate Alleloscope and illustrate
172 this benchmark design. The heatmap in Fig. 2b gives a bird's-eye-view of Alleloscope's
173 allele-specific copy number estimates for this sample, showing it to have a relatively
174 simple genome with few CNAs, which is expected given its MSI subtype. However,
175 detailed inspection of the four chromosomes carrying clear CNA events, chr7, chr8, chr20,
176 and chr21, reveals complexity at the allelic level. For each event, the scatter plots of
177 $(\hat{\theta}_{ir}, \hat{\rho}_{ir})$ estimated by Alleloscope and colored by haplotype-specific CNA state
178 assignment are shown in Fig. 2c. Note that the $(\hat{\theta}_{ir}, \hat{\rho}_{ir})$ clusters fall almost directly on
179 top of the expected canonical values (e.g. (1/2, 1) for diploid, (2/3, 1.5) for 1 copy gain of
180 major haplotype). Interestingly, chromosomes 7, 8, and 21 each show subclonal clusters
181 with differing allelic ratios adding up to the same total copy number, which would not have
182 been detectable without allele-specific estimation. We denote the major haplotype of a
183 region by "M", and the minor haplotype by "m". The chromosome 7 amplification exhibits
184 two tumor subclones with mirrored-subclonal CNAs (MMm and Mmm), each subclone
185 amplifying a different haplotype. Such a mirrored-subclonal CNA configuration is also
186 observed for the deletion on chromosome 21 (M- and m-). The chromosome 8
187 amplification exhibits four tumor subclones differentiated by their haplotype profiles—
188 MMm, Mmm, MMmm, and MMMm.

189 Comparing against the whole genome haplotypes derived from linked-reads, the phasing
190 accuracy is 98% for the chr21 deletion, ~90% for the two clonal amplifications (on chr8
191 and chr20), and 79% for the subclonal chr7 amplification (shown in the titles of the scatter
192 plots of Fig. 2c). Fig. 2d shows scatterplots of $\hat{\rho}_{ir}$ against major haplotype proportion
193 computed using haplotypes derived from linked-read sequencing ($\tilde{\theta}_{ir}$), with cells colored

194 by their state assignment given by Alleloscope without using the haplotype information
195 (i.e. each cell has the same color as in Fig. 2c). Comparing the scatterplots in Fig. 2d to
196 their counterparts in Fig. 2c reveals that Alleloscope's estimated cell haplotype profiles
197 are highly concordant with those derived directly with the haplotypes from linked-read
198 WGS. Specifically, the concordance is ~100% across all four events (the concordance for
199 each event is labeled in the scatter plots of Fig. 2d). This shows that the allele-specific
200 copy number estimation by Alleloscope is robust to errors in phasing (e.g. for chr7). We
201 also used two extreme cases—chr2 and chr3 of P5931 to demonstrate the refinement
202 step (Supplementary Fig. 1). For these two regions, linked-read sequencing revealed a
203 mirrored amplification event distinguishing two subclones each comprised of <10 out of
204 700 cells. The differing major haplotype proportions of these two subclones, at $\frac{1}{3}$ and $\frac{2}{3}$,
205 was missed by the original estimation, but accurately detected with the refinement step
206 of Alleloscope.

207 So far, CHISEL is the only other comparable method for allele-specific copy number
208 estimation with scDNA-seq data. Thus, we benchmarked Alleloscope against CHISEL by
209 comparing each method's estimates under default settings to the same method's
210 estimates using the true haplotype phasing information provided by linked-read WGS.
211 Thus, each method's performance is assessed using its own "gold standard" computed
212 with linked-read WGS haplotypes. We verified that CHISEL's gold standard is highly
213 similar to Alleloscope's gold standard, and thus, with the help of linked-read WGS
214 haplotypes both methods can achieve accurate estimation that is close to the truth. The
215 sensitivity and specificity of each method on the five samples are given in Table 1. For
216 CHISEL, there is an extra "correction step" where the inferred clones are used to generate

217 consensus allele-specific copy number profiles for all cells within each clone. We show
218 CHISEL's results with and without this correction step, as for most samples, it increases
219 sensitivity (but decreases specificity), but for (P5931) it makes the results worse. The
220 results obtained using different external phasing datasets and different block sizes for
221 CHISEL also suggest that CHISEL's accuracy depends on choices for these inputs
222 (Supplementary Table 2). Detailed heatmaps comparing Alleloscope and CHISEL using
223 P5931 as an example are given in Supplementary Fig. 2-3, with more discussion given in
224 Supplementary Results. Overall, Alleloscope maintains high accuracy (both sensitivity
225 and specificity) across the samples analyzed, substantially improving upon the current
226 state-of-the-art.

227 **Further performance assessment via data downsampling and simulations**

228 We also compared Alleloscope and CHISEL on the breast cancer sample that was
229 analyzed in detail by Zaccaria et al[10]. This sample does not have true phasing information,
230 but was sequenced at much higher coverage, and thus, we downsampled the reads to
231 50% and 25% of the original depth and compared the estimates obtained from the
232 downsampled datasets to those obtained from the original dataset. Heatmaps of the
233 genome-wide profiles estimated by CHISEL and Alleloscope on the original dataset, the
234 50% subsampled dataset and the 25% subsampled dataset are shown in Supplementary
235 Fig. 4. For CHISEL, the clone-corrected outputs are shown since they are more accurate
236 than the uncorrected ones. The accuracy of Alleloscope and CHISEL are shown in Table
237 1. We see that at 25% of the original depth, Alleloscope has a specificity of 82% as
238 compared to CHISEL's 65%, and sensitivity of 95% as compared to CHISEL's 64%.

239 Finally, we used simulations to investigate the performance of Alleloscope over a grid of
240 experimental parameters (number of cells, total per cell coverage, and total coverage at
241 heterozygous SNP sites, Supplementary Fig. 5). Across experimental design settings,
242 Alleloscope attains higher accuracy for deletions than amplifications, which is expected
243 due to the larger change in both coverage ratio and major haplotype proportion ($\hat{\theta}$) that
244 accompanies deletions. Phasing for deletions (and other LOH events) is accurate across
245 the board, but for amplifications it is low at 10% and 5% clonal frequencies. However, the
246 CNA state assignment accuracy is, to a certain extent, robust to phasing error, increasing
247 steadily with coverage, number of heterozygous SNPs, and number of cells.

248 **Alleloscope finds pervasive occurrence of polyclonal CNA regions differentiated**
249 **by haplotype ratios**

250 We start by considering copy neutral LOH events, which are well-known drivers of cancer
251 evolution that can only be identified through allele-specific copy number analysis. We use
252 the colorectal adenocarcinoma sample P6198 as illustration. The copy number heatmap
253 in cellranger (Fig. 3a) reveals that most of the genome is amplified for this sample, but
254 with few regions of normal copy number. One of these regions, chromosome 5, is shown
255 by cellranger to be copy neutral for its entire length. The bulk VAF clearly separates this
256 chromosome into two main regions, a normal region followed by a copy-neutral LOH (Fig.
257 3b). Concordantly, Alleloscope reveals a cluster centered at $(\rho, \theta) = (1,1)$ corresponding
258 to copy-neutral LOH only for the region on the right (Fig. 3b). Similarly, alleloscope also
259 revealed copy neutral LOH events in this sample on chr11 and chr16. Using haplotypes
260 from matched linked-read WGS, all of these LOH events are validated and the per-cell
261 LOH state assignment accuracy of Alleloscope is nearly 100% (Supplementary Fig. 6).

262 Next, we survey the overall landscape of single cell allele-specific copy number profiles
263 among the samples analyzed. Detailed segmentation plots and heatmaps for all samples,
264 listed in Supplementary Table 1, are in Supplementary Fig. 7-13. Across most of the
265 samples, we observe a high prevalence of complex subclonal CNAs indicated by multiple
266 clusters differentiated by allelic ratios for the same genomic region. Prototypical examples
267 from P6198, SNU601, P6335 and BC10X are shown in Fig. 3c. In some regions, such as
268 chromosome 9 of SNU601, 3q of SNU601, 2q and 16p of BC10x, we see as many as
269 seven subclonal clusters. In many cases there are multiple clusters corresponding to the
270 same total copy number but varying in allelic dosage. Minor subclones carrying deletion
271 of one haplotype can be easily masked by dominant subclones carrying amplifications of
272 the other haplotype in a bulk analysis or a single cell analysis that relies on total coverage.
273 Overall, the high subclonal diversity in these genomic regions reveal an aspect of
274 intratumor heterogeneity that was previously underreported.

275 Recurrent chromosomal instability events, affecting both haplotypes and producing
276 gradients in haplotype dosage, is a common theme across all samples analyzed.
277 Consider, for example, the region on chromosome 9 of P6198 (Fig. 3c), which reveals 7
278 subpopulations of cells: besides the normal cell cluster and the dominant tumor cell
279 cluster with the haplotype profile MMm, there is a small cluster of cells with copy neutral
280 LOH, two small subclones at four chromosome copies and two more at five chromosome
281 copies. This produces major haplotype ratios of $\{\frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{4}{5}, 1\}$ in different cells. Another
282 example of such complexity is chromosome 3q of SNU 601 and chromosome 2q of BC10x,
283 which share a similar pattern: two mirrored-subclonal CNAs (MMm, mmM) at total copy
284 number of 3, as well as mirrored-subclonal CNAs (MMMm, MMmm, mMMM) at total copy

285 number of 4, producing a gradient of haplotype ratios $\{\frac{1}{4},\frac{1}{3},\frac{1}{2},\frac{2}{3},\frac{3}{4}\}$. Interrogating the
286 evolutionary route by which such diversity was achieved, Alleloscope reveals that a whole
287 genome doubling event is likely to have taken place early in the development of BC10x
288 and P6198, but not in the development of SNU601 (see Supplementary Fig. 4&6). Thus,
289 the subclones at 2q in BC10x and 3q in SNU601 must have evolved through different
290 evolutionary routes: In BC10x, the early whole-genome doubling produced the cluster
291 MMmm, from which the other clusters of different haplotype profiles were most likely
292 derived through successive loss and gene conversion events. On the contrary, the
293 clusters on 3q of SNU601 were most likely a result of successive amplification events
294 starting from the normal haplotype profile Mm. The observation of such similar clonal
295 configurations in allele-specific copy number, in two samples coming from different cancer
296 types (breast and gastric), suggests that convergent evolution or similar evolutionary
297 pressures, where such haplotype dosage gradients serve as an important substrate for
298 selection in tumor evolution, may be at play. Large studies involving more samples would
299 be needed to characterize the evolutionary significance of such complex clonal patterns.

300 **Juxtaposition of single cell copy number and chromatin remodeling events by**
301 **integrative scATAC-seq analysis**

302 To demonstrate the integrative analysis of scATAC-seq data with Alleloscope, we first
303 consider two basal cell carcinoma samples with matched whole-exome sequencing (WES)
304 data[39]. Using the matched WES data, the genome of each sample was first segmented
305 into regions of homogeneous bulk copy number (Fig. 4a, middle panel shows the
306 segmentation for SU008). Alleloscope was then applied to the scATAC-seq data to derive

307 allele-specific copy number estimates of each cell in each region. Scatterplots of $(\hat{\rho}, \hat{\theta})$
308 for five example CNA regions and one control region (chr12) from SU008 are shown in
309 Fig. 4a. For this sample, peak profiles characterizing chromatin accessibility separated
310 the cells confidently into three main clusters: 308 tumor cells, 259 fibroblasts and 218
311 endothelial cells with the cell type identity retrieved from the original study. Since the
312 normal fibroblast and epithelial cells are not expected to carry broad copy number events,
313 they provide a "null scenario" for comparison to the tumor cells. Density contours for each
314 cell type are shown in the $(\hat{\rho}, \hat{\theta})$- scatterplots (Fig. 4a). The $(\hat{\rho}, \hat{\theta})$ values clearly separate
315 the tumor cells from the normal cells for each CNA region, with the tumor cell cluster
316 positioned at canonical points in each case, indicating that these statistics used by
317 Alleloscope can accurately distinguish amplifications and loss-of-heterozygosity events
318 in scATAC-seq data. In particular, Alleloscope differentiated the cells that carry the
319 putative copy neutral LOH events in regions 4a, 6b, and 15b through shifts in major
320 haplotype proportion. Note that normal cells, which are not expected to carry broad
321 chromosome-scale CNAs, exhibit chromosome-level deviations in total coverage due to
322 broad chromatin remodeling but with no significant difference in their $\hat{\theta}_{ir}$ distribution, as
323 exemplified by the chr6b region. The coverage in large genomic bins across the genome
324 was further investigated for the two normal cell types in scATAC-seq data (Supplementary
325 Fig. 14; Supplementary Methods). Comparing one normal cell type against the other,
326 many regions exhibit broad shifts in relative coverage. Thus, relying solely on shifts in
327 coverage, without complementary shifts in major haplotype proportion, would lead to false
328 positive copy number detections for scATAC-seq data.

