Alleloscope: Integrative analysis of single cell haplotype-divergent copy number alterations and chromatin accessibility changes reveals novel clonal architecture of cancers

Chi-Yun Wu[1,2], Billy T. Lau[3,4], Heonseok 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.

Introduction

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

First, consider copy number aberrations (CNAs), through which we have derived much of our current understanding of the relationship between genome instability and tumor evolution[4]. Total copy number profiling, which estimates the sum of the copy numbers of the two homologous chromosomes, is inadequate to characterize some types of cancer genomic aberrations. Such events include the pervasively occurring copy-neutral loss of heterozygosity (LOH)[5-8], intriguing "mirrored events"[9, 10] where a given tumor may have cancer cells carrying amplification of one haplotype are intermingled with cancer cells carrying amplification of the other haplotype, and the even more complex alterations that are only detectable through allele-specific analysis[11]. While the importance of allele-specific copy number has been emphasized in bulk DNA sequencing analysis[5-8, 11], most single-cell CNV analysis considers only total copy number due to low per-cell coverage[12-19]. Recently, Zaccaria et al. developed CHISEL[10], a method for single-cell allele-specific copy number analysis, but requires externally phased haplotypes based on large

reference cohorts. Despite these advances, there remain many missing details about the genomic landscape of allelic imbalances when considering single cells.

Epigenetic modifications are also an important genomic feature of cancer. Analysis of chromatin structure is feasible with a variety of methods including transposase-accessible chromatin sequencing (ATAC-seq). This approach is applied either with conventional bulk-based or single-cell sequencing. Subsequently, analysis of chromatin structure has shown that epigenetic remodeling modulates the plasticity of cells in cancer[20-24], leads to stem-like properties[25-27] and generates therapeutic resistance[28-31]. Since copy number alterations involve large gains and losses of available chromatin, we expect the chromatin accessibility of a region to be influenced by the changes in underlying copy number. Current scATAC-seq studies estimate total copy number profiles by smoothing the read coverage and normalizing the signals against a control cell population, yet this appropriate control is often difficult to identify [23, 32]. Currently, there is no method for reliable total or allele-specific copy number profiling in scATAC-seq data, and thus, how to disentangle the effects of CNA and chromatin remodeling in shaping the epigenetic landscape remains a challenge.

Addressing these challenges, we present Alleloscope, a method for **allel**e-**s**pecific **cop**y number **e**stimation and multi**o**mic profiling in **s**ingle **c**ells. Alleloscope does not rely on external phasing information, and can be applied to scDNA-seq data or to scATAC-seq data with sample-matched bulk DNA sequencing data. We first apply Alleloscope on scDNAseq data from four gastric cancer samples, four colorectal cancer samples, and a breast cancer sample[10, 12, 33]. For three of the gastrointestinal cancer samples, results are extensively validated by 10x linked-read sequencing which provides accurate phasing

information[34-36]. In these datasets, Alleloscope accurately identifies LOH and mirrored-subclonal amplification events, and 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[23] and a complex polyclonal gastric cancer cell line that we analyzed by scDNA-seq. In these samples, we evaluate the accuracy of Alleloscope in genotyping and clone assignment and demonstrate its application to the integrative analysis of CNA and chromatin accessibility.

Results

Overview of Alleloscope allele-specific copy number estimation

First, we briefly overview Alleloscope's method for allele-specific copy number estimation (Figure 1). Clone assignment and integration with peak signals in scATAC-seq data will be described later. Alleloscope relies on two types of data features: coverage, derived from all reads that map to a given region, and allelic imbalance, derived from allele-informative reads that cover heterozygous loci in the region. We start with some essential definitions. For a given single nucleotide polymorphism (SNP) site, we refer to its mean coverage across cells as *bulk coverage* and its mean variant allele frequency (VAF = ratio of alternative allele read count to total read count) across cells as its *bulk VAF*. Between

the two parental haplotypes, we define the term “major haplotype” as the haplotype with higher mean count across cells. Note that a haplotype may be the “major haplotype” of a sample, but be the haplotype with lesser copy number within some cells. For each individual cell i, in any given CNA region, we define two key parameters: (1) the major haplotype proportion (θ_i), defined as the count of the major haplotype divided by the total copy number for the region, and (2) total copy fold change (ρ_i), defined as the ratio of the total copy number of the region in the given cell relative to that in normal cells.

The genotyping algorithm starts by segmenting the genome into regions of homogeneous allele-specific copy number using both the bulk coverage and bulk VAF profiles (Fig.1 Step 2). This can be achieved by multiple existing algorithms, which may be combined to increase detection sensitivity, see Methods for details. In our analyses of scATAC-seq data, the segmentation relied on the matched scDNA-seq data or the whole-exome sequencing data, which ensures that the putative CNA regions considered for genotyping are not confounded by the broad chromatin remodeling that occur in cancer.

Now consider each putative CNA region. An expectation-maximization (EM) based algorithm is used to iteratively phase each SNP and estimate the major haplotype proportion (θ_i) for each cell (Fig. 1, step 3). For each SNP j, let $I_j \in \{0,1\}$ be the indicator of whether the reference allele of SNP j is a component of the major haplotype. An initial estimate $\hat{I}_j^{(0)}$ is first derived from the bulk VAF profile. Then, in iteration t, Alleloscope computes $\hat{\theta}_i^{(t)}$ by pooling counts across sites within the region, weighted by the current phasing $\hat{I}_j^{(t)}$, then updates the estimate of I_j based on $\hat{\theta}_i^{(t)}$ by pooling counts across cells. The estimates of θ_i and I_j usually converge within a few iterations as described in the

Methods. With paired scDNA-seq and scATAC-seq data derived from the same sample, I_j values are estimated from scDNA-seq data which can be used to compute θ_i for each cell in the scATAC-seq data. This step enables integration of the two data types.

The estimated major haplotype proportions ($\hat{\theta}_i$'s), along with coverage that is preliminarily normalized by total cell read count ($\tilde{\rho}_i$), are then used to identify a set of normal cells and diploid regions (Fig. 1, Step 4). This information is used to estimate a relative coverage change ($\hat{\rho}_i$) for each cell within each CNA region. If cell i's true allele-specific copy numbers are homogeneous within the given region, then its true value of (θ_i, ρ_i) should belong to a set of canonical points displayed in Step 5 of Figure 1. Thus, the estimated values $(\hat{\rho}_i, \hat{\theta}_i)$ are clustered across cells and associated with one of the canonical values to yield the cell-level haplotype profiles for the CNV region. The cell-specific haplotype profiles across different sequencing platforms serve as the base for subsequent multi-omics analysis in scATAC-seq data (Fig. 4b).

Figure 1.

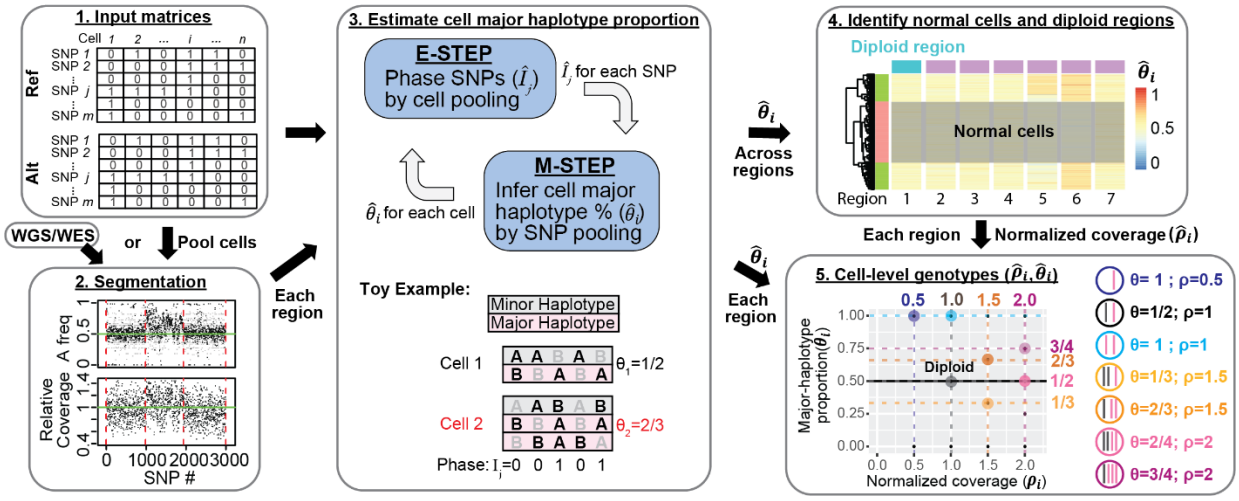

Fig. 1: Overview of allele-specific copy number estimation of single cells with Alleloscope. 1. The algorithm operates on raw read count matrices for reference allele (Ref) and alternative allele (Alt) computed from single cell DNA or ATAC sequencing. 2. First, we obtain a segmentation of the genome 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. For each region derived from the segmentation, simultaneously phase SNPs ($\hat{I}_j$) and estimate cell major haplotype proportion ($\hat{\theta}_i$) by expectation maximization (EM) algorithm. Since we are focusing on only one region, the region indicator is suppressed in our notation here. 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 $\hat{\theta}_i$ 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 ($\hat{\theta}_i$) are shown. 4. For region r let $\{\hat{\theta}_{ir}\}$ be its estimated major haplotype proportions across cells i. Pool data across regions to identify candidate normal cells and candidate normal regions for computing a normalized coverage $\hat{\rho}_{ir}$ for region r in cell i. 5. Alleloscope assigns integer allele-specific copy numbers to each cell for each region based on the ($\hat{\rho}_{ir}$, $\hat{\theta}_{ir}$) pairs.

