KaryoTap Enables Aneuploidy Detection in Thousands of Single Human Cells

Joseph C. Mays; josephcmays@gmail.com; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Sally Mei; Sally.Mei2@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Manjunatha Kogenaru; Manjunatha.Kogenaru@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Helberth M. Quysbertf; Helberth.Quisberth@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Nazario Bosco; nazario.bosco.phd@gmail.com; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA. Current Address: Volastra Therapeutics, New York, NY 10027, USA.

Xin Zhao; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Joy J. Bianchi; Joy.Bianchi@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Aleah Goldberg; aleah.goldberg@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Gururaj Rao Kidiyoor; Gururajrao.Kidiyoor@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Liam J. Holt; liam.holt@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

David Fenyö; david.fenyo@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Teresa Davoli (corresponding author); teresa.davoli@nyulangone.org; Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, New York University Grossman School of Medicine, New York, NY 10016, USA

Abstract

Investigating chromosomal instability and aneuploidy within tumors is essential for understanding tumorigenesis and developing diagnostic and therapeutic strategies. Single-cell DNA sequencing technologies have enabled such analyses, revealing aneuploidies specific to individual cells within the same tumor. However, it has been difficult to scale the throughput of these methods to detect rare aneuploidies while maintaining high sensitivity. To overcome this deficit, we developed KaryoTap, a method combining custom targeted DNA sequencing panels for the Tapestri platform with a computational framework to enable detection of chromosome- and chromosome arm-scale aneuploidy (gains or losses) and copy number neutral loss of heterozygosity in all human chromosomes across thousands of single cells simultaneously. KaryoTap allows detecting gains and losses with an average accuracy of 83% for arm events and 91% for chromosome events. Importantly, together with chromosomal copy number, our system allows us to detect barcodes and gRNAs integrated into the cells' genome, thus enabling pooled CRISPR- or ORF-based functional screens in single cells. As a proof of principle, we performed a small screen to expand the chromosomes that can be targeted by our recently described CRISPR-based KaryoCreate system for engineering aneuploidy in human cells. KaryoTap will prove a powerful and flexible approach for the study of aneuploidy and chromosomal instability in both tumors and normal tissues.

Keywords

Aneuploidy, Chromosomal Instability, Single-Cell DNA Sequencing, Targeted DNA Sequencing, Cancer genomics, Tumor heterogeneity, Copy Number Variants, CNVs

Background

A critical hallmark of cancer initiation and progression is the presence of aneuploidy, or gains and losses of whole chromosomes or chromosome arms, which arise due to mitotic missegregation events(1–3). In addition to aneuploidy, chromosomal instability, characterized by a continuously high rate of these missegregation events among tumor cells, has gained particular interest as a potential driver of tumor progression and metastasis(4,5). Chromosomal instability produces cell populations with heterogeneous aneuploid karyotypes that continuously evolve over time, granting tumor cells an opportunity to adapt to their environment and develop resistance to cancer therapies(6–8). Traditional methods for detecting aneuploidy, such as whole genome sequencing (WGS), rely on bulk averaging of cell populations, effectively masking heterogeneity among individual cells and preventing proper assessment of the extent or rate of chromosomal instability. Single-cell approaches, particularly single-cell DNA sequencing (scDNA-seq), overcome this limitation and can instead detect the full complement of distinct karyotypes present in a tumor(9–13). These data can be used to reassemble the evolution of the tumor's cells, which can provide insights into how aneuploidy and chromosomal instability may drive tumorigenesis or inform treatments for therapeutic resistance(8,11). Methods for modeling aneuploidies of specific chromosomes in cell culture such as KaryoCreate, have also emerged as powerful tools for studying the effects of aneuploidy in cancer(14–16). These methods have benefitted from scDNA-seq, as sequencing individual cells enables the evaluation of the specificity and accuracy of the engineered karyotypes(15,16).

The fundamental challenge of scDNA-seq methods is faithfully and completely sequencing the 6 picograms of DNA within a cell. The uniformity and depth of sequencing coverage determine the sensitivity with which aneuploidy can be detected. Conventional whole-genome amplification (WGA) methods, DOP-PCR(17,18), MDA(19), and MALBAC(20) amplify the genome prior to sequencing, introducing amplification biases and PCR errors that confound results. Finally, these