329 By assigning allele-specific CNA profiles to single cells in scATAC-seq data, Alleloscope
330 allows the integrative analysis of chromosomal instability and chromatin remodeling as
331 follows (Fig. 4b): The scATAC-seq data, paired with bulk or single-cell DNA sequencing
332 data, allows us to detect subclones. In parallel, a peak-by-cell matrix can be computed
333 following standard pipelines. Then, the subclone memberships or CNA profiles can be
334 visualized on the low-dimensional embedding of the peak matrix, and the subclones can
335 be further compared in terms of peak or transcription factor motif enrichment. Precise
336 haplotype profiles for each subclone then allow us to identify significantly
337 enriched/depleted peaks after accounting for copy number differences, thus delineating
338 events that are uniquely attributable to chromatin remodeling.

339 Hierarchical clustering using major haplotype proportion $\hat{\theta}$ identifies the tumor cells from
340 the normal cells for both SU006 (Supplementary Fig. 15) and SU008, and clearly
341 delineates a subclone in SU008 marked by a copy-neutral LOH event on chr4a (Fig. 4c).
342 Focusing on SU008, we call the cell lineage that carries the chr4a LOH event clone-2,
343 and the other lineage clone-1. In parallel, clustering by peaks cleanly separates the tumor
344 cells from the epithelial cells and fibroblasts (Fig. 4d: left), and further, demarcates two
345 distinct clusters in the tumor cells (peaks-1 and peaks-2) (Fig. 4d: middle). What is the
346 relationship between the peaks-1 and peaks-2 clusters obtained from peak signals to the
347 two clones delineated by chr4a LOH? Coloring by chr4a major haplotype proportion ($\hat{\theta}$)
348 on the peaks-derived UMAP shows that the LOH in this region is carried by almost all of
349 the cells in peaks-2 but only a subset of the cells in peaks-1 (Fig. 4d: middle). This can
350 also be clearly seen in the density of $\hat{\theta}$ (Fig. 4d: right): While $\hat{\theta}$ is heavily concentrated
351 near 1 for peaks-2, it is bimodal for peaks-1. Since clone-1 and clone-2 are differentiated

352 by a copy-neutral event, this separation by peaks into two clusters is not driven by broad
353 differences in total copy number. Since clone-2 is split into two groups of distinct peak
354 signals, we infer that the chromatin remodeling underlying the divergence of the peaks-2
355 cells is likely to have occurred in the clone-2 lineage, after the chr4a LOH event (Fig. 4e).
356 In this way, Alleloscope analysis of this scATAC-seq data set allowed us to overlay two
357 subpopulations defined by peak signals with two subpopulations defined by a subclonal
358 copy-neutral LOH, and infer their potential temporal order.

359 **Integrative analysis of clonal evolution and altered chromatin accessibility for a**
360 **complex polyclonal gastric cancer cell line**

361 The gastric cancer cell line SNU601 exhibits complex subclonal structure, as evidenced
362 by multiple multiallelic CNA regions (chr3e and chr18a are shown in Fig. 3c). In addition
363 to scDNA-seq, we also performed scATAC-seq on this sample to profile the chromatin
364 accessibility of 3,515 cells at mean coverage of 73,845 fragments per cell. This allows us
365 to compare the allele-specific copy number profiles obtained by scATAC-seq with those
366 given by scDNA-seq and integrate the two data types in a multi-omic characterization of
367 this complex tumor.

368 First, we segmented the genome and estimated the allele-specific copy number profiles
369 of single cells at each segment for both the scATAC-seq and scDNA-seq data, following
370 the procedure in Fig. 1 with some modifications due to the lack of normal cells to use as
371 control for this sample (see methods). Fig. 5a shows the relative total coverage, pooled
372 across cells from scDNA-seq. Fig. 5b shows $(\hat{\rho}, \hat{\theta})$-scatterplots for five example CNA
373 regions in scDNA-seq and scATAC-seq. Compared to the scATAC-seq data, the scDNA-

374 seq data has about 8-fold higher total read coverage and 7-fold higher heterozygous site
375 coverage per cell. Thus, while subclones corresponding to distinct haplotype profiles are
376 cleanly separated in the scDNA-seq data, they are much more diffuse in the scATAC-seq
377 data. Yet, cluster positions in scATAC-seq roughly match those in scDNA-seq. As
378 expected, the $(\hat{\rho}, \hat{\theta})$-scatterplots reveal the high level of chromosomal instability in this
379 sample, with each region exhibiting multiple clusters of different haplotype structures that
380 indicate the existence of subclones carrying mirrored events and, for some regions, the
381 variation of haplotype dosage over a gradient across cells.

382 Fig. 5c shows the hierarchical clustering of cells from scDNA-seq based on their allele-
383 specific copy number profiles, revealing the subclonal structure and the co-segregating
384 CNA events that mark each subclone. For each cell in each region, Alleloscope also
385 produces a confidence score for its assignment to different haplotype profiles
386 (Supplementary Fig. 16). Based on visual examination of the confidence scores at the
387 marker regions in both the ATAC and DNA sequencing data sets, we identified 6
388 subclones for further investigation (Clones 1-6 labeled at the right of the heatmap). The
389 allele-specific copy number profiles allow us to manually reconstruct the probable
390 evolutionary tree relating these 6 clones under the following three rules:

391 (1) Parsimony: The tree with the least number of copy number events is preferred.

392 (2) Monotonicity: For a multi-allelic region with escalating amplifications (e.g. Mm, MMm,
393 MMMm), the haplotype structures were produced in a monotonic order (e.g. Mm→
394 MMm→ MMMm) unless a genome doubling event occurred.

395 (3) Irreversibility of LOH: Once a cell completely loses an allele (i.e. copy number of that
396 allele becomes 0), it can no longer gain it back.

397 The evolutionary tree, thus derived, is shown in Fig. 6b. The mirrored-subclonal
398 amplifications on chr3q, the deletion on chr4p, and the multiallelic amplification on chr20q
399 allowed us to infer the early separation of clones 3-6 from clones 1-2. Subclones 3-6 are
400 confidently delineated by further amplifications on chr3q, chr20q, chr11, chr13, and chr17.
401 Note that high chromosomal instability led to concurrent gains of 1q and 7p in both the
402 Clone 1-2 and Clone 3-6 lineages. We also observed a large number of low-frequency
403 but high-confidence CNA events indicating that ongoing chromosomal instability in this
404 population is spawning new sporadic subclones that have not had the chance to expand.

405 We now turn to scATAC-seq data, focusing on the 10 marker regions which, together,
406 distinguish Clones 1-6: chr1b, 3b-d, 4b, 7a, 11b, 13b, and 20b-c. The $(\hat{\rho}, \hat{\theta})$ values
407 computed by Alleloscope allows us to directly assign allele-specific copy number profiles
408 to each cell for each region, as well as subclone labels to each cell, with posterior
409 confidence score. The subclone assignment utilizes a Bayesian mixture model that pools
410 information across the 10 marker regions. Despite the low accuracy in per-region
411 genotyping, when information is pooled across the 10 marker regions, 81.6% of the 2,753
412 cells after filtering can be assigned to a subclone with >95% posterior confidence
413 (Supplementary Fig. 17, the number of ATAC cells confidently assigned to each clone
414 are shown in Fig. 6a.). These subclone assignments for each cell, and cell-level haplotype
415 profiles for each region, can now be integrated with peak-level signals.

416 Following the scheme in Fig. 4b, we computed the Uniform Manifold Approximation and
417 Projection (UMAP) coordinates for the scATAC-seq cells based on their peak profiles,
418 which enables us to visualize the difference in chromatin accessibility profiles among
419 individual cells (Fig. 6a). UMAP scatterplots colored by clone assignment show that the
420 6 clones exhibit marked differences in their chromatin accessibility profiles (Fig. 6a): While
421 Clone 1 and Clone 2 are concentrated at the top half of the UMAP, Clones 3-5 are
422 positioned almost exclusively at the bottom half. Clone 6, which exhibits more variance,
423 is also significantly enriched at the bottom half of the UMAP. Among Clones 3-5, Clone 3
424 has a distinct chromatin accessibility profile that is mostly concentrated at the bottom tip,
425 Clone 4 is positioned higher, while Clone 5 contains cells that are similar to both clones
426 3 and 4. We expect some of these peak-level differences to be driven by CNAs.

427 To delineate the peaks that differ between clones, and to distinguish peak differences
428 that are not accountable by CNAs, we developed a statistical test to adjust for underlying
429 copy numbers to identify differentially accessibility peaks (DAPs) between the cell
430 populations on the two branches (Fig. 6b). The test is based on a generalized log-
431 likelihood ratio test with specific adjustment terms for the copy number profile of each of
432 the two subclones (Methods). The same test, but neglecting coverage adjustment, was
433 also performed for comparison. Along each branch, three numbers are shown for DAPs—
434 1. DAPs identified only before adjustment; 2. DAPs identified both before and after
435 adjustment; 3. DAPs identified only after adjustment (Fig. 6b). For the smaller subclones
436 (Clone 3,4,5), low coverage limits the detection power and thus limits the DAP counts in
437 both categories. Yet, juxtaposing DAP and CNA events along the tumor phylogeny yields
438 insights: Along most lineages, adjustment for CNA-induced broad shifts in coverage

439 removes a large number of peaks whose differential-accessibility signal is explainable by
440 the underlying change in copy numbers. Interestingly, CNA adjustment also identifies new
441 DAPs whose signals were obscured by a change in copy number in the opposite direction.
442 This argues for the importance of CNAs as a mechanism underlying subclonal differences
443 in chromatin accessibility in this tumor.

444 Nevertheless, along some branches we find a large number of DAPs remaining after
445 adjustment (i.e. DAPs identified both before and after adjustment), and thus must be due
446 to other mechanisms. Two example DAPs of this latter category are shown as insets in
447 Fig. 6b, with full list given in Supplementary Table 3. The first example is a peak at the
448 transcription start site (TSS) of the REC8 gene, which is located on chr14 where no
449 apparent CNAs were observed across the six major subclones. The TSS of REC8 is open
450 in clones 3-6 but closed in clones 1-2 (p-value<0.0001). REC8 is a gene encoding a
451 meiosis-specific cohesion component that is normally suppressed in mitotic proliferation,
452 and its role in cancer has recently gained increasing attention and controversy: While Yu
453 et al.[40] found the expression of this gene to suppress tumorigenicity in a gastric cancer
454 cell line, McFarlane et al.[41] postulated that it may be broadly activated in some cancers
455 where it generates LOH by reductional segregation. The opening of the TSS of REC8,
456 stably maintained in Clones 3-6, suggests that meiotic processes may underlie the
457 increased chromosomal instability of this multiclonal lineage. The second example is a
458 peak at the TSS of the WWOX gene, located on chr16, which is significantly depleted in
459 Clone 3 (p-value<0.0001). Although chr16 has LOH across all tumor cells, there are no
460 detectable subclonal differences, and thus we don't expect the decrease in accessibility
461 at WWOX for subclone 3 to be due to a large copy number event. Since WWOX is a well-

462 known tumor suppressor whose down-regulation is associated with more advanced
463 tumors[42, 43], its decrease in accessibility suggests a more aggressive phenotype for Clone
464 3. Overall, these two examples show how Alleloscope can be used to dissect the roles of
465 CNA and chromatin-level changes in the identification of gene targets for follow-up study.

466 **Discussion**

467 Despite recent advances in the application of single cell sequencing to cancer, we are
468 still far from understanding the diversity of genomes that are undergoing selection at the
469 single cell level. Notably, little is yet known about the intratumor diversity of allelic
470 configurations within CNA regions, and to what extent the diversity of cells in chromatin
471 accessibility can be attributed to diversity in allele-specific copy number. Here, we
472 presented Alleloscope, a new method for allele-specific copy number estimation that can
473 be applied to single cell DNA and ATAC sequencing data (separately or in combination).
474 Through a combination of matched linked-reads whole genome sequencing,
475 downsampling-based benchmark experiments, and simulations, we comprehensively
476 assessed the accuracy of Alleloscope and benchmarked against CHISEL, which is
477 currently the only other method for scDNA-seq allele-specific copy number profiling.
478 Detailed discussion about the two methods can be found in Supplementary Results.