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

First, we assessed the phasing and genotyping accuracy of Alleloscope in scDNA-seq data using matched linked-read whole-genome sequencing data on three gastrointestinal tumor samples: P5931, P6335 and P6198. Linked-read sequencing, in which one derives reads from individual high molecular weight DNA molecules, provides variants that can be phased into extended haplotypes covering Mb[34-36]. As a result, one obtains accurate, Mb-scale haplotype information from cancer genome. To evaluate the accuracy of phasing, we compared the haplotypes estimated by Alleloscope to the haplotypes obtained from linked-read WGS. Specifically, we used the WGS haplotype to evaluate the allele-specific copy number estimation for each cell and to assess the impact of phasing errors on genotyping accuracy (Fig. 2a).

Figure 2b shows the results for the gastric cancer sample from P5931, whose genome-wide copy number profile indicates clear CNA events on four chromosomes—chr7, chr8, chr20, and chr21. For each event, the scatter plots of $(\hat{\theta}_i, \hat{\rho}_i)$ estimated by Alleloscope and colored by haplotype profiles, are shown in Fig. 2c. Note that the $(\hat{\theta}_i, \hat{\rho}_i)$ clusters fall almost directly on top of the expected canonical values (e.g. (1/2, 1) for diploid, (2/3, 1.5) for 1 copy gain of major haplotype), also shown in the plots. Interestingly, chromosomes 7, 8, and 21 each show subclonal clusters have differing allelic ratios but the same total copy number, which would not be detectable for this sample without allele-specific estimation for the same region. We denote the major haplotype of a region by “M”, and the minor haplotype by “m”. The chromosome 7 amplification exhibits two tumor

subclones with mirrored-subclonal CNAs (MMm and Mmm), each subclone amplifying a different haplotype. Such a mirrored-subclonal CNA configuration is also observed for the deletion on chromosome 21 (M- and m-). The chromosome 8 amplification exhibits as four tumor subclones with different haplotype profiles— MMm, Mmm, MMmm, and MMMm.

We compared the phasing estimated by Alleloscope ($\hat{I}_j$) against the whole genome haplotypes. The phasing accuracy is 98% for the deleted region (chr21), ~90% for the two clonal amplifications (on chr8 and chr20), and 79% for the subclonal chr7 amplification (shown in the titles of the scatter plots of Fig. 2c). Moreover, we evaluated the phasing accuracy for some of the somatic alterations. Figure 2d shows scatterplots of $\hat{\rho}_i$ against major haplotype proportion computed using haplotypes derived from linked-read sequencing ($\tilde{\theta}_i$), with the same coloring as Figure 2c. Comparing the scatterplots in Figure 2d to their counterparts in Figure 2c reveals that Alleloscope's estimated cell haplotype profiles are highly concordant with those derived with the cancer haplotypes from linked-read WGS. Specifically, the concordance is ~100% across all four events (the concordance for each event is labeled in the scatter plots of Figure 2d). This shows that the genotyping algorithm in Alleloscope is robust to errors in phasing (e.g. for chr7). Similar analysis performed for P6335 is given in Supplementary Fig. 1.

Figure 2.

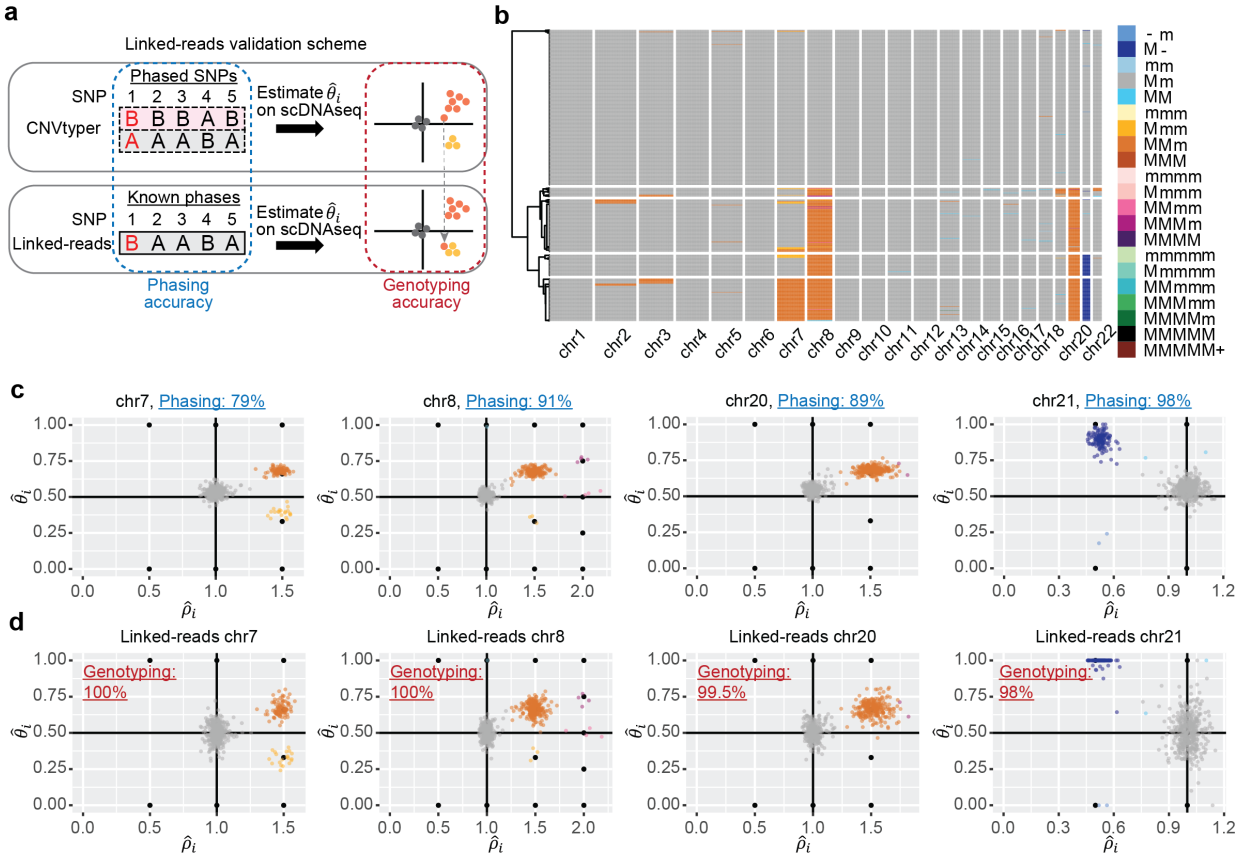

Fig. 2: Validation of the Alleloscope results on the P5931 gastric cancer patient sample and linked-reads sequencing data. (a) Illustration of the validation scheme using linked-reads sequencing data. Phasing accuracy and genotyping accuracy are used to access performance of the method. (b) Hierarchical 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) $(\hat{\rho}_{ri}, \hat{\theta}_{ri})$ 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 $\hat{\theta}_i$ estimated using known SNP phases from matched linked-reads sequencing data, colored by the haplotype profiles assigned in (c) using Alleloscope without the given phasing information. Genotyping accuracy is labeled in the plots.

Since copy neutral LOH events, common in cancer genomes, can only be identified through allele-specific copy number analysis. We examined the accuracy of Alleloscope specifically for copy-neutral LOH events with a colorectal adenocarcinoma from P6198. This tumor sample had a conventional WGS profile revealing several copy-neutral LOH regions that were not evident when considering the copy number heatmap in cellranger (Fig. 3a). Chromosome 5 presents an illustrative example: The bulk VAF clearly separates this chromosome into two main regions, a normal region followed by a copy-neutral LOH (Fig. 3b). Concordantly, Alleloscope reveals a cluster centered at $(\rho, \theta) = (1,1)$ corresponding to copy-neutral LOH only for the region on the right (Fig. 3b), which cleanly separates the tumor cells from normal cells. Comparing this tumor's haplotype profiles derived using the haplotypes from linked-read WGS showed that the accuracy of Alleloscope for copy-neutral LOH events is nearly 100% (Supplementary Fig2).

Figure 3.

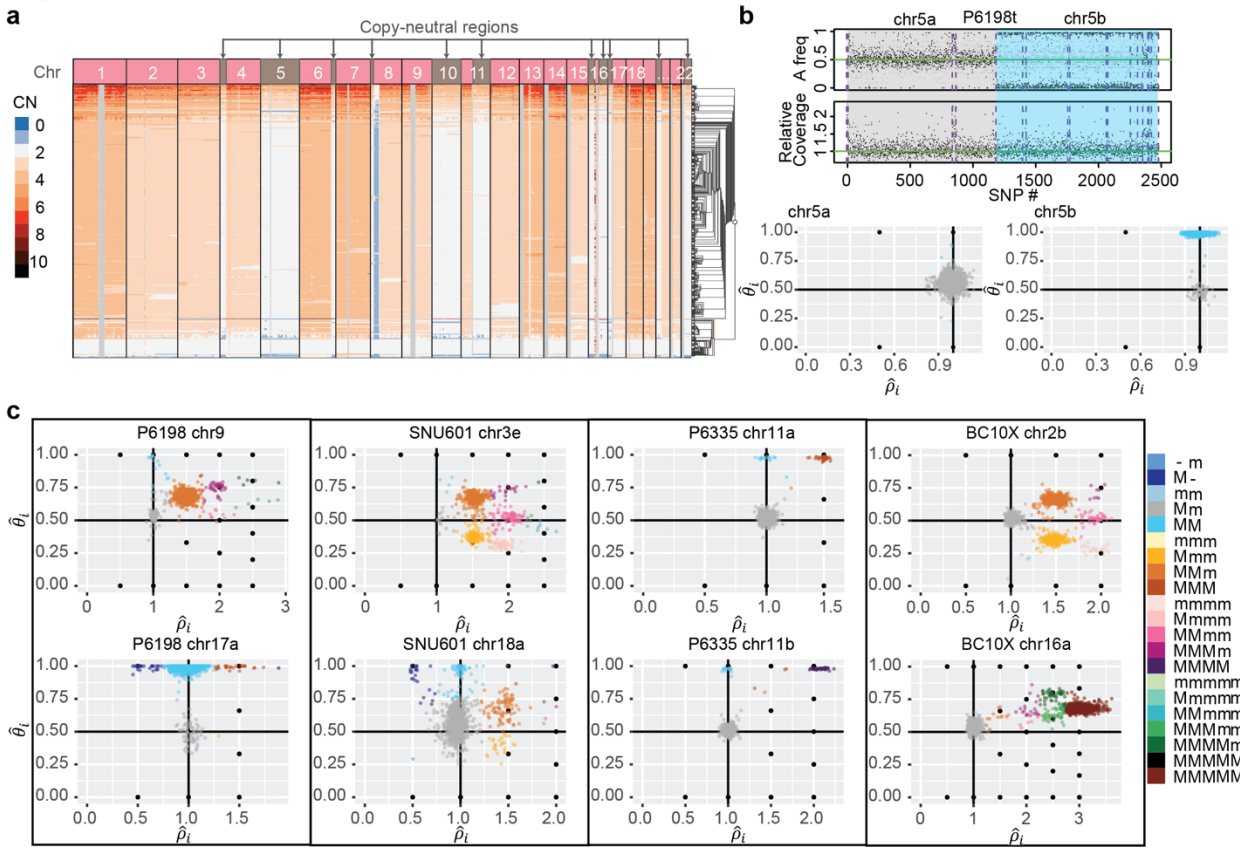

Fig. 3: Across multiple cancer types, Alleloscope detects loss-of-heterozygosity events and multi-allelic copy number aberrations, delineating complex subclonal structure which are invisible to total copy number analysis. (a) The Cell Ranger hierarchical clustering result for P6198t with copy-neutral 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.