methods rely on the partitioning of cells into individual wells or tubes, which limits throughput, preventing the cost-efficient sequencing of enough cells to identify rare aneuploidy events in a large population(21). More recent methods vastly improve throughput by using microfluidic partitioning(22), combinatorial indexing(23), and liquid handling robots(24), allowing hundreds of thousands of cells to be sequenced at once. However, these methods suffer from uneven amplification across the genome and require sufficiently deep sequencing per cell, jeopardizing the confident detection of aneuploidy in individual cells(25). Furthermore, the need for custom hardware, expensive liquid handlers, and complicated protocols makes these methods difficult to adopt for most laboratories(21). We note that in scDNA-seq methods based on (untargeted) WGS there is a natural bias in the sensitivity of detecting whole chromosome or arm-level gains or losses across chromosomes depending on their size. In fact, the sensitivity of aneuploidy detection depends on the total number of reads per chromosome thus is lower for smaller chromosomes compared to larger ones as they contain a smaller proportion of the total reads. However, for the purpose of evaluating aneuploidy (gains and losses of whole chromosomes or chromosome arms) and chromosomal instability (rate of chromosome missegregation), each chromosome counts as a single entity. Thus, scDNA-seq methods based on WGS necessitate a high number of reads (and thus increase in cost) to achieve sufficient sensitivity of detection across all chromosomes.

To address the need for a high-throughput method for detecting chromosome-scale aneuploidy across the human genome that maintains high sensitivity at a cost-efficient sequencing depth, we turned to the Tapestri platform from Mission Bio, a droplet-based targeted scDNA-seq solution that allows for the sequencing of hundreds of genomic loci across thousands of cells in one experiment(26). The platform is commonly used to detect tumor hotspot mutations in cancer driver genes and has not yet been utilized to identify aneuploidy (gains and losses of whole chromosomes or chromosome arms) across the genome. While targeted sequencing is typically

used to identify mutations in single nucleotides(27), we reasoned that we could use the relative sequencing depth of targeted loci to detect chromosome- and chromosome arm-scale aneuploidy in individual cells. Here, we describe KaryoTap, a method combining custom targeted Tapestri panels of PCR probes covering all human chromosomes with a Gaussian mixture model framework for DNA copy number detection, thus enabling the accurate detection of aneuploidy in all chromosomes in several thousand cells across different cell lines. Additionally, we included probes that detect lentiviral-integrated CRISPR guide RNAs to enable functional studies, and DNA barcodes to enable sample multiplexing. To enhance usability, we also developed a companion software package for R, *karyotapR*, which enables the straightforward copy number analysis, visualization, and exploration of the data produced by our custom panels. KaryoTap allows detecting gains and losses with an average accuracy of 83% for arm events and 91% for chromosome events. By overcoming the limitations of current methods, this system will be a valuable tool for investigating the evolution and consequences of aneuploidy and chromosomal instability in human tumors, in addition to other healthy and diseased tissues, such as normal tissues during physiological aging or clonal hematopoiesis of indeterminate potential.

Results

Design of a Custom Targeted Panel for Detecting Chromosome-Scale Aneuploidy

To detect DNA copy number across the human genome in single cells, we designed a custom panel (Version 1; CO261) for the Tapestri system comprising 330 PCR probes that target and amplify specific loci across all 22 autosomes and the X chromosome (**Fig 1; Table S1; Additional File 1;** for the Version 3 panel, the Y chromosome was also included; see below). The number of probes targeting each chromosome was proportional to the size of the chromosome (e.g., 24 probes for chr1, 5 for chr22) to achieve a roughly uniform density of ~1-2 probes per 100 megabases across all chromosomes. Loci were selected to cover regions carrying single

nucleotide polymorphisms (SNPs) with major allele frequencies of 0.5-0.6 such that cells from different lines or individuals sequenced in the same experiment could be identified by their distinct genotypes (see Methods).

Population-Level Aneuploidy Detection Across Human Cell Lines

The Tapestri system partitions individual cells into aqueous droplets and generates barcoded amplicons from the loci targeted by the supplied probe panel. Both cell-specific sequencing read counts and variant allele frequencies (VAFs) for each probe are generated by processing sequencing reads from these amplicons. To test whether our Version 1 panel could detect whole-chromosome aneuploidy in individual cells, we performed a Tapestri scDNA-seq experiment on a pool of five cell lines with varying karyotypes, mixed in equal proportion. The pool consisted of retinal pigment epithelial cells hTERT RPE-1 (hereafter RPE1; +10q, XX) as a near-diploid reference population(6,28), and four aneuploid colon cancer cell lines: LS513 (+1q, +5, +7, +9, +13, +13, XY), SW48 (+7, +14, XX), LoVo (+5p, +7, +12, XY), and CL11 (-6, +7, +7, -17, -18, -22, XY). The bulk (i.e., population-averaged) karyotypes for each cell line were determined by whole genome sequencing (WGS) (**Fig S1A**) and confirmed for RPE1 by G-banded karyotyping (**Fig S1B**).