479 We applied Alleloscope to a panel of breast, colorectal, and gastric tumors and cancer
480 cell lines, where it revealed an unprecedented level of allelic heterogeneity within
481 hypermutable CNA regions. In these regions, subclones reside on a gradient of allelic
482 ratios that is unobservable in total copy number analysis. In simple cases, these
483 hypermutable regions contain mirrored subclones, as previously identified[9, 10], but are

484 often much more complex. We observed multiple instances of convergent evolution
485 involving recurrent CNA events affecting the same region, some verified by linked read
486 sequencing. In accordance with the findings in Watkins et al.[44], we found using
487 Alleloscope that chromosomal instability drives the formation of subclones not only in
488 primary tumors but also after metastasis. Alleloscope was also applied to a scDNA-seq
489 dataset[15] generated from a different sequencing protocol (Supplementary Fig. 18;
490 Supplementary Methods), further demonstrating the general applicability of the approach.

491 Having established the allelic complexity of CNAs at single cell resolution, we next applied
492 Alleloscope to scATAC-seq data, thus enabling the combined study of clonal evolution
493 and chromatin accessibility. First, we considered the analysis of a public data set
494 consisting of two basal cell carcinoma samples, for which matched bulk whole-exome
495 sequencing data was used for initial genome segmentation upon which single cell CNA
496 genotyping was then conducted in the scATAC-seq data. Here we showed that
497 Alleloscope can detect amplifications, deletions, and copy-neutral LOH events accurately
498 in scATAC-seq data, and was able to find a subclone delineated by a copy-neutral LOH
499 event. Juxtaposing this subclone assignment with peak signals allowed us to detect a
500 wave of genome-wide chromatin remodeling in the lineage carrying the LOH. Next, we
501 applied Alleloscope to a complex polyclonal gastric cancer cell line with matched scDNA-
502 seq data. By overlaying peak signals with subclones delineated by allele-specific copy
503 number estimates, we developed a statistical test to adjust for copy numbers in identifying
504 DAPs. Comparison between the testing results before and after copy number adjustment
505 suggests that much of the intratumor heterogeneity in chromatin accessibility can be

506 attributed to CNAs. Focusing on subclone-enriched peaks with copy number adjustment
507 allowed the prioritization of genes for downstream follow-up.

508 Alleloscope can potentially be applied to the integration of single cell data of other
509 modalities, for example scATAC-seq and scRNA-seq data, to investigate the relationships
510 between clonal evolution, chromatin remodeling, and transcriptome. To facilitate
511 experimental design for single cell omics sequencing protocols, we investigated the
512 performance of Alleloscope under different experimental parameters (number of cells,
513 total per cell coverage, and total coverage at heterozygous SNP sites). Coverage at
514 heterozygous SNP sites is an especially important consideration for the design of scRNA-
515 seq and scATAC-seq studies, for which shifts in total coverage is an unreliable proxy for
516 underlying DNA copy number. For scATAC-seq, the lower heterozygosity within peak
517 regions led to lower number of reads mapping to heterozygous loci as compared to
518 scDNA-seq, and this resulted in noisier subclone detection. Most of the current scRNA-
519 seq technologies only sequence either the 3' or 5' end of the mRNA transcripts, which
520 limits the number of heterozygous SNP sites covered by reads. The latest developments
521 in single cell long read sequencing[45-47] and single cell multimodal sequencing[48] herald
522 new analysis opportunities with this method.

523 **Methods**

524 **ScDNA-seq Data Sets and Pre-processing**

525 The ten 10x scDNA-seq datasets analyzed in this study are summarized in
526 Supplementary Table 1. P5931, P5847 and P6461 scDNA-seq data were generated
527 using the method described in the previous study[35]. We applied the Cell Ranger DNA

528 pipeline (https://support.10xgenomics.com/single-cell-dna/software/overview/welcome;
529 beta version: 6002.16.0) for sample demultiplexing, read alignment, CNA calling and
530 visualization. Most data were aligned to the GRCh38 reference genome except for the
531 two processed BC10x datasets (GRCh37). For the tumor samples with a matched
532 normal sample, the GATK HaplotypeCaller was used to call heterozygous SNPs on the
533 matched normal samples. Otherwise, SNPs were retrieved on the tumor sample
534 themselves. Next, we applied VarTrix to efficiently generate SNP-by-cell matrices for
535 both reference and alternative alleles of the identified SNPs.

536 To select high-quality SNPs, we filtered out the SNPs with <5 reads for P5846 and
537 P5847, <10 reads for P5915 and P5931, <15 reads for P6335 and P6461, <20 reads for
538 P6198, SNU601, and ther BC10X samples based on the number of SNP detected for
539 each sample. Additionally, SNPs located in the centromeres and telomeres were
540 excluded. To exclude cells that might undergo apoptosis or cell cycles, noisy cells
541 labeled by the Cell Ranger tool were excluded.

542 **Single-cell ATAC Data sets, Sequencing and Preprocessing**

543 The scATAC-seq datasets analyzed in this study are summarized in Supplementary
544 Table 4. We used the following procedure to generate the SNU601 scATAC-seq
545 dataset.About 400,000 cells were washed with RPMI media and centrifuged (400g for 5
546 min at 4°C) twice. The supernatant was removed and resuspended with chilled PBS +
547 0.04% BSA solution. The resuspended pellet was centrifuged (400g for 5min at 4°C),
548 and the supernatant was removed again. Then 100 μL of chilled Lysis Buffer (10 mM
549 Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.1% Nonidet P40 Substitute,

550 0.1% Tween-20 and 0.01% digitonin) was added and carefully mixed 10 times. The tube
551 was incubated on ice for 7 min. After incubation, 1 mL of chilled Wash Buffer (10 mM
552 Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA and 0.1% Tween-20) was added
553 and mixed 5 times followed by centrifugation of nuclei (500g for 5 min at 4°C). After
554 removing the supernatant, nuclei were resuspended in chilled Nuclei Buffer (10X
555 Genomics), filtered by Flowmi Cell Strainer (40uM) and counted using a Countess II FL
556 Automated Cell Counter. Then the nuclei were immediately used to generate scATAC-
557 seq library.

558 ScATAC-seq library was generated using the Chromium Single Cell ATAC Library & Gel
559 Bead Kit (10X Genomics) following the manufacturer's protocol. We targeted 3000
560 nuclei with 12 PCR cycles for sample index PCR. Library was checked by 2% E-gel
561 (Thermofisher Scientific) and quantified using Qubit (Thermofisher Scientific).
562 Sequencing was performed on Illumina NextSeq500 using NextSeq 500/550 High
563 Output Kit v2.5 (Illumina).

564 Raw sequencing reads of the SNU601 scATAC-seq sample was de-multiplexed with the
565 10x Genomics Cell Ranger ATAC Software (v.1.2.0; https://support.10xgenomics.com/single-
566 cell-atac/software/pipelines/latest/algorithms/overview) and aligned to the GRCh38 reference
567 genome. The aligned scATAC-seq data of the two pre-treatment basal cell carcinoma
568 samples (BCC;SU006 and SU008) were retrieved from GSE129785[26]. To obtain all
569 potential SNPs for the SU006 and SU008 samples, GATK Mutect2 was used to call
570 single-nucleotide variants (SNVs) on the bam files deduplicated by the Picard toolkits of
571 both the t-cell dataset and the tumor dataset from the same tumor. All SNVs from the
572 paired tumor-normal datasets were combined and the read counts of these SNPs were

573 quantified for each cell in the tumor scATAC-seq dataset. For the SNU601 scATAC-seq
574 data, we instead quantified the read counts of the two alleles of the SNPs more reliably
575 called from the paired normal scDNA-seq data. Then the SNP-by-cell matrices for both
576 reference and alternative alleles were generated using Vartrix for all the scATAC-seq
577 datasets. To select potential germline SNPs, we further filtered out the SNVs <5 reads
578 for the SU008 sample and <10 reads for the SU006 sample. SNPs with extreme VAF
579 values <0.1 or >0.9 were also excluded for both samples. Because we used the phasing
580 information from the paired scDNA-seq data to assist the estimation of the haplotype
581 structures for the SNU601 scATAC-seq data, we did not filter out SNPs with extreme
582 VAF values for the scATAC-seq data. To improve the quality of the downstream
583 analysis for SNU601 scATAC-seq data, we filtered out cells <5 reads and the SNPs <5
584 reads.

585 **Linked-read sequencing and data processing**

586 The five samples with the linked-read sequencing data were acquired as surgical
587 resections following informed consent under an approved institutional review board
588 protocol from Stanford University. Samples were subjected to mechanical and
589 enzymatic dissociation as previously described, followed by cryopreservation of
590 dissociated cells[33].

591 Cryofrozen cells were rapidly thawed in a bead bath at 37 ºC. Cell counts were obtained
592 on a BioRad TC20 cell counter (Biorad, Hercules, CA) using 1:1 trypan blue dilution.
593 Between 1.5-2.5 million total cells were washed twice in PBS. Centrifugation was
594 carried out at 400g for 5 minutes. PBS was removed and cell pellets were frozen at -

595 80ºC. DNA extraction was carried out on cell pellets following thawing using either
596 MagAttract HMW DNA Kit (P5931) or AllPrep DNA/RNA Mini Kit (Qiagen Inc.,
597 Germantown, MD, USA) as per manufacturer's protocol. Quantification was carried out
598 using Qubit (Thermofisher Scientific).

599 Sequencing libraries were prepared from DNA using Chromium Genome Reagent Kit
600 (v2 Chemistry) (10X Genomics, Pleasanton, CA, USA) following manufacturer's
601 instructions. Sequencing was performed using Illumina HiSeq or NovaSeq sequencers
602 using 150x150 bp paired end sequencing and i7 index read of 8 bp. Long Ranger (10X
603 Genomics; v2.2.0) was used to align reads, call and phaseSNPs, indels and structural
604 variants.

605 **Segmentation**

606 The first step of Alleloscope is to segment the genome into regions with different CNA
607 profiles. The appropriate segmentation algorithm depends on what samples are
608 available. First, matched bulk DNA sequencing data (WGS/WES) or pseudo-bulk data
609 from scDNA-seq data can be segmented using FACLON[5], a segmentation method that
610 jointly models the bulk coverage and bulk VAF profiles, if a matched normal sample is
611 available. To accommodate segments from rare subclones, methods that integrate
612 shared cellular breakpoints in CNA detection for scDNA-seq[49] can improve sensitivity. If
613 a normal sample is unavailable, Alleloscope instead uses an HMM-based segmentation
614 method. The HMM method, which operates on the binned counts of pooled cells,
615 assumes a Markov transition matrix on four hidden states representing deletion, copy-
616 neutral state, single-copy amplification and double-copy amplification:

$$\begin{pmatrix} 1-3t & t & t & t \\ t & 1-3t & t & t \\ t & t & 1-3t & t \\ t & t & t & 1-3t \end{pmatrix}$$, where $t = 1 \times 10^{-6}$ by default. The segmentation plots in the study were generated by using the HMM algorithm. With the paired sample, the P6198 tumor sample was segmented using FALCON on the SNPs >30 reads with default parameters. Besides FALCON and HMM methods, users can choose the methods that work best for specific datasets. For example, we also tried ASCAT[11] on the P6198 sample for comparison (Supplementary Fig. 10).

Whole-exome sequencing (WES) data processing

The WES data of the two paired tumor-normal samples (SU006 and SU008) were obtained from PRJNA533341. Raw fastq files were aligned to the GRCh37 reference genome using bwa-mem[50] with duplicate reads removed using the Picard toolkits[51]. The copy number calls of paired normal-tumor samples were obtained using Varscan2[52]. Then the HMM algorithm was applied to segment the genome.