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

We used Alleloscope to analyze an additional set of cancers including three colon cancer, two gastric cancer, a gastric cancer cell line, and a breast cancer with detailed segmentation plots and heatmaps for genome-wide allele-specific copy number profiles in the supplementary figures 3-9. Across most of the samples, we observed a high prevalence of complex subclonal CNAs indicated by multiple clusters of different haplotype structures within a given genomic region with prototypical examples from P6198, SNU601, P6335 and BC10X shown in Figure 3c. In some regions, such as chromosome 9 of SNU601, 3q of SNU601, 2q and 16p of BC210x, we see as many as seven subclonal clusters for a single event. In many cases we note multiple clusters corresponding to the same total copy number but varying in allelic dosage. Minor subclones carrying deletion of one haplotype can be easily masked by dominant subclones carrying amplifications of the other haplotype in a conventional sequencing analysis without the benefit of single cell resolution or an analysis that considers only copy number without allelic information. Overall, the high subclonal diversity in these genomic regions prone to chromosomal instability reveal an aspect of intratumor heterogeneity that was previously undetectable.

Recurrent chromosomal instability events affecting both haplotypes and producing gradients in haplotype dosage, is a common theme across all samples analyzed. Consider, for example, the region on chromosome 9 of P6198, which reveals 7 subpopulations of cells: besides the normal cell cluster and the dominant tumor cell cluster with the haplotype profile MMm, there is a small cluster of cells with copy neutral

LOH, two small subclones at four chromosome copies and two more at five chromosome copies. This produces major haplotype ratios of $\{\frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{4}{5}, 1\}$ in different cells, possibly conferring different fitness values. Another example of such complexity is chromosome 3q of SNU 601 and chromosome 2q of BC10x, which share a similar pattern: two mirrored-subclonal CNAs (MMm, mmM) at total copy number of 3, as well as mirrored-subclonal CNAs (MMMm, MMmm, mMMM) at total copy 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 evolutionary route by which such diversity was achieved, Alleloscope reveals that a whole genome doubling event is highly likely to have taken place early in the development of BC10x and P6198, but not in the development of SNU601 (see Supplementary Fig. 3&5). Thus, the subclones at 2q in BC10x and 3q in SNU601 must have evolved through different evolutionary routes: In BC10x, the early whole-genome doubling produces the cluster MMmm, from which the other clusters of different haplotype profiles were most likely derived through successive loss and gene conversion events. On the contrary, the clusters on 3q of SNU601 were most likely a result of successive amplification events starting from the normal haplotype profile Mm. The fact that different evolutionary routes, in two different cancer types (breast and gastric) evolved to have such similar allelic-specific copy number patterns imply that such haplotype dosage gradients may be serve as an important substrate for selection in tumor evolution.

Another recurring theme is the co-occurrence of LOH of one parental haplotype and amplification of the other haplotype for the same event, producing co-evolving subclones that carry multiple copies of the remaining haplotype. Figure 3c shows such examples for chromosome 17p of P6198 and chromosome 11p, 11q of P6335. For 17p of P6198, a

gene conversion leading to copy-neutral LOH is most likely the early event, followed by separate loss and gain events that lead to the clusters M- and MMM.

Sometimes, extreme instability of a chromosome region leads to clones with LOH coexisting with clones with amplification of the lost haplotype. For example, this is what occurred for SNU601's chr18a (Figure 3c). Such clones may be missed in a conventional bulk WGS analysis or even with single cell analysis which does not take into account allelic information. The co-existence of such subclones suggests that the selective fitness of a tumor cell may rely not only the loss of tumor suppressors on the deleted haplotype, but also on the dosage of the remaining haplotype which may contain genes conducive to survival. Using the procedure in Figure 2a, we validated our findings of these multiallelic subclones in P6198 and P6335 by comparing to the paired linked-reads sequencing data. Phasing accuracy is high for all LOH and amplification event types that create an allelic imbalance (Supplementary Fig. 2).

Juxtaposition single cell copy number and chromatin remodeling events in cancers

Next, we applied Alleloscope to the multi-omic analysis of scATAC-seq data from two basal cell carcinoma with matched WES data[37]. Using matched whole-exome sequencing data, the genome of each sample was first segmented into regions of homogeneous bulk copy number (Fig. 4a, middle panel shows the segmentation for SU008). Alleloscope was then applied to the scATAC-seq data to derive allele-specific copy number estimates of each cell in each region. Scatterplots of $(\hat{\rho}, \hat{\theta})$ for five example CNA regions and 1 control region (chr12) from SU008 are shown in Fig. 4a. For this sample, peak profiles characterizing chromatin accessibility separated the cells confidently into three main

clusters: 308 tumor cells, 259 fibroblasts and 218 endothelial cells. Since normal cells are not expected to carry broad copy number events, we compared the $(\hat{\rho}, \hat{\theta})$ values of the tumor cells against those of the fibroblast and epithelial cells to assess our genotyping accuracy. Density contours for each cell type are shown in the $(\hat{\rho}, \hat{\theta})$- scatterplots (Fig. 4a). The $(\hat{\rho}, \hat{\theta})$ values clearly separate the tumor cells from the normal cells for each CNV region, with the tumor cell cluster positioned at canonical points, indicating that these statistics used by Alleloscope can accurately distinguish amplifications and loss-of-heterozygosity events in scATAC-seq data. In particular, Alleloscope differentiated the cells that carry copy neutral LOH events through shifts in major haplotype proportion. Note that normal cells, which are not expected to carry broad chromosome-scale CNVs, exhibit chromosome-level deviations in total coverage due to broad chromatin remodeling as exemplified by the chr6b region. Furthermore, many regions with no CNA signal in bulk WES data also exhibit shifts in aggregate coverage in ATAC data, but with no significant difference in their $\hat{\theta}_i$ distribution. Thus, relying solely on shifts in coverage, without complementary shifts in major haplotype proportion, would lead to false positive copy number detections for scATAC-seq data.

By assigning allele-specific CNA profiles to single cells in scATAC-seq data, Alleloscope allows the integrative analysis of chromosomal instability and chromatin remodeling as follows (Figure 4b): The scATAC-seq data, paired with bulk or single-cell DNA sequencing data, allows us to detect subclones In parallel, a peak-by-cell matrix can be computed following standard pipelines. Then, the subclone memberships or CNA profiles can be visualized on the low-dimensional embedding of the peak matrix, and the subclones can be further compared in terms of peak or transcription factor motif enrichment. Precise

haplotype profiles for each subclone then allows us to identify significantly enriched/depleted peaks after accounting for copy number differences, thus delineating events that are uniquely attributable to chromatin remodeling.

Hierarchical clustering using major haplotype proportion $\hat{\theta}$ identifies the tumor cells from the normal cells for both SU006 (Supplementary Fig. 10) and SU008, and clearly delineates a subclone in SU008 marked by a copy-neutral LOH event on chr4a (Fig. 4c). Focusing on SU008, we call the cell lineage that carries the chr4a LOH event clone-2, and the other lineage clone-1. In parallel, clustering by peaks cleanly separates the tumor cells from the epithelial cells and fibroblasts (Fig. 4d: left), and further, demarcates two distinct clusters in the tumor cells (peaks-1 and peaks-2) (Fig. 4d: middle). What is the relationship between the peaks-1 and peaks-2 clusters obtained from peak signals to the two clones delineated by chr4a LOH? Coloring by chr4a major haplotype proportion ($\hat{\theta}$) on the peaks-derived UMAP shows that the LOH in this region is carried by almost all of the cells in peaks-2 but only a subset of the cells in peaks-1 (Fig. 4d: middle). This can also be clearly seen in the density of $\hat{\theta}$ (Fig. 4d: right): While $\hat{\theta}$ is heavily concentrated near 1 for peaks-2, it is bimodal for peaks-1. Since clone-1 and clone-2 are differentiated by a copy-neutral event, this separation by peaks into two clusters is not driven by copy number differences but by chromatin remodeling. The pattern of overlap between the LOH-derived clones and the peaks-derived clusters indicate that, barring convergent evolution, most of the chromatin remodeling that led to the divergence of the peaks-2 cells must have occurred in the clone-2 lineage, after the chr4a LOH (Fig. 4e). Thus, Alleloscope analysis of this scATAC-seq data set allowed us to overlay two

subpopulations defined by peak signals with two subpopulations defined by a subclonal copy-neutral LOH, and infer their temporal order.

Figure 4.