SNP Phasing and Single-cell Allele Profile Estimation per region

For each region r after segmentation, an expectation-maximization (EM)-based method is used to iteratively phase each SNP and estimate cell-specific allele-specific copy number states for scDNA-seq and scATAC-seq datasets. “Major haplotype” is defined as the haplotype with higher aggregate copy number in the sample. Let I_j indicate whether the reference allele of SNP j is located on the major haplotype and θ_{ir} denote major haplotype proportion of cell i for region r. The EM model iterates the E-step and the M-step. The complete log likelihood of the model is

637 $$l(\theta_{ir}) = \sum_{j=1}^{n} logP(A_{ij}, B_{ij}|\theta_{ir})$$

638 $$= \sum_{j=1}^{n} \{\left[A_{ij} log\theta_{ir} + B_{ij}\log(1-\theta_{ir})\right] I_j + \left[B_{ij} log\theta_{ir} + A_{ij}\log(1-\theta_{ir})\right](1-I_j)\}$$

639 where A_{ij} and B_{ij} are the observed read counts for the reference and alternative alleles
640 of cell i on SNP j. In the E-step, we calculate the expected value of the posterior
641 probability of the hidden variable I_j.

642 $$E_{\hat{\boldsymbol{\theta}}_{\boldsymbol{ir}}{}^{(t)}}\left[I_j \middle| A_{ij}, B_{ij}\right] = \hat{I}_j^{(t)} = \frac{\prod_i \hat{\theta}_{ir}^{(t)^{A_{ij}}}\left(1-\hat{\theta}_{ir}^{(t)}\right)^{B_{ij}}}{\prod_i \hat{\theta}_{ir}^{(t)^{A_{ij}}}\left(1-\hat{\theta}_{ir}^{(t)}\right)^{B_{ij}} + \prod_i\left(1-\hat{\theta}_{ir}^{(t)}\right)^{A_{ij}}\hat{\theta}_{ir}^{(t)^{B_{ij}}}}$$

643 where $\hat{\theta}_{ir}^{(t)}$ is the parameter from the tth iteration. In the M-step, $\hat{\theta}_{ir}$ is updated by solving

644 $$\hat{\theta}_{ir}^{(t+1)} = argmax_{\theta_{ir}} E\left[l(\theta_{ir})\middle| A_{ij}, B_{ij}, \hat{\theta}_{ir}^{(t)}\right]$$

645 $$= \frac{\sum_j [A_{ij}\hat{I}_j^{(t)} + B_{ij}(1-\hat{I}_j^{(t)})]}{\sum_j [A_{ij}\hat{I}_j^{(t)} + B_{ij}(1-\hat{I}_j^{(t)})] + \sum_j [A_{ij}(1-\hat{I}_j^{(t)}) + B_{ij}\hat{I}_j^{(t)}]}$$

646 where $\hat{\theta}_{ir}^{(t)}$ and $\hat{\theta}_{ir}^{(t+1)}$ are from two successive iterations of EM. An initial estimate $\hat{I}_j^{(0)}$ is
647 first derived from the bulk VAF profile, and the two steps are iterated until converge. The
648 maximum number of SNPs was set as 30,000 by default. For the SNU601 scATAC-seq
649 dataset, the phases estimated in the paired scDNA-seq dataset ($\hat{I}_j$'s) were directly
650 applied to estimate the $\hat{\theta}_{ir}$'s of the cells in the scATAC-seq data. To have enough

651 power, we included the cells with ≥ 20 reads covering the identified SNPs for each
652 region.

653 **Selecting normal cells and normal regions for single-cell Coverage Normalization**

654 To compute the relative coverage change for each cell in region r ($\hat{\rho}_{ir}$), normal cells and
655 diploid regions are required for normalization. The $\hat{\theta}_{ir}$'s inferred from the EM-based
656 algorithm are used to help identify normal cells and diploid regions. To identify normal
657 cells, the cells were divided into k largest groups ($k = 5$ by default) by hierarchical
658 clustering. The cth cluster is considered normal if it has the minimum distance calculated
659 by

660
$$\sum_{r=1}^{R} \left| \frac{\sum_{\hat{\theta}_{ir} \in S_c} \hat{\theta}_{ir}}{n_c} - 0.5 \right|^2$$

661 where S_c represents $\hat{\theta}_{ir}$ values of the cells in the cth cluster, and n_c is total cell number
662 in the cth cluster. All cells in the cth cluster are considered as candidate normal cells.

663 Putative diploid regions are next identified in each cluster by considering both $\hat{\theta}_{ir}$ and
664 coverage. Alleloscope computes the first measurement (d_{cr}) as the sum $\hat{\theta}_{ir}$ distance of
665 the cells in the cth cluster for each region r.

666
$$d_{cr} = \sum_{\hat{\theta}_{ir} \in S_c} \left| \hat{\theta}_{ir} - 0.5 \right|^2$$

667 Because regions with both haplotypes equally amplified can also have small d_{cr},
668 adjusted raw coverages ($\tilde{\rho}_{ir}$; defined as below) of cell i and region r are also
669 considered in diploid region selection.

670 $$\tilde{\rho}_{ir} = \frac{N_{ir}}{N_i} \times \frac{L_r}{LL},$$

671 where for cell i and region r, N_{ir} is the total read counts and N_i is the total read counts
672 across regions. L_r is length of the region r and LL is total length of the genome. For
673 region r, cells with $\tilde{\rho}_{ir}$ values larger than the 99th percentile are assigned the $\tilde{\rho}_{ir}$ values
674 equal to the 99th percentile across the cells. The second measurement (m_{cr}) used to
675 select diploid regions in the cth cluster is the mean $\tilde{\rho}_{ir}$ for each region r

676 $$m_{cr} = \frac{\sum_{\tilde{\rho}_{ir} \in S_c} \tilde{\rho}_{ir}}{n_c}$$

677 where S_c here represents $\tilde{\rho}_{ir}$ values of the cells in the cth cluster. Then Alleloscope
678 identifies potential diploid regions for each cluster by raking the sums of d_{cr} ranks and
679 m_{cr} ranks. Excluding the cth cluster identified as the normal group, Alleloscope
680 proposed a list of candidate diploid regions across the clusters by selecting the majority
681 region.

682 Because coverage on scATAC-seq data is confounded by epigenetic signals,
683 chromosome 22 for SU008 and chromosome 18 for SU006 were directly selected as
684 normal regions based on the matched WES data. Cells were considered as normal or
685 tumor cells based on the epigenetic signals. For the SNU601 scATAC-seq dataset,

686 chromosome 10 was selected as the normal region based on the paired scDNA-seq
687 data.

688 **Cell-level CNA state estimation**

689 The cell-level allele-specific copy number profiles are defined by both relative coverage
690 change ($\hat{\rho}_{ir}$) and major haplotype proportion ($\hat{\theta}_{ir}$) for region r. After the normal cells and
691 a control region are identified, cell-specific relative coverage change in region r is
692 calculated as

693 $$\hat{\rho}_{ir} = \frac{N_{ir}}{N_{i0}} / \, median(\frac{N_{0r}}{N_{00}})$$

694 where N_{ir} is total read counts in region r and N_{i0} is total read counts in a reference
695 region for cell i. N_{0r} is a vector denoting total read counts in region r of all identified
696 normal cells and N_{00} is a vector denoting total read counts in the same reference region
697 of all identified normal cells. For P6461 with no diploid region, $\hat{\rho}_{ir}$ was adjusted by
698 timing 2 for the abnormal cells. Because SNU601 is a tumor cell line with no normal
699 cells in the dataset, N_{0r} and N_{00} were calculated from the cells in the normal P6198
700 sample instead for the scDNA-seq data. For SNU601 scATAC-seq data, we aligned the
701 distribution of the $\hat{\rho}_{ir}$ values in paired scDNA-seq data to the distribution of the $\frac{N_{ir}}{N_i}$
702 values for each region to get the normalized $\hat{\rho}_{ir}$ in the scATAC-seq data. The
703 normalized $\hat{\rho}_{ir}$ values for the scATAC-seq data were computed by

704 $$\hat{\rho}_{ir}^{atac} = \frac{N_{ir}}{N_{i0}} / \, median(\frac{N_{ir}}{N_{i0}}) \times \, median(\hat{\rho}_{ir}^{dna})$$

705 Cells with extreme $\hat{\rho}_{ir}$ values larger than the 99th percentile and smaller than the first
706 percentile across the cells were excluded for each region. With the $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ pairs, cells
707 in the scDNA-seq data can be classified into haplotype profile g with the expected
708 (ρ_g, θ_g) values based on minimum distance. With noisier signals in the scATAC-seq
709 data, the haplotype structures identified in the paired scDNA-seq data can help guide
710 the genotyping for each region. For region r, the posterior probability that cell i carries
711 haplotype profile g was

712 $$P\left(GT_{ir} = g_r \middle| \hat{\rho}_{ir}, \hat{\theta}_{ir}\right) = \frac{P\left(\hat{\rho}_{ir}, \hat{\theta}_{ir} \middle| GT_{ir} = g_r\right)\pi_{g_r}}{\sum_{g_{r'}} P\left(\hat{\rho}_{ir}, \hat{\theta}_{ir} \middle| GT_{ir} = g'\right)\pi_{g'_r}},$$

713 where g_r denotes the haplotype profile observed in region r in the paired scDNA-seq
714 data and π_{g_r} denotes the prior probability that a randomly sampled cell carrying the g_r
715 haplotype profile. A uniform prior was used for π_{g_r}. In the formula,

716 $$P\left(\hat{\rho}_{ir}, \hat{\theta}_{ir} \middle| GT_{ir} = g_r\right) = P\left(\hat{\rho}_{ir} \middle| \mu = \rho_g; \sigma_\rho = 0.25\right) \times P\left(\hat{\theta}_{ir} \middle| \mu = \theta_g; \sigma = \sqrt{\frac{\hat{\theta}_{ir}\left(1 - \hat{\theta}_{ir}\right)}{n_{ir}}}\right),$$

717 where n_{ir} is total read counts in region r for cell i. For scATAC-seq data, the haplotype
718 profile of cell i in region r was estimated by maximizing the above posterior probability.
719 The haplotype profiles of each region are visualized using different colors in scatter
720 plots for both scDNA-seq and scATAC-seq data with the confidence scores calculated
721 using the distance of the points to the canonical centers and the standard deviations.

722 **Validations using matched linked-read sequencing data for each region**

723 We performed validations on five samples using matched linked-read sequencing data.
724 Two strategies were used for the validations. First, to assess the phasing accuracy of
725 region *r*, the largest haplotype block within region r in the matched linked-read
726 sequencing data was selected for comparison, ensuring that the included SNPs are
727 phased with respect to one another. Then the $\hat{I}_j$ estimates for SNPs within region r in
728 the scDNA-seq data were converted to binary indicator $\tilde{I}_j$

729 $$\tilde{I}_j = 1(\hat{I}_j > 0.5) \in \{0,1\}$$

730 Using the major/minor haplotype setting, all SNPs in the largest haplotype block in the
731 matched linked-read sequencing data were placed accordingly to retried I_j as the gold
732 standard. Comparing the binary estimates $\tilde{I}_j$ and I_j, the phasing accuracy for region r
733 was computed as follows:

734 $$ACC_{r,phasing} = \frac{\sum 1\,(\tilde{I}_j = I_j; \tilde{I}_j \in I_j)}{\sum 1(\tilde{I}_j \in I_j)},$$

735 Secondly, we evaluated the accuracy for cell-level CNA state estimation for region r. To
736 do this, for each cell in region r we compared the estimated haplotype profiles ($\hat{\theta}_{ir}$) to
737 the haplotype profiles (θ_{ir}) obtained by plugging in known phases provided by the
738 matched linked-read sequencing data as the gold standard. Similarly, the haplotype
739 block with the largest number of SNPs was used here. The θ_{ir}'s, used as the gold
740 standard, were retrieved by directly plugging in known phases I_j, provided by the
741 matched linked-read sequencing data (explained in the first strategy). Then the
742 accuracy for cell-level CNA state estimation of region r is computed as

743 $$ACC_{r,CNAstate} = \frac{\sum_i 1(\hat{\theta}_{ir} = \theta_{ir})}{\sum_i i}$$

744 **Simulations and Power Analysis**

745 For a simulated region, let n be the number of cells, m be the number of heterozygous
746 SNPs, θ be the major haplotype proportion, and μ_i be the total coverage of cell i sampled
747 from the cells on chr7 in the P5931 tumor sample. For cell i, we simulated total coverage
748 of SNP j (μ_{ij}) using a Poisson distribution

749 $$\mu_{ij} \sim Poisson(\mu_i),$$

750 where $i = 1 \sim n$. Parallelly, phases of SNP j (I_j) were simulated under a Bernoulli
751 distribution

752 $$I_j \sim Bernoulli(0.5),$$

753 where I_j indicates whether a reference allele is on the major haplotype for SNP j, and
754 j=1~m. Using μ_{ij} and I_j, simulated read counts of reference alleles of SNP j in cell i (A_{ij})
755 were simulated under a Binomial distribution

756 $$A_{ij} \sim Binomial(\mu_{ij}, p_{ij}),$$

757 where p_{ij} is the proportion of the reference allele at loci j in cell i with the values shown in
758 the following table

p_{ij}	cell i with CNA	cell i without CNA
$I_j = 1$	θ	0.5
$I_j = 0$	$1 - \theta$	0.5

759

760 Then simulated read counts of alternative alleles of SNP j in cell i (B_{ij}) were retrieved by

761 $$B_{ij} = \mu_{ij} - A_{ij}$$

762 In the first simulation used to illustrate distribution of the estimates from Alleloscope, we
763 fixed the cell number n to be 1,000, the SNP number m to be 10,000 which are typical in
764 real datasets. θ was set to be 1 and 0.66 for cells carrying deletion and one-copy
765 amplification respectively with the purity equal to 0.5. On the simulated A_{ij} and B_{ij}
766 matrices Alleloscope estimated phases for each SNP and CNA states for each cell.
767 Distribution of the estimated values versus the true values are visualized using boxplots.

768 To explore the effects of SNP numbers, cell coverage, cell numbers, and purity, power
769 analysis was performed for one-copy deletion and one-copy amplification scenarios. We
770 assessed the accuracy for phasing and cell-level CNA state estimation under the
771 following scenarios: SNP numbers from 1,000 to 50,000, mean coverage from 0.01 to
772 0,05 for each cell, cell number from 500 to 2500. For different scenarios, we assessed
773 the effect of three purity: 0.5, 0.1, and 0.01, reflecting from larger subclones to rare
774 subclones. All parameters remained the same as those in the previous paragraph except
775 for the parameters that were assessed. Phasing accuracy was calculated by comparing
776 true I_j's and estimated $\hat{I}_j$'s in the region. If $\hat{I}_j \geq 0.5$, the values were considered as 1.
777 Otherwise, the values were considered 0. On the other hand, the accuracy of cell CNA
778 state estimation was the clustering accuracy using the estimated $\hat{\theta}_i$ values. Cells with $\hat{\theta}_i$
779 values smaller than the midpoints between true θ of normal cells ($\theta_o = 0.5$) and true θ of
780 carriers ($\theta_{del} = 1$; $\theta_{amp} = 0.66$) were considered as normal cells; otherwise, cells were

781 considered as carriers. The clustering accuracy was calculated by comparing the clusters
782 to the true cell states.

783 **Cell Lineage Reconstruction**

784 For scDNA-seq data, cell-specific haplotype profiles from Alleloscope across the
785 genome are used to reconstruct cell lineage trees. The "Gower's distance" is calculated
786 using "cluster" R package on the nominal haplotype profiles between cells. The
787 'pheatmap' R package was used to perform hierarchical clustering on the distance using
788 the "ward.D2" method. Because fewer SNPs lead to higher variance in $\hat{\theta}_{ir}$'s, we by
789 default included segments with more than 2,000 SNPs identified. For visualization, each
790 segment was plotted with its length proportional to 5,000,000 bins, and the heights of
791 the clustering tree were log-transformed.

792 To explore tumor subclonal structure in the two BCC scATAC-seq datasets, we instead
793 clustered the cells using $\hat{\theta}_{ir}$ values, which are orthogonal to the CNA signals based on
794 total coverage, across the segments with more than 500 SNPs. Then hierarchical
795 clustering is performed on the Euclidean distance using the "ward.D2" method.

796 Because the subclones for the SNU601 sample were identified from the scDNA-seq
797 data, we adopted a supervised strategy to assign each cell in the SNU601 scATAC-seq
798 dataset into different subclones. First, we identified 10 marker regions-- chr1b, 3b-d, 4b,
799 7a, 11b, 13b, and 20b-c that help to differentiate the cells into the six major subclones
800 based on the subclone specific copy number profiles from the scDNA-seq data.
801 Combining the haplotype profiles across the ten regions for each cell enables

802 assignment of the cells into one of the six subclones with high confidence. The posterior
803 probability of cell i coming from clone k was

804 $$P\left(Clone_i = k \middle| \hat{\rho}_i, \hat{\theta}_i\right) = \frac{P\left(\hat{\rho}_i, \hat{\theta}_i \middle| Clone_i = k\right)\pi_k}{\sum_{k'} P\left(\hat{\rho}_i, \hat{\theta}_i \middle| Clone_i = k'\right)\pi_{k'}},$$

805 where $k \in \{1{\sim}6\}$ for the six clones and π_k is the prior probability that a randomly
806 sampled cell coming from the kth clone, which was set to uniform in the analysis.. In the
807 formula,

808 $$P\left(\hat{\rho}_i, \hat{\theta}_i \middle| Clone_i = k\right) = \prod_x P(\hat{\theta}_{ix}, \hat{\rho}_{ix} | Clone_i = k)$$

809 $$= \prod_x \phi\left(\frac{\hat{\theta}_{ix} - \theta_{kx}}{\sqrt{n_{ix}\theta_{kx}(1-\theta_{kx})}}\right)\phi\left(\frac{\hat{\rho}_{ix} - \rho_{kx}}{\sigma_\rho}\right),$$

810 where x indexes the ten marker regions; $\hat{\theta}_{ix}$ and $\hat{\rho}_{ix}$ are the estimated major haplotype
811 proportion and relative coverage for cell i in the scATAC-seq data; θ_{kx} and ρ_{kx} are the
812 "known values" for specific haplotype profiles for clone k derived from the paired
813 scDNA-seq data; and n_{ix} is the number of total read counts in the xth marker region for
814 cell i. Each cell was assigned into one of the six subclones by maximizing the posterior
815 probability with the confidence score being the posterior probability of the assigned
816 clone.

817 **ScATAC-seq data analysis**

818 To investigate the relationships between allele-specific CNAs and chromatin
819 accessibility, we processed the peak signals in addition to the allele-specific CNAs for

820 each cell in scATAC-seq data. For the two public BCC samples, the peak-by-cell
821 matrices were obtained from GSE129785. We log-transformed the count matrices,
822 selected the peaks in >10% cells, regressed out cell total coverage for each peak by
823 linear regression, and projected the cells onto the UMAP plot using genome-wide peak
824 signals[53]. The cell type identify for each cluster was retrieved from the previous study[26].
825 To further explore intratumor heterogeneity, we selected the cells labeled as tumor
826 cells, repeated the pre-processing steps described above, and projected the tumor cells
827 onto the UMAP plot. Then DNA level information and epigenetic signals for each cell
828 were integratively visualized.

829 For the SNU601 scATAC-seq dataset, scATAC-pro[54] was used to call peaks and
830 generate the peak-by-cell matrix. We first filtered out the cells that have proportions of
831 fragments on the detected peaks <0.4 and or total peaks outside of the range
832 15,000~100,000, and filtered out the peaks observed in less than 10% of cells. After the
833 matrix was log-transformed, we regressed out cell total coverage for each peak by
834 linear regression. Using Louvain clustering, the cells in two clusters with extreme total
835 fragments were removed. After cell filtering, we repeated the process and projected the
836 cells onto the UMAP plot. Then the clonal assignment based on the DNA information
837 and the peak signals were integrated at the single-cell level

838 **Differentially accessible peaks (DAPs) identification with copy number**
839 **adjustment for scATAC-seq data**

840 To identify differentially accessible peaks (DAPs) after accounting for copy number
841 differences between two clones, we developed a statistical test based on a generalized

842 log-likelihood ratio (LLR) statistic. We first define some necessary terms: Given a
843 segmentation of the genome, for a given clone c, let $\theta_c = \{(L_R, Z_R)\}$ be the copy number
844 profile, which can be expressed as a vector of region lengths (L_R) and average copy
845 numbers (Z_R) across all regions R. For peak k, we define function $f_k(\theta_c)$ to be the
846 proportion of genomic DNA in the peak region based on the copy number profile of
847 clone c, computed as follows:

848 $$f_k(\theta_c) = \frac{L_{r(k)} Z_{r(k)}}{\sum_r L_r Z_r} \times \frac{L_k}{L_{r(k)}},$$

849 where $r(k)$ is the region r to which peak k belongs and L_r is the length of peak k.

850 For peak *k* and clone c, let Y_{ck} be the total read count across the cells in the clone, and
851 let N_c be the total read count across all peaks summed across the cells in the clone. Y_{cg}
852 can be modeled by Binomial sampling,

853 $$Y_{ck} \sim Binomial(N_c,\ f_k(\theta_c) p_{ck}),$$

854 where, p_{ck} is the relative probability of detecting reads in the peak k region after
855 adjusting for the copy numbers. Between clones with different copy number profiles,
856 differences in p_{ck} is evidence for chromatin remodeling. Thus, pairwise comparisons
857 were performed for each branch (clone c_1 versus clone c_2) and for each peak, using the
858 generalized likelihood ratio test with the hypothesis $H_0: p_{c_1 k} = p_{c_2 k} = p_{0k}$ and $H_1: p_{c_1 k} \neq$
859 $p_{c_2 k}$.

860 Let $l(p)$ denote the log-likelihood function of Y_{ck} under this model, and $\hat{p}$ denote the
861 MLE. The LLR for GLRT follows the Chi-squared distribution as below with derivation of
862 MLE estimates in the Supplementary Methods.

863 $$LLR_k = -2\log\frac{\max\limits_{p_{0k}\in p} l(p_{0k})}{\max\limits_{p_{c_1k};\, p_{c_2k}\in p} l(p_{c_1k,}, p_{c_2k,})} = 2\ell(\hat{p}_{c_1k}, \hat{p}_{c_2k}) - 2\ell(\hat{p}_{0k}) \sim X_1^2,$$

864 Using the GLRT, a peak is considered as a DAP if its FDR adjusted p-value<0.01 for
865 the pairwise clonal comparison.

866 Simultaneously, the GLRT under the binomial model without copy number adjustment
867 was also performed for comparison:

868 $$Y_{ck} \sim Binomial(N_c,\ p_{ck}),$$

869 The same criterion was used to identify DAPs for each pairwise clonal comparison
870 (FDR adjusted p-values<0.01). Each DAP was further mapped to the genes that are
871 potentially regulated based on the $\pm$ 2,000 bp distance on the genome.

872 **<u>Data availability</u>**

873 The patient scDNA-seq and linked-read sequencing data generated for this study are
874 available under dbGAP identifier phs001711. The scATAC-seq dataset is available in the
875 National Institute of Health's SRA repository under accession PRJNA674903. There are
876 no restrictions on data availability or use. The other patient scDNA-seq data were
877 obtained from dbGAP under accession phs001818.v3.p1[35] and phs001711[12]. The cell
878 line scDNA-seq dataset was from the Sequence Read Archive (SRA) under accession

879 PRJNA498809. The public scATAC-seq data and whole exome sequencing data were

880 obtained from the SRA under accession PRJNA532774[26] and PRJNA533341[39].

881 **Code availability**

882 Alleloscope is available on GitHub at https://github.com/seasoncloud/Alleloscope.

883 **Reference**

884 1. Baylin, S.B. & Jones, P.A. A decade of exploring the cancer epigenome - biological and
885 translational implications. *Nat Rev Cancer* **11**, 726-734 (2011).
886 2. Sandoval, J. & Esteller, M. Cancer epigenomics: beyond genomics. *Curr Opin Genet Dev*
887 **22**, 50-55 (2012).
888 3. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306-313 (2012).
889 4. Burrell, R.A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences
890 of genetic heterogeneity in cancer evolution. *Nature* **501**, 338-345 (2013).
891 5. Chen, H., Bell, J.M., Zavala, N.A., Ji, H.P. & Zhang, N.R. Allele-specific copy number
892 profiling by next-generation DNA sequencing. *Nucleic Acids Res* **43**, e23 (2015).
893 6. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor
894 sequencing data. *Ann Oncol* **26**, 64-70 (2015).
895 7. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations
896 from tumor whole-genome sequence data. *Genome Res* **24**, 1881-1893 (2014).
897 8. Shen, R. & Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity
898 analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**, e131 (2016).
899 9. Jamal-Hanjani, M. et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl*
900 *J Med* **376**, 2109-2121 (2017).
901 10. Zaccaria, S. & Raphael, B.J. Characterizing allele- and haplotype-specific copy numbers
902 in single cells with CHISEL. *Nat Biotechnol* (2020).
903 11. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U*
904 *S A* **107**, 16910-16915 (2010).
905 12. Andor, N. et al. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines
906 reveals rules of in vitro evolution. *NAR Genom Bioinform* **2**, lqaa016 (2020).
907 13. Bakker, B. et al. Single-cell sequencing reveals karyotype heterogeneity in murine and
908 human malignancies. *Genome Biol* **17**, 115 (2016).
909 14. Garvin, T. et al. Interactive analysis and assessment of single-cell copy-number
910 variations. *Nat Methods* **12**, 1058-1060 (2015).
911 15. Kim, C. et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated
912 by Single-Cell Sequencing. *Cell* **173**, 879-893 e813 (2018).
913 16. Laks, E. et al. Clonal Decomposition and DNA Replication States Defined by Scaled
914 Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221 e1222 (2019).
915 17. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94
916 (2011).