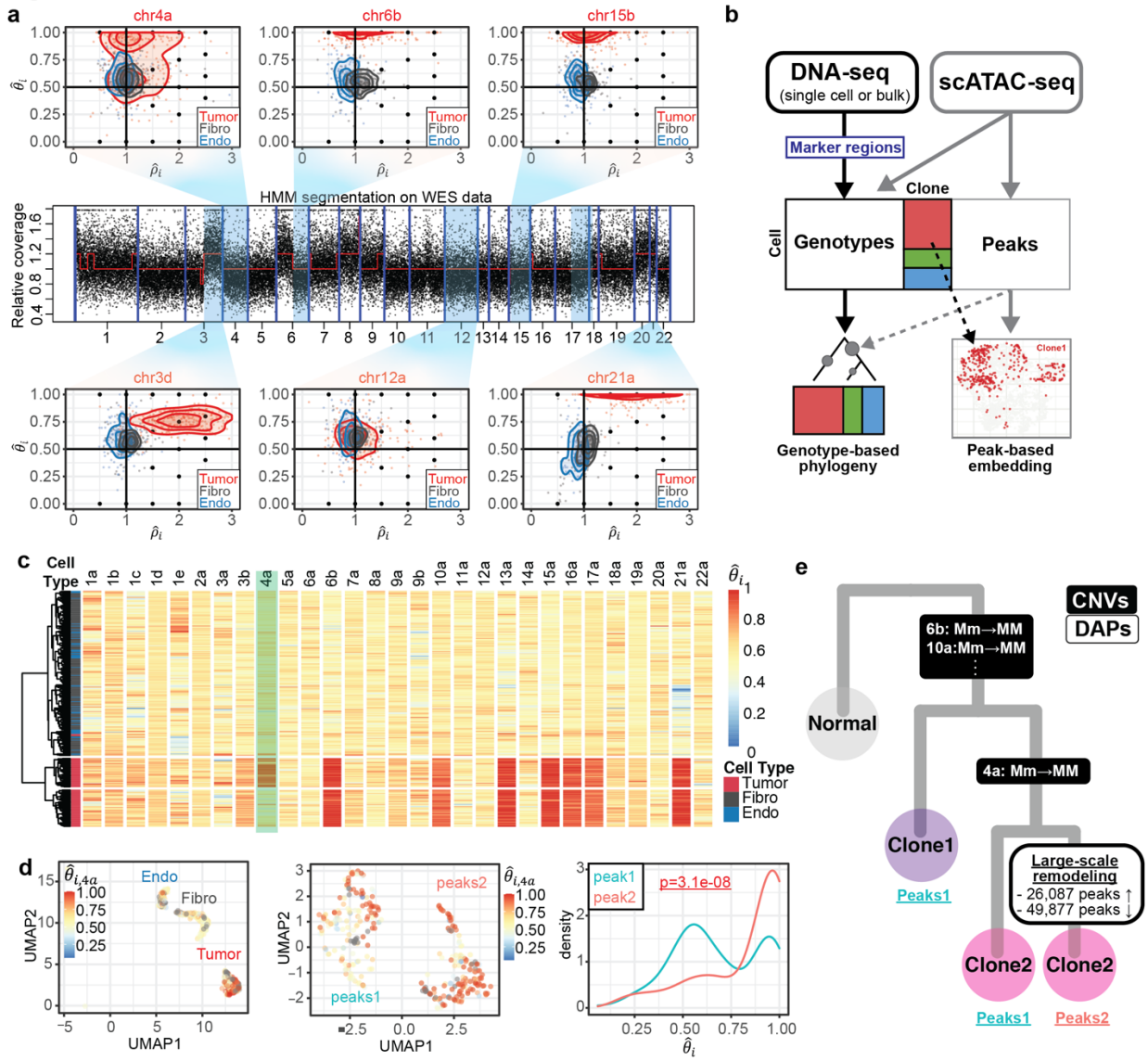

Fig. 4: Alleloscope multiomic analysis of scATAC-seq data of a basal cell carcinoma sample (SU008[23]). (a) Genotype profiles for six example regions for cells in scATAC-seq data. The regions are taken from segmentation of matched whole exome sequencing (WES) data. Each dot represents a cell-specific ($\hat{\rho}_i$, $\hat{\theta}_i$) pair. Cells are colored by annotation derived from peak signals[23], 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 chr4a major haplotype proportion ($\hat{\theta}_i$) and genome-wide peak profile. Left: UMAP projection of the 788 cells in the dataset by their genome-wide peak profile, colored by $\hat{\theta}_i$. The cell type annotation (endothelial, fibroblasts, and tumor cells) is labeled in the plot. Middle: UMAP projection of only the 308 tumor cells by their genome-wide peak profile shows two well-separated clusters: peaks1 and peaks2. Right: Density of $\hat{\theta}_i$ values for the peaks1 and peaks2 subpopulations. (e) Intratumor

heterogeneity of SU008 is shaped by a subclonal LOH of chr4a followed by subsequent genome-wide chromatin remodeling leading to three subpopulations: Clone 1 which does not carry the chr4a LOH (peaks cluster 1), Clone 2 carrying the chr4a LOH (peaks cluster 1), and remodeled clone 2 (peaks cluster 2).

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

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

First, we segmented the genome and estimated the allele-specific copy number profiles of single cells at each segment for both the scATAC-seq and scDNA-seq data, following the procedure in Figure 1 with some modifications due to the lack of normal cells to use as control for this sample. Figure 5a shows the relative total coverage, pooled across cells from scDNA-seq. Figure 5b shows $(\hat{\rho}, \hat{\theta})$-scatterplots for five example CNA regions in scDNA-seq and scATAC-seq. Compared to the scATAC-seq data, the scDNA-seq data has about 8-fold higher total read coverage and 7-fold higher heterozygous site coverage per cell. Thus, while subclones corresponding to distinct haplotype profiles are cleanly separated in the scDNA-seq data, they are much more diffuse in the scATAC-seq data. Yet, cluster positions in scATAC-seq roughly match those in scDNA-seq. As expected,

the $(\hat{\rho}, \hat{\theta})$-scatterplots reveal the high level of chromosomal instability in this sample, with each region exhibiting multiple clusters of different haplotype structures that indicate the existence of subclones carrying mirrored events and, for some regions, the variation of haplotype dosage over a gradient across cells.

Figure 5c shows the hierarchical clustering of cells from scDNA-seq based on their allele-specific copy number profiles, revealing the subclonal structure and the co-segregating CNA events that mark each subclone. For each cell in each region, Alleloscope also produces a confidence score for its assignment to different haplotype profiles (Supplementary Fig. 11). Based on visual examination of the confidence scores at the marker regions, we identified 6 subclones for further investigation (Clones 1-6 labeled at the right of the heatmap). The allele-specific copy number profiles allow us to manually reconstruct the probable evolutionary tree relating these 6 clones under the following three rules:

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

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

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

The evolutionary tree, thus derived, is shown in Figure 6b. The mirrored-subclonal amplifications on chr3q, the deletion on chr4p, and the multiallelic amplification on chr20q

allowed us to infer the early separation of clones 3-6 from clones 1-2. Subclones 3-6 are confidently delineated by further amplifications on chr3q, chr20q, chr11, chr13, and chr17. Note that high chromosomal instability led to concurrent gains of 1q and 7p in both the Clone 1-2 and Clone 3-6 lineages. We also observed a large number of low-frequency but high-confidence CNA events indicating that ongoing chromosomal instability in this population is spawning new sporadic subclones that have not had the chance to expand.

We now turn to scATAC-seq data, focusing on the 10 marker regions which, together, distinguish Clones 1-6: chr1b, 3b-d, 4b, 7a, 11b, 13b, and 20b-c. The $(\hat{\rho}, \hat{\theta})$ values computed by Alleloscope allows us to directly assign allele-specific copy number profiles to each cell for each region, as well as subclone labels to each cell, with posterior confidence score. The subclone assignment utilizes a Bayesian mixture model that pools information across the 10 marker regions. Despite the low accuracy in per-region genotyping, when information is pooled across the 10 marker regions, 81.6% of the 2,753 cells after filtering can be assigned to a subclone with >95% posterior confidence (Supplementary Fig. 12, the number of ATAC cells confidently assigned to each clone are shown in Figure 6a.). These subclone assignments for each cell, and cell-level haplotype profiles for each region, can now be integrated with peak-level signals.

Figure 5.

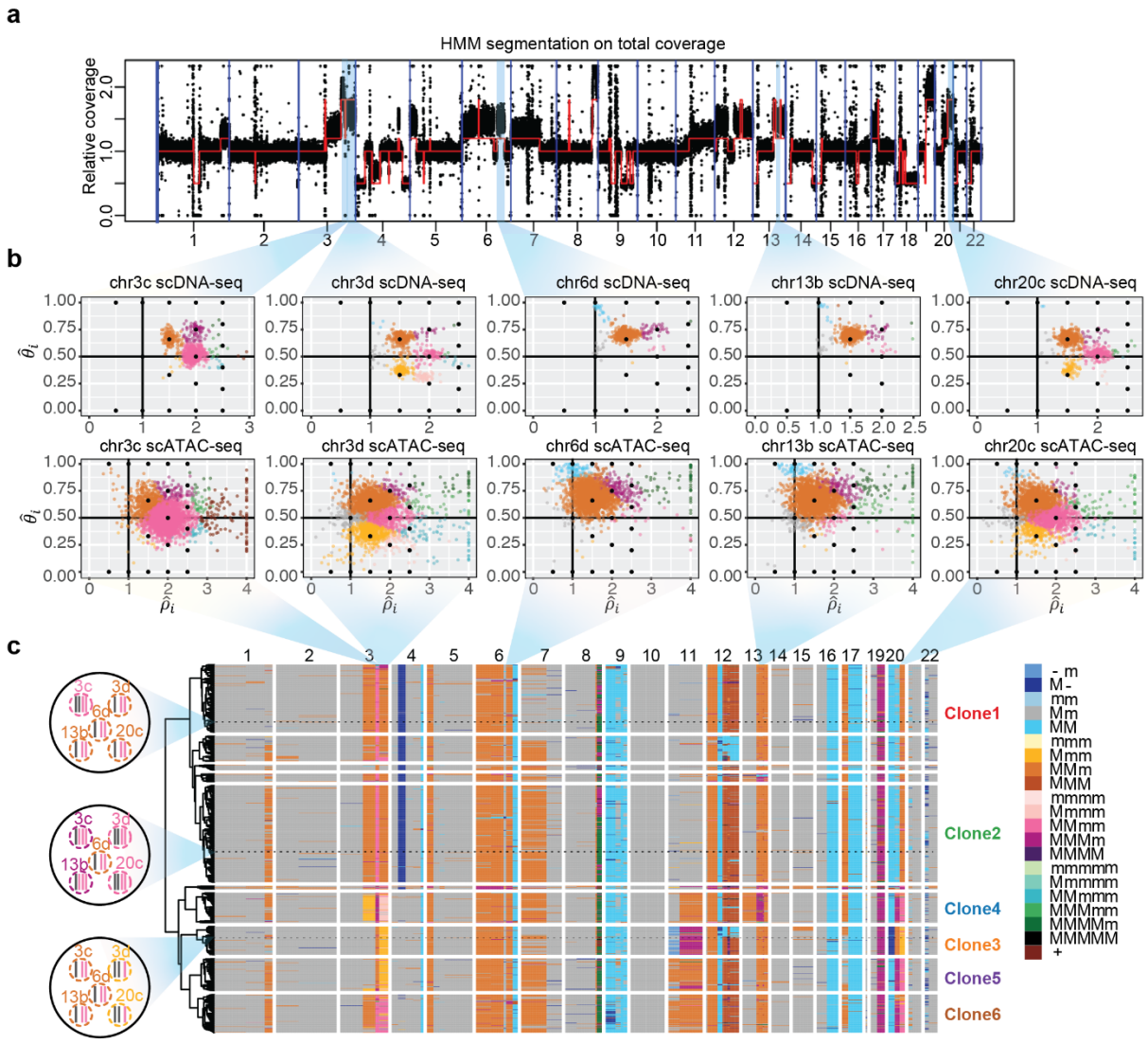

Fig. 5: Alleloscope analysis of scDNA-seq and scATAC-seq data reveals complex subclonal heterogeneity in the SNU601 gastric cancer cell line. (a) Genome segmentation using HMM on the pooled total coverage profile computed from scDNA-seq data.
(b) Single cell allele-specific copy number profiles $(\hat{\theta}, \hat{\rho})$ for five regions in scDNA-seq and scATAC-seq data. Cells are colored by haplotype profiles according to legend in Figure 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 Figure 5b, for three example cells, are shown in the left. The haplotype structures for the 5 regions in Figure 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.

Following the scheme in Figure 4b, we computed the Uniform Manifold Approximation and Projection (UMAP) coordinates for the scATAC-seq cells based on their peak profiles, which gives a two-dimensional visualization of the geometry of the chromatin accessibility landscape of this sample (Fig. 6a). UMAP scatterplots colored by clone assignment show that the 6 clones exhibit marked differences in their chromatin accessibility profiles (Fig. 6a): While Clone 1 and Clone 2 are concentrated at the top half of the UMAP, Clones 3-5 are positioned almost exclusively at the bottom half. Clone 6, which exhibits more variance, is also significantly enriched at the bottom half of the UMAP. Among Clones 3-5, Clone 3 has a distinct chromatin accessibility profile that is mostly concentrated at the bottom tip, Clone 4 is positioned higher, while Clone 5 contains cells that are similar to both clones 3 and 4. We expect some of these peak-level differences to be driven by CNAs.

To delineate the peaks that differ between clones, and to distinguish peak differences that are not accountable by CNAs, we identified differential accessibility peaks (DAPs) across each split of the tree (Fig. 6b) by performing pairwise Chi-square tests for peak enrichment between the cell populations on the two branches. The DAPs are categorized into two groups —1. DAPs lying in CNA regions for which the direction of change aligns with the direction of change of DNA coverage, and 2. DAPs not in CNA regions and DAPs in CNA regions that don't align in directionality of change with DNA coverage. The number of DAPs in both groups are shown along each branch (Fig. 6b). For the smaller subclones (Clone 3,4,5), low coverage limits the detection power and thus limits the DAP counts in both categories. Yet, juxtaposing DAP and CNA events along the tumor phylogeny yields insights: Along most lineages, a significant proportion of DAPs are attributable to CNAs

(p-values shown along each branch), and CNA events drive a substantial 36.3% of all of the DAPs identified. This argues for the importance of CNAs as a mechanism underlying subclonal differences in chromatin accessibility in this tumor.

Nevertheless, along some branches we find a large number of DAPs not attributable to broad CNAs, and thus must be due to other mechanisms. Two example DAPs of this latter category are shown as insets in Figure 6b, with full list given in Supplementary Table 1. The first example is a peak at the transcription start site (TSS) of the REC8 gene, which is located on chr14 where no apparent CNAs were observed across the six major subclones. The TSS of REC8 is open in clones 3-6 but closed in clones 1-2 (p-value<0.0001). REC8 is a gene encoding a meiosis-specific cohesion component that is normally suppressed in mitotic proliferation, and its role in cancer has recently gained increasing attention and controversy: While Yu et al.[38] found the expression of this gene to suppress tumorigenicity in a gastric cancer cell line, McFarlane et al.[39] postulated that it may be broadly activated in some cancers where it generates LOH by reductional segregation. The opening of the TSS of REC8, stably maintained in Clones 3-6, suggests that meiotic processes may underlie the increased chromosomal instability of this multiclonal lineage. The second example is a peak at the TSS of the WWOX gene, located on chr16, which is significantly depleted in Clone 3 (p-value<0.0001). Although chr16 has LOH across all tumor cells, there are no detectable subclonal differences, and thus we don't expect the decrease in accessibility at WWOX for subclone 3 to be due to a large copy number event. Since WWOX is a well-known tumor suppressor whose down-regulation is associated with more advanced tumors[40, 41], its decrease in accessibility suggests a more aggressive phenotype for Clone 3. Overall, these two examples show

how Alleloscope can be used to dissect the roles of CNA and chromatin-level changes in the identification of gene targets for follow-up study.

Figure 6.

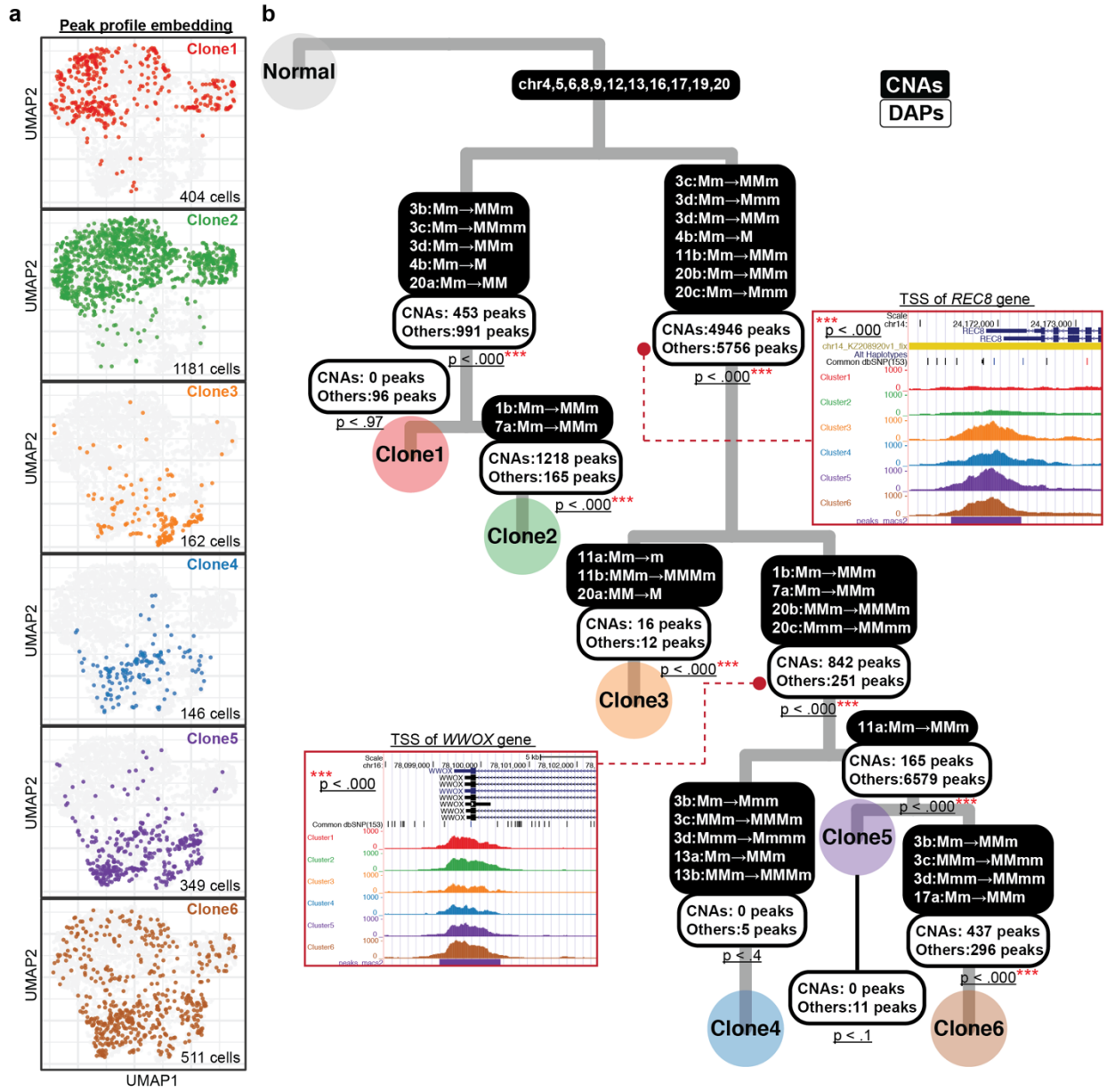

Fig. 6: Integrative analysis of allele-specific copy number and chromatin accessibility for SNU601 ATAC sequencing data. (a) UMAP projection of genome-wide scATAC-seq peak profile on 2,753 cells. The same group of cells were clustered into one of the six subclones based on their allele-specific copy 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 alternations (CNAs) and differential 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.