917 18. Velazquez-Villarreal, E.I. et al. Single-cell sequencing of genomic DNA resolves sub-
918 clonal heterogeneity in a melanoma cell line. *Commun Biol* **3**, 318 (2020).
919 19. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome
920 sequencing. *Nature* **512**, 155-160 (2014).
921 20. Klemm, S.L., Shipony, Z. & Greenleaf, W.J. Chromatin accessibility and the regulatory
922 epigenome. *Nat Rev Genet* **20**, 207-220 (2019).
923 21. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition
924 of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-
925 binding proteins and nucleosome position. *Nat Methods* **10**, 1213-1218 (2013).
926 22. Buenrostro, J.D. et al. Single-cell chromatin accessibility reveals principles of regulatory
927 variation. *Nature* **523**, 486-490 (2015).
928 23. Corces, M.R. et al. The chromatin accessibility landscape of primary human cancers.
929 *Science* **362** (2018).
930 24. Granja, J.M. et al. Single-cell multiomic analysis identifies regulatory programs in
931 mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458-1465 (2019).
932 25. Litzenburger, U.M. et al. Single-cell epigenomic variability reveals functional cancer
933 heterogeneity. *Genome Biol* **18**, 15 (2017).
934 26. Satpathy, A.T. et al. Massively parallel single-cell chromatin landscapes of human
935 immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925-
936 936 (2019).
937 27. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring
938 transcription-factor-associated accessibility from single-cell epigenomic data. *Nat*
939 *Methods* **14**, 975-978 (2017).
940 28. Cabal-Hierro, L. et al. Chromatin accessibility promotes hematopoietic and leukemia
941 stem cell activity. *Nat Commun* **11**, 1406 (2020).
942 29. Dravis, C. et al. Epigenetic and Transcriptomic Profiling of Mammary Gland
943 Development and Tumor Models Disclose Regulators of Cell State Plasticity. *Cancer*
944 *Cell* **34**, 466-482 e466 (2018).
945 30. Guilhamon, P. et al. Single-cell chromatin accessibility in glioblastoma delineates cancer
946 stem cell heterogeneity predictive of survival. *bioRxiv*, 370726 (2020).
947 31. Pan, D. et al. A major chromatin regulator determines resistance of tumor cells to T cell-
948 mediated killing. *Science* **359**, 770-775 (2018).
949 32. Qu, K. et al. Chromatin Accessibility Landscape of Cutaneous T Cell Lymphoma and
950 Dynamic Response to HDAC Inhibitors. *Cancer Cell* **32**, 27-41 e24 (2017).
951 33. Shaffer, S.M. et al. Rare cell variability and drug-induced reprogramming as a mode of
952 cancer drug resistance. *Nature* **546**, 431-435 (2017).
953 34. Shu, S. et al. Synthetic Lethal and Resistance Interactions with BET Bromodomain
954 Inhibitors in Triple-Negative Breast Cancer. *Mol Cell* **78**, 1096-1113 e1098 (2020).
955 35. Sathe, A. et al. The cellular genomic diversity, regulatory states and networking of the
956 metastatic colorectal cancer microenvironment. *bioRxiv* (2020).
957 36. Bell, J.M. et al. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer
958 aneuploidy. *Nucleic Acids Res* **45**, e162 (2017).
959 37. Greer, S.U. et al. Linked read sequencing resolves complex genomic rearrangements in
960 gastric cancer metastases. *Genome Med* **9**, 57 (2017).
961 38. Zheng, G.X. et al. Haplotyping germline and cancer genomes with high-throughput
962 linked-read sequencing. *Nat Biotechnol* **34**, 303-311 (2016).

963 39. Yost, K.E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade.
964 *Nat Med* **25**, 1251-1259 (2019).
965 40. Yu, J. et al. REC8 functions as a tumor suppressor and is epigenetically downregulated in
966 gastric cancer, especially in EBV-positive subtype. *Oncogene* **36**, 182-193 (2017).
967 41. McFarlane, R.J. & Wakeman, J.A. Meiosis-like Functions in Oncogenesis: A New View
968 of Cancer. *Cancer Res* **77**, 5712-5716 (2017).
969 42. Aqeilan, R.I. et al. Loss of WWOX expression in gastric carcinoma. *Clin Cancer Res* **10**,
970 3053-3058 (2004).
971 43. Baryla, I., Styczen-Binkowska, E. & Bednarek, A.K. Alteration of WWOX in human
972 cancer: a clinical view. *Exp Biol Med (Maywood)* **240**, 305-314 (2015).
973 44. Watkins, T.B.K. et al. Pervasive chromosomal instability and karyotype order in tumour
974 evolution. *Nature* (2020).
975 45. Gupta, I. et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands
976 of cerebellar cells. *Nat Biotechnol* (2018).
977 46. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error
978 corrected Nanopore single cell transcriptome sequencing. *Nat Commun* **11**, 4025 (2020).
979 47. Singh, M. et al. High-throughput targeted long-read single cell sequencing reveals the
980 clonal and transcriptional landscape of lymphocytes. *Nat Commun* **10**, 3120 (2019).
981 48. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat*
982 *Methods* **17**, 11-14 (2020).
983 49. Wang, R., Lin, D.Y. & Jiang, Y. SCOPE: A Normalization and Copy-Number Estimation
984 Method for Single-Cell DNA Sequencing. *Cell Syst* **10**, 445-452 e446 (2020).
985 50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler
986 transform. *Bioinformatics* **25**, 1754-1760 (2009).
987 51. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for
988 analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
989 52. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery
990 in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).
991 53. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and
992 projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
993 54. Yu, W., Uzun, Y., Zhu, Q., Chen, C. & Tan, K. scATAC-pro: a comprehensive
994 workbench for single-cell chromatin accessibility sequencing data. *Genome Biol* **21**, 94
995 (2020).
996

997 **Acknowledgements**

998 The work is supported by the National Institutes of Health [P01HG00205ESH to B.T.L.,
999 S.M.G. and H.P.J., 5R01-HG006137-07 and 1U2CCA233285-01 to C-Y.W. and to N.R.Z.,
1000 1R35HG011292-01 to B.T.L.]. Additional support to HPJ came from the Research Scholar

1001 Grant, RSG-13-297-01-TBG from the American Cancer Society, Clayville Foundation and
1002 the Gastric Cancer Foundation.

1003 **Author contributions**

1004 C.-Y.W. and N.R.Z. conceived the computational methods and designed the study with
1005 help from H.P.J. C.-Y.W. developed and implemented the computational methods and
1006 conducted all data analyses. B.T.L. helped in data interpretation. B.T.L., H.K. and A.S.
1007 performed all related sample preparation and sequencing. S.M.G. performed data pre-
1008 processing and coordinated data transfer. H.P.J. advised all experiments and data
1009 collection. C.-Y.W., N.R.Z, and H.P.J. wrote the manuscript. All authors read and
1010 approved the final manuscript.

1011 **Competing interests**

1012 The authors declare no competing interests.

1013

Figure 1.

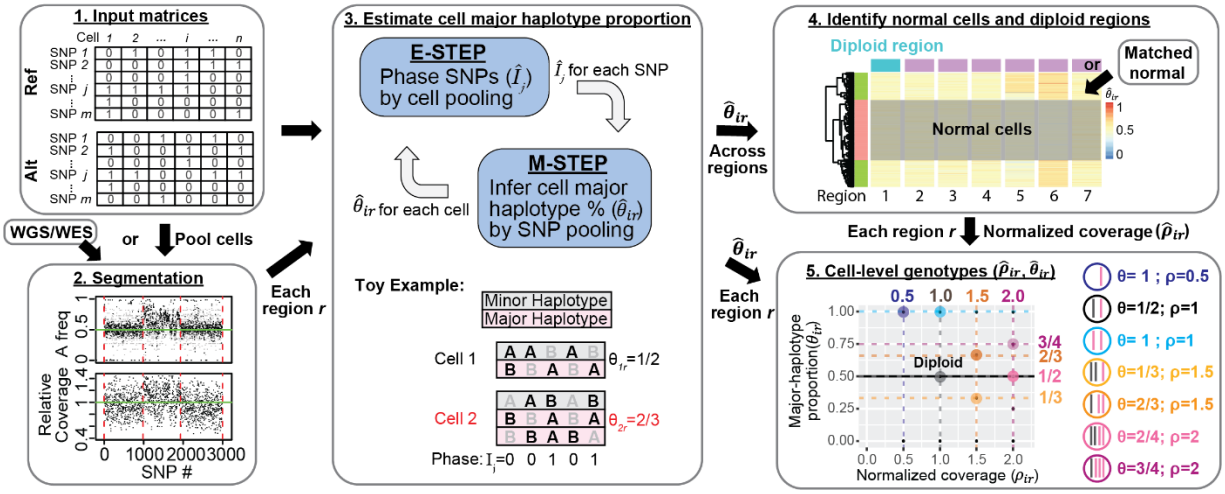

Fig. 1: Overview of allele-specific copy number estimation of single cells with Alleloscope. 1. The algorithm operates on raw read count matrices for reference allele (Ref) and alternative allele (Alt) computed from single cell DNA or ATAC sequencing. 2. First, we obtain a segmentation of the genome based on sample-matched whole genome or whole exome sequencing data using FALCON[5]. If scDNA-seq is available, cells can be pooled to derive a pseudo-bulk. 3. We first define $\hat{\theta}_{ir}$ as the major haplotype proportion for cell i for region r derived from the segmentation. Then for each region r, Alleloscope simultaneously phase SNPs ($\hat{I}_j$) and estimate cell major haplotype proportion ($\hat{\theta}_{ir}$) by expectation maximization (EM) algorithm. In the E-step, information is pooled across cells to estimate the phasing of each SNP. In the M-step, information is pooled across all SNPs in the region are pooled to estimate the major haplotype proportion $\hat{\theta}_{ir}$ for each cell. The toy example shows a scenario with two cells for a region containing 5 SNPs, with cell 2 carrying an amplification of the major haplotype (in pink). For each cell and each SNP, alleles that are observed in a sequenced read are bolded in black (we assume that only one read is observed, reflecting the sparsity of the data). The true phase (I_j) of the SNPs and the true major haplotype proportion (θ_{ir}) are shown. 4. After $\hat{\theta}_{ir}$ are estimated for each region r, $\hat{\theta}_{ir}$'s are pooled across all regions to identify candidate normal cells and candidate normal regions for computing a normalized coverage $\hat{\rho}_{ir}$ for region r in cell i. Matched normal can be specified if available. 5. Alleloscope assigns integer allele-specific copy numbers to each cell for each region based on the ($\hat{\rho}_{ir}$, $\hat{\theta}_{ir}$) pairs.

Figure 2.