Discussion

Despite the recent advances in the application of single cell sequencing to cancer, we are still far from understanding the diversity of genomes that are undergoing selection at the single cell level. Notably, little is yet known about the intratumor diversity of allelic configurations within CNV regions, and to what extent the diversity of cells in chromatin accessibility can be attributed to diversity in allele-specific copy number. We presented Alleloscope, a new method for allele-specific copy number estimation that can be applied to single cell DNA and ATAC sequencing data (separately or in combination). First, on scDNA-seq data of 9 samples from 3 different tumor types, with phasing validation by linked-read sequencing on three samples, Alleloscope revealed an unprecedented level of allelic heterogeneity within hypermutable CNA regions. In these regions, subclones reside on a gradient of allelic ratios that is unobservable in total copy number analysis. In simple cases, these hypermutable regions contain mirrored subclones, as previously identified[9,10], but are often much more complex. We observed multiple instances of recurrent CNA events, some verified by linked read sequencing, where the same region is mutated multiple times during the evolution of the tumor, arriving at the same haplotype profile in distinct clones. In accordance with the findings in Watkins et al.[42], we found using Alleloscope that chromosomal instability drives the formation of subclones not only in primary tumors but also after metastasis.

Having established the allelic complexity of CNAs at single cell resolution, we next applied Alleloscope to scATAC-seq data, thus enabling the combined study of clonal evolution and chromatin accessibility. First, we considered the analysis of a public data set consisting of two basal cell carcinoma samples, for which matched bulk whole-exome

sequencing data was used for initial genome segmentation upon which single cell CNA genotyping was then conducted in the scATAC-seq data. Here we showed that Alleloscope can detect amplifications, deletions, and copy-neutral LOH events accurately in scATAC-seq data, and was able to find a subclone delineated by a copy-neutral LOH event. Juxtaposing this subclone assignment with peak signals allowed us to detect a wave of genome-wide chromatin remodeling in the lineage carrying the LOH. Next, we applied Alleloscope to a complex polyclonal gastric cancer cell line with matched scDNA-seq data. We found, by overlaying peak signals with subclones delineated by allele-specific copy number estimates, that much of the intratumor heterogeneity in chromatin accessibility can be attributed to CNAs. Focusing on subclone-enriched peaks outside of CNA regions allowed the prioritization of genes for downstream follow-up.

Alleloscope can potentially be applied to the integration of single cell data of other modalities, for example scATAC-seq and scRNA-seq data, to investigate the relationships between clonal evolution, chromatin remodeling, and transcriptome. To facilitate experimental design for single cell omics sequencing protocols, we investigated the performance of Alleloscope under different scenarios (number of cells, total per cell coverage, and total coverage at heterozygous SNP sites), see Supplementary Methods. As expected, accuracy is a function of all three quantities (Supplementary Fig. 13). Coverage at heterozygous SNP sites is especially important for scRNA-seq and scATAC-seq data, for which shifts in total coverage is an unreliable proxy for underlying DNA copy number. For scATAC-seq, the lower heterozygosity within peak regions led to lower number of reads mapping to heterozygous loci as compared to scDNA-seq, and this resulted in noisier subclone detection. Most of the current scRNA-seq technologies only

sequence either the 3' or 5' end of the mRNA transcripts, which limits the number of heterozygous SNP sites covered by reads. The latest developments in single cell long read sequencing[43-45] and single cell multimodal sequencing[46] herald new analysis opportunities with this method.

Methods

ScDNA-seq Data Sets and Pre-processing

Table 1 summaries the nine 10x scDNA-seq samples analyzed in this study:

Sample	Cancer type	Source	Paired normal	Linked -reads	Coverage per cell	Cell number	
P5846	Gastric	Primary tissue	Yes	No	454,806	510	[33]
P5847	Gastric	Primary tissue	Yes	No	422,134	715	[33]
P5915	Colorectal	Liver meta	Yes	No	126,2629	233	[33]
P5931	Gastric	Primary tissue	Yes	Yes	730,932	796	[12]
P6198	Colorectal	Liver meta	Yes	Yes	532,343	2,271	[33]
P6335	Colorectal	Omentum meta	No	Yes	564,058	953	[33]
P6461	Colorectal	Liver meta	Yes	No	483,524	1,242	[33]
SNU601	Gastric	Ascites meta	No	No	565,648	1,531	[12]
BC10x	Breast	Primary tissue	No	No	781,506	1,916	[10]

The Cell Ranger DNA pipeline (https://support.10xgenomics.com/single-cell-dna/software/) automates sample demultiplexing, read alignment, CNA calling and visualization. We first applied the tool to process the sequencing data (beta version: 6002.16.0) using the GRCh38 reference genome. The output bam files from the tool contain all information for later analysis. If the tumor samples had a matched normal sample, the GATK HaplotypeCaller was used to reliably call heterozygous SNPs on the matched normal samples. Otherwise, SNPs were retrieved on the tumor sample themselves. Next, we applied VarTrix, a software tool for extracting single cell variant information from the 10x barcoded bam files (https://github.com/10XGenomics/vartrix), to

efficiently generate two SNP-by-cell matrices for both reference alleles and alternative alleles of the SNPs called in the previous step.

To include high-quality SNPs in the later analysis, we filtered out the SNPs with <5 reads for P5846 and P5847, <10 reads for P5915 and P5931, <15 reads for P6335 and P6461, <20 reads for P6198 and SNU601, and <40 for BC10X samples based on the number of SNP detected for each sample. Additionally, SNPs located in the regions of repetitive sequences such as centromeres and telomeres were excluded. To exclude cells that might undergo apoptosis or cell cycles, the cells labeled noisy from the metadata output by the Cell Ranger tool were excluded.

Single-cell ATAC Data sets, Sequencing and Preprocessing

Table 2 summaries the scATAC-seq samples analyzed in this study:

Sample	**Cancer type**	**Source**	**Matched DNA**	**Coverage per cell**	**Cell number**	**Ref**
SU006	Basal cell carcinoma	Primary tissue	Yes	41,368	2771	[23]
SU008	Basal cell carcinoma	Primary tissue	Yes	36,057	788	[23]
SNU601	Gastric	Ascites meta	Yes	73,845	3614	-

The scATAC-seq dataset for the SNU601 sample was generated in this study. About 400,000 cells were washed with RPMI media and centrifuged (400g for 5 min at 4°C) twice. The supernatant was removed and chilled PBS + 0.04% BSA solution was added. The resuspended pellet was added to a 2ml microcentrifuge tube and centrifuged (400g for 5min at 4°C). After removing the supernatant without disrupting the pellet, 100 μL of chilled Lysis Buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.1%

Nonidet P40 Substitute, 0.1% Tween-20 and 0.01% digitonin) was added and carefully mixed 10 times. The tube was incubated on ice for 7 min. After incubation, 1 mL of chilled Wash Buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA and 0.1% Tween-20) was added and mixed 5 times followed by centrifugation of nuclei (500g for 5 min at 4°C). After removing the supernatant carefully, nuclei were resuspended in chilled Nuclei Buffer (10X Genomics), filtered by Flowmi Cell Strainer (40uM) and counted using a Countess II FL Automated Cell Counter. Then the nuclei were immediately used to generate scATAC-seq library.

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

Raw sequencing reads of the SNU601 scATAC-seq sample was de-multiplexed with the 10x Genomics Cell Ranger ATAC Software (v.1.2.0; https://support.10xgenomics.com/single-cell-atac/software/pipelines/latest/algorithms/overview) and aligned to the human GRCh38 reference genome. The aligned scATAC-seq data of the two pre-treatment basal cell carcinoma samples (SU006 and SU008) were downloaded from the Gene Expression Omnibus under accession GSE129785[23]. To obtain all potential SNPs for the SU006 and SU008 samples, GATK Mutect2 was used to call all single-nucleotide variants (SNVs) on the deduplicated bam files by the Picard toolkits of both the t-cell dataset and the tumor dataset from the same tumor. All SNVs from the paired tumor-normal datasets were

combined and the read counts of these SNPs were quantified for each cell in the tumor scATAC-seq dataset. The pre-filtered cell barcodes for the two public scATAC-seq datasets were retrieved from the previous study[23]. For the SNU601 scATAC-seq data, we instead quantified the read counts of the two alleles of the SNPs more reliably called from the paired normal scDNA-seq data. Like scDNA-seq, we applied VarTrix to generate two SNP-by-cell matrices for both reference alleles and alternative alleles of all the SNVs for all the scATAC-seq datasets. To obtain a SNP set including only SNVs that are more possible to be germline SNPs, we further filtered out the SNVs <20 reads for the SU008 sample and <30 reads for the SU006 sample. SNPs with extreme VAF values <0.1 or >0.9 were also excluded for both samples. Since we used the phasing information from the paired scDNA-seq data to assist the estimation of the haplotype structures for the SNU601 scATAC-seq data, we instead filtered out the cells <5 reads and the SNPs <5 reads to improve quality of the downstream analysis.

Linked-reads sequencing and data processing

The three samples with the linked-reads sequencing data were acquired as surgical resections following informed consent under an approved institutional review board protocol from Stanford University. Samples were subjected to mechanical and enzymatic dissociation as previously described, followed by cryopreservation of dissociated cells (DOI: 10.1101/2020.09.01.273672, 10.1158/1078-0432.CCR-19-3231)

Cryofrozen cells were rapidly thawed in a bead bath at 37 ºC. Cell counts were obtained on a BioRad TC20 cell counter (Biorad, Hercules, CA) using 1:1 trypan blue dilution. Between 1.5-2.5 million total cells were washed twice in PBS. Centrifugation was carried

out at 400g for 5 minutes. PBS was removed and cell pellets were frozen at -80 ºC. DNA extraction was carried out on cell pellets following thawing using either MagAttract HMW DNA Kit (P5931) or AllPrep DNA/RNA Mini Kit (Qiagen Inc., Germantown, MD, USA) as per manufacturer's protocol. Quantification was carried out using Qubit (Thermofisher Scientific).

Sequencing libraries were prepared from DNA using Chromium Genome Reagent Kit (v2 Chemistry) (10X Genomics, Pleasanton, CA, USA) as per manufacturer's instructions. Sequencing was performed using Illumina HiSeq or NovaSeq sequencers using 150x150 bp paired end sequencing and i7 index read of 8 bp. Long Ranger (10X Genomics) version 2.2.0 was used to perform read alignment to GRCh38, calling and phasing of SNPs, indels and structural variants.

Segmentation

The first step of Alleloscope is to segment the genome into regions with different CNA profiles. The appropriate segmentation algorithm depends on what samples are available. First, matched bulk DNA sequencing data (WGS/WES) or pseudo-bulk data from scDNAseq data can be segmented using FACLON[5], a segmentation method that jointly models the bulk coverage and bulk VAF profiles, if a matched normal sample is available.To accommodate segments from rare subclones, methods that integrate shared cellular breakpoints in CNA detection for scDNA-seq data such as SCOPE[47] can improve sensitivity. Since FALCON requires a matched normal sample or a sufficiently large set of normal cells, if these are not available then Alleloscope instead relies on a HMM-based segmentation method. The HMM method, which operates on the binned counts of pooled

cells, assumes a Markov transition matrix on four hidden states representing deletion, copy-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 $t = 1 \times 10^{-6}$ as default. Emission probabilities follow a normal distribution with means equal to {1.8, 1.2, 1, 0.5} and standard deviations equals to 0.2. All the scDNA-seq samples were segmented using the HMM algorithm. With the paired sample, the P6198 tumor sample was segmented using FALCON on the 1,399,650 SNPs >30 reads across 2,271 cells with all the default parameters.

Whole-exome sequencing (WES) data processing

The WES data of the two paired tumor-normal samples (SU006 and SU008) were obtained from the Sequence Read Archive under accession PRJNA533341. Raw fastq files were aligned to the GRCh37 reference genome using bwa-mem[48] with duplicate reads removed using the Picard toolkits[49]. The copy number calls of paired normal-tumor samples were obtained using Varscan 2[50]. To perform allele-specific copy number analysis on the WES using FALCON, GATK HaplotypeCaller[49] was used to call SNPs on both tumor and normal samples. Then FALCON was used to segment each chromosome based on the read counts of the reference alleles and alternative alleles of the SNPs overlapped between the paired tumor-normal samples.

SNP Phasing and Single-cell Allele Profile Estimation

For each region 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 all scDNA-seq and scATAC-seq data sets. Recall that by “major haplotype” we refer to 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_i denote major haplotype proportion of cell i. The EM model iterates the expectation step and the maximization step. The complete log likelihood of the model is

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

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

where A_{ij} and B_{ij} are the observed read counts for the reference and alternative alleles of cell i on SNP j. In the E-step, we first calculate the expected value of the posterior probability of the hidden variable I_j to construct a lower bound for optimization

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

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

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

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

Where $\hat{\theta}_i^{(t)}$ and $\hat{\theta}_i^{(t+1)}$ are from two successive iterations of EM. The two steps are iterated until converge. To speed up the EM process, we limited the maximum number of SNPs in a region to be 30,000 in our analysis. For the SNU601 scATAC-seq dataset, since the phase were estimated in the paired scDNA-seq dataset with higher depth, we directly applied the estimated $\hat{I}_j$'s from scDNA-seq data to estimate the $\hat{\theta}_i$'s of the cells in the scATAC-seq data. To improve the estimation results, cells with <20 read counts covering the identified SNPs were excluded for each region.

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

Let r represent a region in the genome after segmentation. To compute the relative coverage change for each cell in region r ($\hat{\rho}_{ir}$), normal cells and diploid regions identified within the sample are required for normalization. After major haplotype proportions for each cell in each region $\hat{\theta}_{ir}$'s are inferred from the EM-based algorithm, the estimates are used to identify normal cells and diploid regions under a hierarchical clustering of all cells. To identify normal cells, the dendrogram tree is first cut into k largest groups (we used $k = 5$ which worked well across samples). The cluster with normal cells is identified by selecting the cth cluster with the minimum distance calculated by

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

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

Putative diploid regions are next identified in each cluster. Similar to normal cell identification, Alleloscope computes the first measurement (d_{cr}) as the sum $\hat{\theta}_i$ distance of the cells in the cth cluster for each region r

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

Since amplified regions with both haplotypes equally amplified can also have small sum $\hat{\theta}_i$ distance, adjusted raw coverages are also considered in diploid region selection. The adjusted raw coverage of cell i in region r ($\tilde{\rho}_{iR}$) is computed by

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

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

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

where S_c here represents $\tilde{\rho}_i$ values of the cells in the cth cluster. To identify diploid regions, d_{cr} and m_{cr} are both ranked from the smallest to the largest for each cluster c. Alleloscope shows a list of potential diploid regions for each cluster by raking the sums of d_{cr} ranks and m_{cr} ranks. Excluding the cth cluster identified as the normal group,

Alleloscope proposed a list for the candidate diploid regions across the clusters by selecting the majority region.

Since coverage on scATAC-seq data is confounded by the epigenetic signals, chromosome 22 for SU008 and chromosome 18 for SU006 were directly selected as normal regions based on the WES data. Individual cells were classified into normal and tumor cells based on the epigenetic signals on the scATAC-seq data. For the SNU601 scATAC-seq dataset, chromosome 10 was selected as the normal region based on the paired scDNA-seq data.

Cell-level Genotyping

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

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

where N_{ir} is total read counts in region r and N_{i0} is total read counts in a reference region of cell i. N_{0r} is a vector denoting total read counts in region r of all identified normal cells and N_{00} is a vector denoting total read counts in the same reference region r of all identified normal cells. Since SNU601 is a tumor cell line with no normal cells in the dataset, N_{0r} and N_{00} were calculated from the cells in the matched normal P6198 sample as a substitute for the scDNA-seq data. For SNU601 scATAC-seq data, we aligned the

distribution of the $\hat{\rho}_{ir}$ values in paired scDNA-seq data to the distribution of the $\frac{N_{ir}}{N_i}$ values for each region to get the normalized $\hat{\rho}_{ir}$ in the scATAC-seq data. The normalized $\hat{\rho}_{ir}$ values for the scATAC-seq data were computed by

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

Next, cells with extreme $\hat{\rho}_{ir}$ values larger than the 99th percentile and smaller than the first percentile across the cells are considered outliers and excluded for each region. With the $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ pairs, cells in the scDNA-seq data can be classified into the haplotype profiles (g) with the expected (ρ_g, θ_g) values based on minimum distance. Although signals in the scATAC-seq data are much noisier, the haplotype structures identified in the paired scDNA-seq data can help to guide the genotyping for each region. In region r, the posterior probability of cell i carrying a haplotype profile observed in region r in the paired scDNA-seq data was

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

where g_r denotes the haplotypes profiles observed in region r in the paired scDNA-seq data and π_{g_r} denotes the prior probability that a randomly sampled cell carrying the g_r haplotype profile. A uniform prior can be used for π_{g_r} in the absence of external information. In the formula,

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

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

Validations using paired linked-reads sequencing data

We validated our algorithm using paired linked-reads sequencing data with two strategies in one gastric cancer patient sample and two colorectal cancer patient samples. First, the phasing accuracy was assessed by comparing the estimated SNP phases on the scDNA-seq data and the known phases of the same SNPs from the linked-reads sequencing data in individual regions. In the linked-reads sequencing data, SNPs within the same phase sets are phased with respect to one another, while those between different SNP sets are not. Therefore, we compared the phases of the SNPs overlapping between our estimated SNP set and the phase set with the largest numbers of SNPs in the linked-reads sequencing data for each region. The reference alleles of the overlapping SNPs with $\hat{I}_j$ >0.5 are estimated to be on the major haplotype. Otherwise, the reference alleles of the overlapping SNPs with $\hat{I}_j$ <0.5 are estimated to be on the minor haplotype. The SNPs with $\hat{I}_j$=0.5 are excluded. By comparing estimated phases and known phases from the linked-reads sequencing data of the overlapping SNPs, the phasing accuracy was computed for each region.

Secondly, we evaluated the genotyping accuracy by comparing the estimated haplotype profiles of each cell and the haplotype profiles inferred from the linked-reads sequencing data in individual regions. In the linked-reads sequencing data, the phase set with the largest numbers of SNPs was selected. The known phases of the overlapping SNPs between the phase set and the estimated SNPs were used to infer θ_i for each cell. Cell-level haplotype profiles using θ_i's from linked-reads sequencing data were considered as gold standard. By comparing the estimated haplotype profiles from $\hat{\theta}_i$ and the haplotype profiles from θ_i, genotyping accuracy was computed for each region. If the number of overlapping SNPs for an amplified region is smaller than 5000, the phase sets were combined from the largest to the smallest to reduce variance of θ_i's inferred from linked-reads sequencing data. The estimated SNP phases ($\hat{I}_j$) were used as templates to combine separate phase sets.

Cell Lineage Reconstruction

To investigate the tumor subclonal structure for scDNA-seq data, cell-specific haplotype profiles from Alleloscope across the genome were used to reconstruct cell lineage trees. The "Gower's distance" is calculated using "cluster" R package on the nominal haplotype profiles between cells. Then hierarchical clustering is performed on the distance using the "ward.D2" method. Since variance of $\hat{\theta}_i$'s are higher when fewer SNPs are located in a segment, we included the segments with more than 2,000 SNPs identified. The clustering result is visualized using the 'pheatmap' R package with the five largest clusters separated by marginal lines. Each segment was plotted with based on its length

proportional to 5,000,000 bins. The heights of the clustering tree were log-transformed for easier visualization.