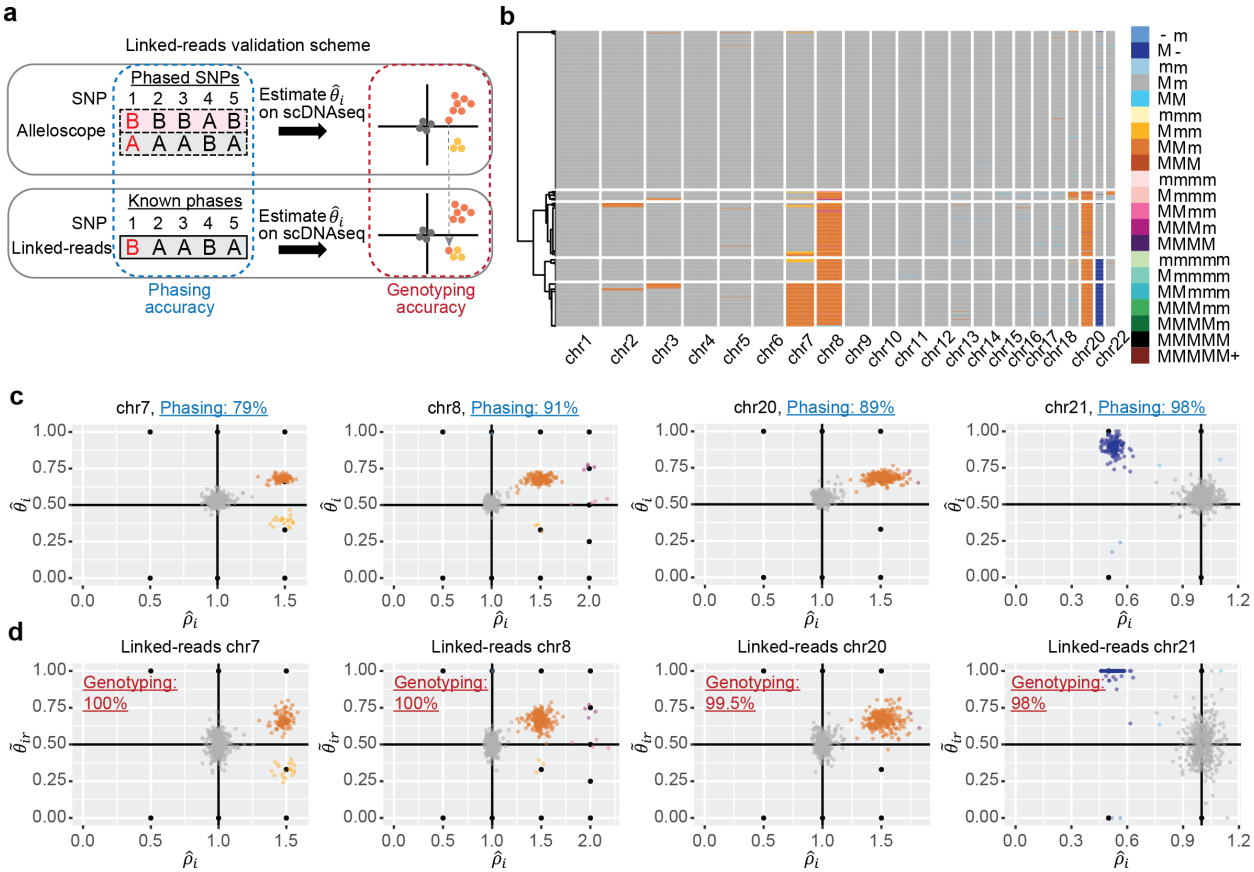

Fig. 2: Validation of the Alleloscope results on the P5931 gastric cancer patient sample and linked-read sequencing data. (a) Illustration of the validation scheme using linked-read sequencing data. Phasing accuracy and accuracy in allele-specific copy number state estimation are used to access performance of the method. (b) Hierarchical clustering of cells in the P5931t sample based on allele-specific copy numbers given by Alleloscope, showing normal cells and 4 main clones, as well as a number of small clones marked by highly confident low-frequency mutations. M: Major haplotype, m: minor haplotype. (c) $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ estimated by Alleloscope for four regions, colored by the inferred haplotype profile. Note that clusters fall on canonical points corresponding to discrete allele-specific copy number configurations. Phasing accuracy for each region is shown in the plot title. In the color legend, M and m represent the "Major haplotype" and "minor haplotype" respectively. (d) Similar to (c), with $\tilde{\theta}_{ir}$ estimated using known SNP phases from matched linked-read sequencing data, colored by the haplotype profiles assigned in (c) using Alleloscope without the given phasing information. Accuracy in allele-specific copy number state estimation (Genotyping accuracy) is labeled in the plots.

Figure 3.

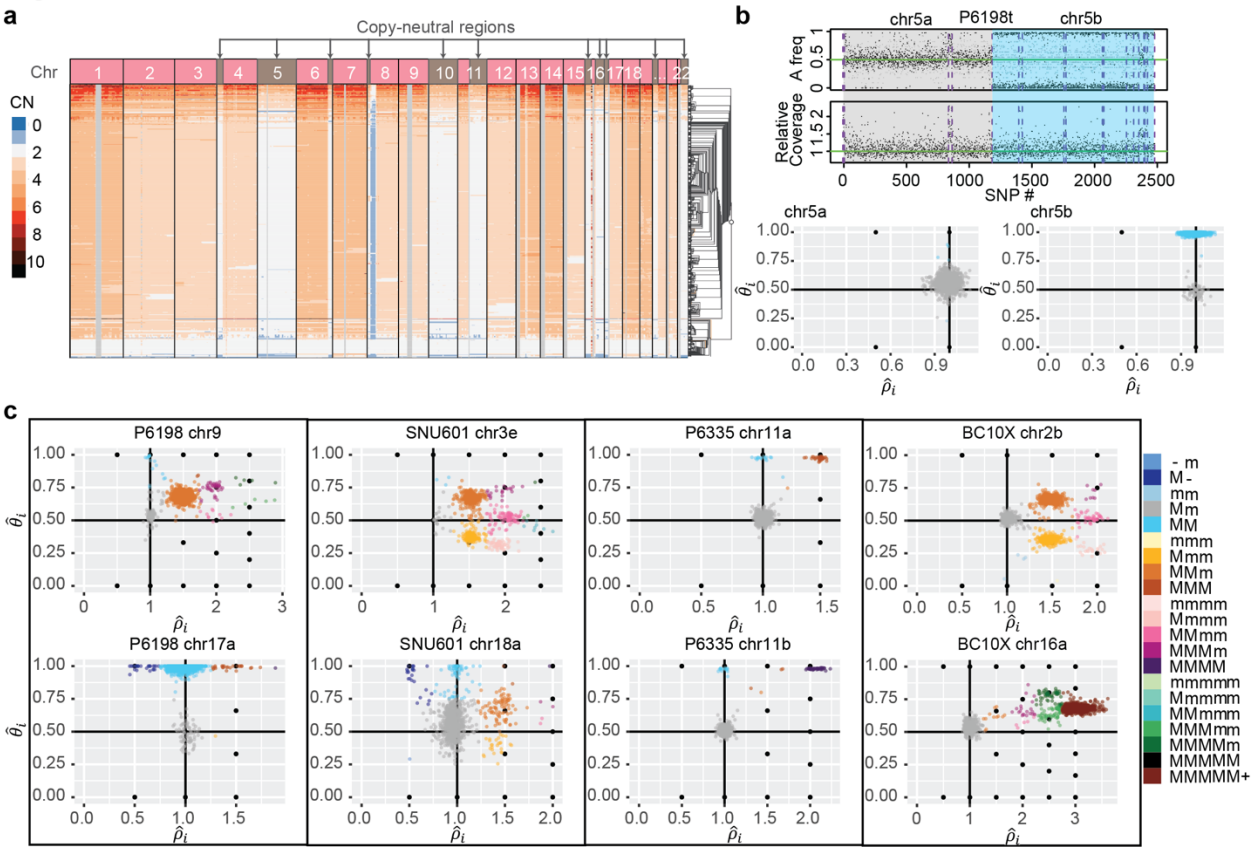

Fig. 3: Across multiple cancer types, Alleloscope detects loss-of-heterozygosity events and multi-allelic copy number aberrations, delineating complex subclonal structure which are invisible to total copy number analysis. (a) The Cell Ranger hierarchical clustering result for P6198t with copy-neutral regions labeled (total 512 cells). (b) Top: FALCON segmentation of P6198 chr5 into two regions with different allele-specific copy number profiles. Bottom: Detailed haplotype profiles of the two regions from Alleloscope, showing that the first region is diploid across cells and the second region has a loss-of-heterozygosity for a subpopulation of cells. The a and b following the chromosome number denote two ordered segments. (c) Single cell allele-specific estimates $(\hat{\rho}, \hat{\theta})$, colored by assigned haplotype profiles, for select regions in the samples P6198t (metastasized colorectal cancer sample), SNU601 (gastric cancer cell line), P6335 (colorectal cancer sample), and BC10X (breast cancer cell line). In the color legend, M and m represent the "Major haplotype" and "minor haplotype" respectively. The lower-case letters following the chromosome number in the titles denote the ordered genomic segments.

Figure 4.

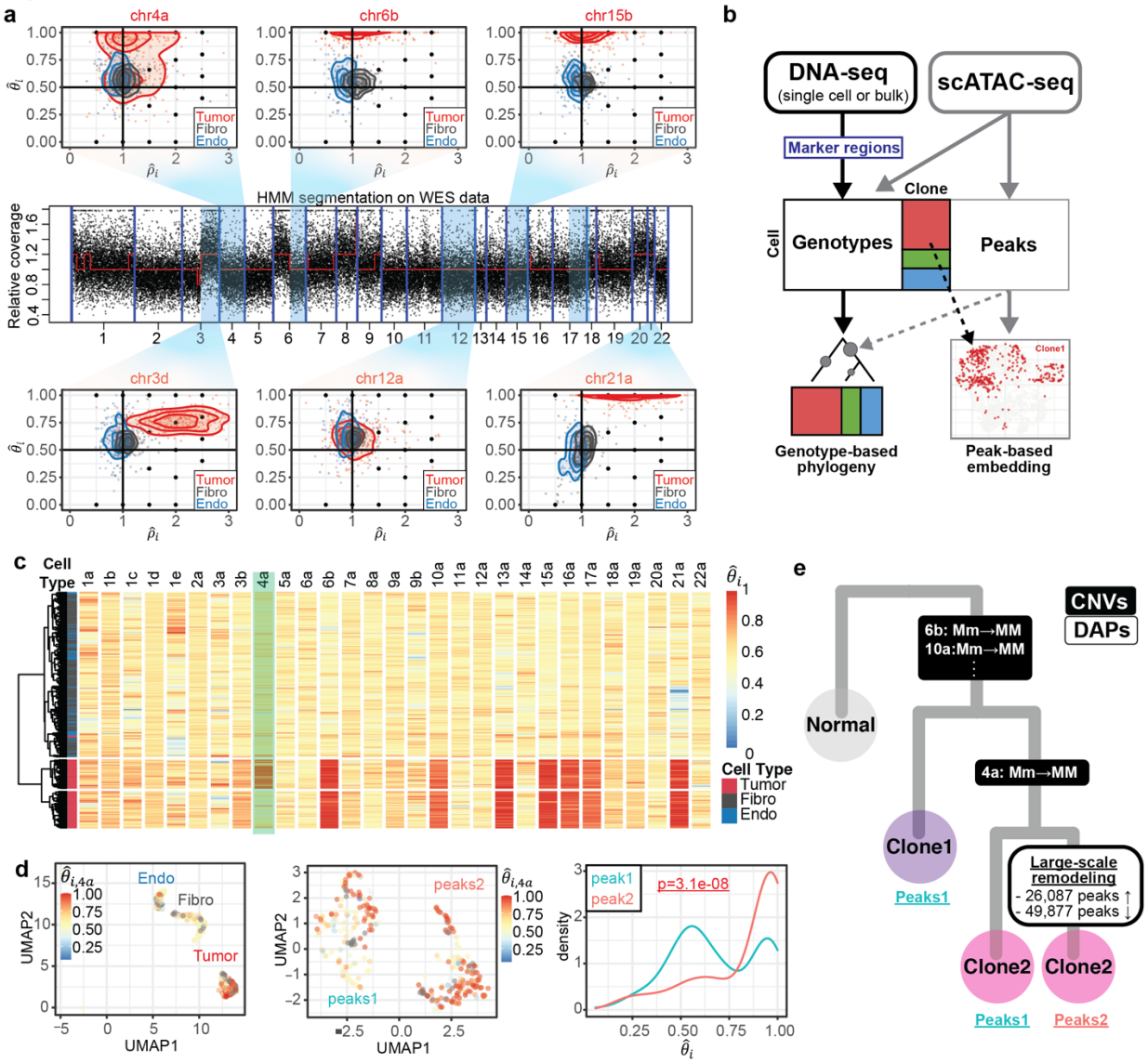

Fig. 4: Alleloscope multiomic analysis of scATAC-seq data of a basal cell carcinoma sample (SU008[26]). (a) Genotype profiles for six example regions for cells in scATAC-seq data. The regions are taken from segmentation of matched whole exome sequencing (WES) data. Each dot represents a cell-specific $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ pair. Cells are colored by annotation derived from peak signals[26], Tumor: tumor cells, Fibro: fibroblasts, Endo: Endothelial cells]. Density contours are computed for each cell type (tumor, fibroblasts, endothelial) separately and shown by color on the plot. The lower-case letters following the chromosome number in the titles denote the ordered genomic segments. (b) Pipeline for multi-omics analysis integrating allele-specific copy number estimates and chromatin accessibility peak signals on ATAC-seq data. (c) Hierarchical clustering of cells by major haplotype proportion $(\hat{\theta})$ allows the separation of tumor cells from normal cells, as well as the differentiation of a subclone within the tumor cells. The marker region on chr4a separating the two tumor subclones is highlighted. (d) Integrated visualization of

1084 chr4a major haplotype proportion ($\hat{\theta}_{ir}$) and genome-wide peak profile. Left: UMAP
1085 projection of the 788 cells in the dataset by their genome-wide peak profile, colored by
1086 $\hat{\theta}_{ir}$. The cell type annotation (endothelial, fibroblasts, and tumor cells) is labeled in the
1087 plot. Middle: UMAP projection of only the 308 tumor cells by their genome-wide peak
1088 profile shows two well-separated clusters: peaks1 and peaks2. Right: Density of $\hat{\theta}_{ir}$
1089 values for the peaks1 and peaks2 subpopulations. (e) Intratumor heterogeneity of SU008
1090 is shaped by a subclonal LOH of chr4a followed by subsequent genome-wide chromatin
1091 remodeling leading to three subpopulations: Clone 1 which does not carry the chr4a LOH
1092 (peaks cluster 1), Clone 2 carrying the chr4a LOH (peaks cluster 1), and remodeled clone
1093 2 (peaks cluster 2).