The tumor subclonal structures were also investigated in the scATAC-seq data. Instead of using the haplotype profiles defined by the $(\hat{\rho}_{ir}, \hat{\theta}_{ir})$ pairs, the cells from the two public basal cell carcinoma were clustered using $\hat{\theta}_{ir}$ values, which are orthogonal to the peak signals based on total coverage, across the segments with more than 500 SNPs. Then hierarchical clustering is performed on the Euclidean distance using the "ward.D2" method and visualized using 'pheatmap' R package with the three largest clusters separated by marginal lines. The heights of the clustering tree were log-transformed for easier visualization.

Since the subclones for the SNU601 sample were identified first from the scDNA-seq data, for this cell we adopted a supervised strategy to assign each cell into different subclones. First, we identified 10 marker regions-- chr1b, 3b-d, 4b, 7a, 11b, 13b, and 20b-c that help to differentiate the cells into the six major subclones based on the subclone specific copy number profiles from the scDNA-seq data. Combining the haplotype profiles across the ten regions for each cell enables assignment of the cells into one of the six subclones with higher confidence. The posterior probability of cell i coming from clone k was

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

where $k \in \{1{\sim}6\}$ for the six clones and π_k is the prior probability that a randomly sampled cell coming from the kth clone, which can be estimated from the paired scDNA-seq data

or set to uniform (in our analysis setting to uniform gives very similar results). In the formula,

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

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

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

ScATAC-seq data analysis

To investigate the relationships between allele-specific CNAsand chromatin accessibility, for each cell in scATAC-seq data we processed the peak signals in addition to the allele-specific CNAs. For the two public basal cell carcinoma samples, the peak by cell matrices was obtained from GSE129785. We subset the fragment counts for each peak in the cells from the SU008 sample, regressed out cell total coverage for each peak by linear regression, and projected the cells onto the UMAP plot using genome-wide peak signals[51]. The cell type identify for each cluster was retrieved from the labels in the previous study[23]. To further explore intratumor heterogeneity, we selected the cells labeled as tumor cells,

regressed out cell total coverage, and projected the tumor cells onto the UMAP plot like previously described. Then the DNA level information and epigenetic signals for each cell can be visualize and analyzed together.

For the SNU601 scATAC-seq dataset, scATAC-pro[52] was used to call peaks and generate the peak by cell matrix from the bam file and fragment file output by the Cell Ranger software. We first filtered out the cells that have proportions of fragments on the detected peaks <0.4 and or total peaks outside of the range 15,000~100,000, and filtered out the peaks observed in less than 0.1 of cells. Next, we regressed out cell total coverage for each peak by linear regression, and projected the cells onto the UMAP plot using genome-wide peak signals. Then the clonal assignment based on the DNA information and the peak signals can be integrated at the single-cell level.

Based on the lineage structure from the paired scDNA-seq data, the cells can also be placed in the lineage tree based on their clonal assignment. Under the lineage structure, pairwise comparison using Chi-squared test was performed on the proportion of the to identify differential accessible peaks (DAPs) for each branch. A peak was considered a DAP if the FDR adjusted p-values<0.05. Since copy number alternations are confounding factors that also affect the peak signals, the DAPs were further divided into two groups—1. “CNA” group if the DAPs are in the CNA regions and both signals are positive correlated; 2. “Other” group if the DAPs are not categorized in the first group. A set of DAPs were considered to be enriched in the CNA regions if the p-values<0.05 under the hypergeometric test. This type of analysis enables investigation of the relationships between the two signals. Each DAP was further mapped to the genes that are potentially regulated based on the $\pm$ 2,000 bp distance on the genome. To further visualize the

difference of the peak signals among the six clones, the peak signals were pooled across the cells and normalized by the total cell number in each subclone.

Data availability

All the linked-reads sequencing data and the scATAC-seq dataset are available under accession ###. There are no restrictions on data availability or use. The patient scDNA-seq data were obtained from dbGAP under accession phs001818.v3.p1[33] (all except 5931 scDNA) and phs001711[12] (5931 scDNA). The cell line scDNA-seq dataset was from the Sequence Read Archive (SRA) under accession PRJNA498809. The public scATAC-seq data and whole exome sequencing data were obtained from the SRA under accession PRJNA532774[23] and PRJNA533341[37].

Code availability

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

Reference

1. Baylin, S.B. & Jones, P.A. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer* **11**, 726-734 (2011).
2. Sandoval, J. & Esteller, M. Cancer epigenomics: beyond genomics. *Curr Opin Genet Dev* **22**, 50-55 (2012).
3. Greaves, M. & Maley, C.C. Clonal evolution in cancer. *Nature* **481**, 306-313 (2012).
4. Burrell, R.A., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338-345 (2013).

5. Chen, H., Bell, J.M., Zavala, N.A., Ji, H.P. & Zhang, N.R. Allele-specific copy number profiling by next-generation DNA sequencing. *Nucleic Acids Res* **43**, e23 (2015).

6. Favero, F. et al. Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann Oncol* **26**, 64-70 (2015).

7. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881-1893 (2014).

8. Shen, R. & Seshan, V.E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**, e131 (2016).

9. Jamal-Hanjani, M. et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med* **376**, 2109-2121 (2017).

10. Zaccaria, S. & Raphael, B.J. Characterizing allele- and haplotype-specific copy numbers in single cells with CHISEL. *Nat Biotechnol* (2020).

11. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U S A* **107**, 16910-16915 (2010).

12. Andor, N. et al. Joint single cell DNA-seq and RNA-seq of gastric cancer cell lines reveals rules of in vitro evolution. *NAR Genom Bioinform* **2**, lqaa016 (2020).

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

14. Garvin, T. et al. Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* **12**, 1058-1060 (2015).

15. Kim, C. et al. Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893 e813 (2018).

16. Laks, E. et al. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell* **179**, 1207-1221 e1222 (2019).

17. Navin, N. et al. Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90-94 (2011).

18. Velazquez-Villarreal, E.I. et al. Single-cell sequencing of genomic DNA resolves sub-clonal heterogeneity in a melanoma cell line. *Commun Biol* **3**, 318 (2020).

19. Wang, Y. et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155-160 (2014).

20. Corces, M.R. et al. The chromatin accessibility landscape of primary human cancers. *Science* **362** (2018).

21. Granja, J.M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat Biotechnol* **37**, 1458-1465 (2019).

22. Litzenburger, U.M. et al. Single-cell epigenomic variability reveals functional cancer heterogeneity. *Genome Biol* **18**, 15 (2017).

23. Satpathy, A.T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925-936 (2019).

24. Schep, A.N., Wu, B., Buenrostro, J.D. & Greenleaf, W.J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975-978 (2017).

25. Cabal-Hierro, L. et al. Chromatin accessibility promotes hematopoietic and leukemia stem cell activity. *Nat Commun* **11**, 1406 (2020).

26. Dravis, C. et al. Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. *Cancer Cell* **34**, 466-482 e466 (2018).

27. Guilhamon, P. et al. Single-cell chromatin accessibility in glioblastoma delineates cancer stem cell heterogeneity predictive of survival. *bioRxiv*, 370726 (2020).

28. Pan, D. et al. A major chromatin regulator determines resistance of tumor cells to T cell-mediated killing. *Science* **359**, 770-775 (2018).

29. Qu, K. et al. Chromatin Accessibility Landscape of Cutaneous T Cell Lymphoma and Dynamic Response to HDAC Inhibitors. *Cancer Cell* **32**, 27-41 e24 (2017).

30. Shaffer, S.M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431-435 (2017).

31. Shu, S. et al. Synthetic Lethal and Resistance Interactions with BET Bromodomain Inhibitors in Triple-Negative Breast Cancer. *Mol Cell* **78**, 1096-1113 e1098 (2020).

32. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).

33. Sathe, A. et al. The cellular genomic diversity, regulatory states and networking of the metastatic colorectal cancer microenvironment. *bioRxiv* (2020).

34. Bell, J.M. et al. Chromosome-scale mega-haplotypes enable digital karyotyping of cancer aneuploidy. *Nucleic Acids Res* **45**, e162 (2017).

35. Greer, S.U. et al. Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med* **9**, 57 (2017).

36. Zheng, G.X. et al. Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* **34**, 303-311 (2016).

37. Yost, K.E. et al. Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat Med* **25**, 1251-1259 (2019).

38. Yu, J. et al. REC8 functions as a tumor suppressor and is epigenetically downregulated in gastric cancer, especially in EBV-positive subtype. *Oncogene* **36**, 182-193 (2017).

39. McFarlane, R.J. & Wakeman, J.A. Meiosis-like Functions in Oncogenesis: A New View of Cancer. *Cancer Res* **77**, 5712-5716 (2017).

40. Aqeilan, R.I. et al. Loss of WWOX expression in gastric carcinoma. *Clin Cancer Res* **10**, 3053-3058 (2004).

41. Baryla, I., Styczen-Binkowska, E. & Bednarek, A.K. Alteration of WWOX in human cancer: a clinical view. *Exp Biol Med (Maywood)* **240**, 305-314 (2015).

42. Watkins, T.B.K. et al. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* (2020).

43. Gupta, I. et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. *Nat Biotechnol* (2018).

44. Lebrigand, K., Magnone, V., Barbry, P. & Waldmann, R. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nat Commun* **11**, 4025 (2020).

45. Singh, M. et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. *Nat Commun* **10**, 3120 (2019).

46. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nat Methods* **17**, 11-14 (2020).

47. Wang, R., Lin, D.Y. & Jiang, Y. SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. *Cell Syst* **10**, 445-452 e446 (2020).

48. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

49. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

50. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576 (2012).

51. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

52. Yu, W., Uzun, Y., Zhu, Q., Chen, C. & Tan, K. scATAC-pro: a comprehensive workbench for single-cell chromatin accessibility sequencing data. *Genome Biol* **21**, 94 (2020).

Acknowledgements

The work is supported by the National Institutes of Health [P01HG00205ESH to B.T.L., S.M.G. AND H.P.J., 5R01-HG006137-07 and 1U2CCA233285-01 to C-Y.W. and to N.R.Z.]. Additional support to HPJ came from the Research Scholar Grant, RSG-13-297-01-TBG from the American Cancer Society, Clayville Foundation and the Gastric Cancer Foundation.

Author contributions

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

Competing interests

The authors declare no competing interests.