1094

Figure 5.

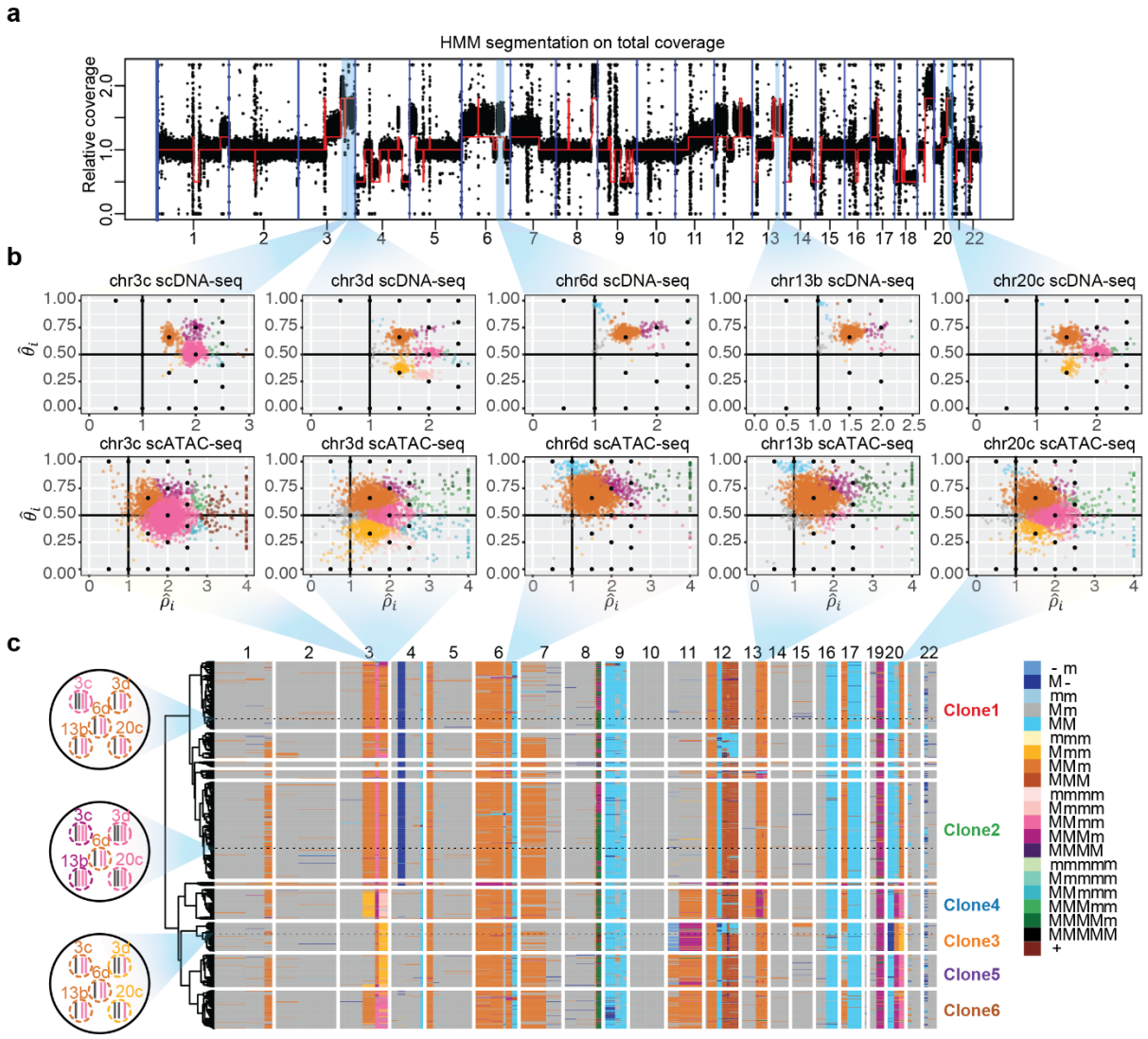

Fig. 5: Alleloscope analysis of scDNA-seq and scATAC-seq data reveals complex subclonal heterogeneity in the SNU601 gastric cancer cell line. (a) Genome segmentation using HMM on the pooled total coverage profile computed from scDNA-seq data. (b) Single cell allele-specific copy number profiles ($\hat{\rho}, \hat{\theta}$) for five regions in scDNA-seq and scATAC-seq data. Cells are colored by haplotype profiles according to legend in Fig. 5c. (c) Tumor subclones revealed by hierarchical clustering of allele-specific copy number profiles from the scDNA-seq data. Genotypes of the five regions shown in Fig. 5b, for three example cells, are shown in the left. The haplotype structures for the 5 regions in Fig. 5b of three cells randomly chosen from Clone 1, 2, and 3, are shown to the left of the heatmap. In the color legend, M and m represent the "Major haplotype" and "minor haplotype" respectively. The six clones selected for downstream analysis in scATAC-seq data are labeled in the plot.

Figure 6.

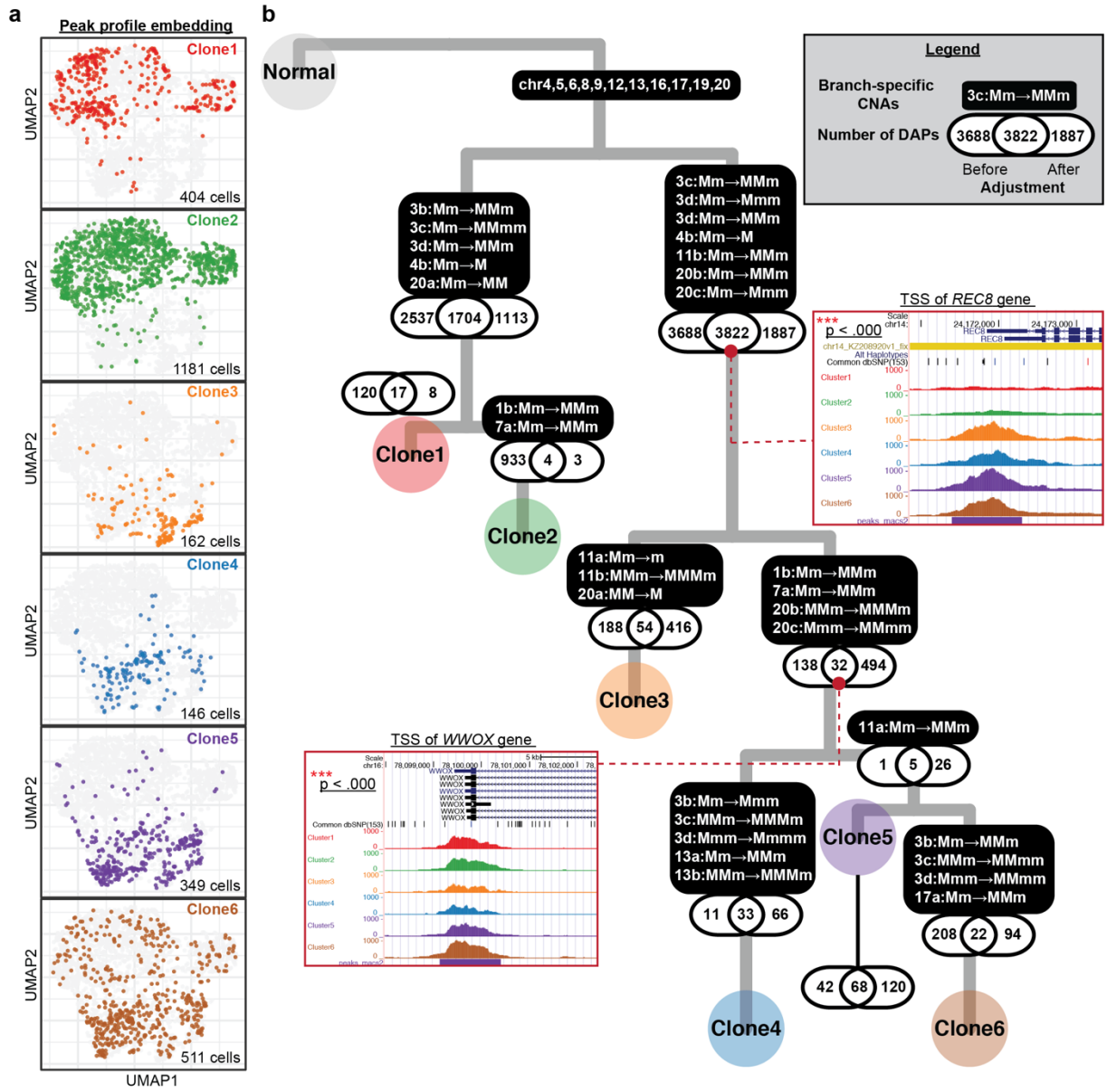

Fig. 6: Integrative analysis of allele-specific copy number and chromatin accessibility for SNU601 ATAC sequencing data. (a) UMAP projection of genome-wide scATAC-seq peak profile on 2,753 cells. The same group of cells were clustered into one of the six subclones based on their allele-specific copy number profiles across the 10 selected regions. Cells in different subclones are labeled with different colors, using the same color scheme as that for the subclone labels in Fig. 4c. The number of cells colored in each UMAP is shown at the bottom-right corners. (b) A highly probable lineage history of SNU601, with copy number alterations (CNAs) and differentially accessibility peaks (DAPs) marked along each branch. P-values of the tests for association between DAPs and CNAs are shown along each branch. For two example DAP genes, pooled peak signals for each subclone are shown as inset plots.

1121 **Table 1. Comparison between Alleloscope and CHISEL using two benchmark**
1122 **strategies.**

Dataset	Method	Sensitivity	Specificity
Benchmark with matched linked-read sequencing data			
P5915	Alleloscope	0.9373	0.9915
	CHISEL (before correction)	0.7284	0.9757
	CHISEL (after correction)	0.7786	0.9722
P5931	Alleloscope	0.9402	0.9986
	CHISEL (before correction)	0.7520	0.9434
	CHISEL (after correction)	0.0112	0.9508
P6198	Alleloscope	0.9433	0.9666
	CHISEL (before correction)	0.9397	0.9311
	CHISEL (after correction)	0.9700	0.9359
P6335	Alleloscope	0.9671	0.9906
	CHISEL (before correction)	0.7858	0.9873
	CHISEL (after correction)	0.8404	0.9943
P6461	Alleloscope	0.8638	0.9904

	CHISEL (before correction)	0.7932	0.9624
	CHISEL (after correction)	0.8143	0.9411
Benchmark with subsampling			
50% subsample	Alleloscope	0.9261	0.9905
	CHISEL (before correction)	0.9267	0.9383
	CHISEL (after correction)	0.9496	0.9446
25% subsample	Alleloscope	0.8249	0.9525
	CHISEL (before correction)	0.6515	0.6412
	CHISEL (after correction)	0.6533	0.6475

1123