Sequencing read counts and variant allele frequencies (VAFs) for 2,986 cells across the five cell lines were recovered from the Tapestri experiment. Dimensional reduction of the cells' VAFs by principal component analysis (PCA) and UMAP revealed 5 major clusters corresponding to the 5 cell lines (**Fig S1C**). The remaining smaller clusters, representing composites of VAF profiles from multiple cells captured in the same droplet, were discarded from further analyses. We estimate the copy number for each probe in each cell as the ratio of read counts relative to a reference population. Here, we used the RPE1 cells as the reference population, which have 2 copies

(diploid) of each chromosome, except for a third copy (triploid) of the chr10q arm translocated to the X chromosome (**Fig S1B**). To identify the cell cluster comprising RPE1 in our data, we calculated the mean VAF across variants for each cluster and compared them to VAFs from published deep WGS of RPE1(29) by PCA (**Fig S1D**). The cells whose mean VAFs clustered closest to those of published RPE1 represent RPE1 cells. The copy number estimates for each probe in each cell for all cells (hereafter, cell-probe scores) were then calculated by taking the ratio of normalized read counts to the median normalized read counts of the RPE1 reference population (see Methods).

The 330 target regions have varying base compositions (**Additional File 1**) and the probes have different optimal melting temperatures but are amplified under the same conditions and thermocycling parameters, introducing technical artifacts from amplification bias(30). The probe-level heatmap (**Fig S1E**) highlights such technical variation between copy number values from intra-chromosomal probes, suggesting that measurements from any individual probe are unlikely to reliably reflect a cell's copy number. To address this, we calculated a single copy number score for each chromosome in each cell (hereafter, cell-chromosome unit) by smoothing the cell-probe scores of all probes targeting the same chromosome. Smoothing was accomplished by calculating the weighted median of the cell-probe copy number values for all probes on a given chromosome; larger weights were assigned to probes whose copy number scores had smaller spreads (see Methods). Heatmap visualization of cell-chromosome copy number scores for the four colon cancer cell lines corresponds with the expected population-level copy number values from bulk WGS (**Fig 2A**). For example, the per-cell and average heatmap intensities indicate correctly that LS513, SW48, and LoVo carry 3 copies of chromosome 7, while CL11 carries 4 copies. Similarly, 1 copy of chromosome 6 could be detected in CL11, indicating a chromosomal loss, and single copies of chromosome X could be detected in LS513, LoVo, and CL11, indicating XY sex chromosomes.

scDNAseq is often used on tumor cells to characterize the copy number heterogeneity that arises from chromosomal instability during tumorigenesis (intratumoral heterogeneity; ITH)(31). Here, we can consider the entire dataset a model of a heterogenous tumor carrying several major subclonal lineages, with each cell line representing a distinct major subclone. Unbiased clustering correctly groups the subclones/cell lines by copy number score into their respective cell lines (**FIG S2A**). In a real tumor, this could be used to distinguish subclones. Furthermore, clustering cells from each line can reveal subclones occurring within a given line. Here, we show that clustering of LoVo cells reveals two subpopulations hallmarked by exclusive gains in chr5p or chr15q (**FIG S2B**). Altogether, these data demonstrate that our Version 1 panel can resolve the average copy number of populations of cells at the whole-chromosome level and distinguish major subclones within a heterogenous cell population using copy number.

Copy Number Estimation in Individual Cells

To account for variation in the distributions of copy number scores, we classified the scores as integer copy number calls using a 5-component Gaussian mixture model (GMM)(32), where each component represents a possible copy number value of 1, 2, 3, 4 or 5 (**Fig 2B**). Using the known copy number for each chromosome in RPE1 and the corresponding distributions of copy number scores, we simulated the expected distributions of copy number scores that would be measured from chromosomes with actual copy numbers 1-5 for each chromosome. We then used Bayes theorem to calculate the posterior probability of each cell-chromosome score belonging to each of the 5 GMM component distributions and assign the cell-chromosome an integer copy number corresponding to the component with the highest posterior probability (**Fig S2C)** (see Methods for details).

To evaluate copy number calling performance, we determined the accuracy of the classifier model, calculated as the proportion of correct calls (i.e., true positives), using the known copy numbers for RPE1 as ground truth. We focused on RPE1 as its karyotype is stable and homogeneous across the population(6); karyotyping confirmed that the line is triploid for chr10q and diploid for all other chromosomes in all metaphases analyzed, with the exception of 3 copies of chr12 in 3% of metaphases (**Fig S1B**). Accuracy, or correctly identifying 2 copies of a chromosome in RPE1, varied between chromosomes and ranged from 95% for chr2 to 49% for chr22 with a mean of 82% (**Fig 2C**). Because we used the RPE1 cells to both fit and test the GMMs, we performed 5-fold cross validation by partitioning them randomly into 5 equally sized subsets and calculated accuracy five times, each time reserving one of the sets from the model generation and using it only to calculate accuracy. The mean absolute deviation of the 5 accuracy measurements for each chromosome ranged from 0.61 to 4.59 percentage points, suggesting that copy number calling performance would be maintained when classifying new data. 18 out of 22 chromosomes had sensitivities of at least 75%; chr10 was excluded from the whole-chromosome analysis because the p and q arms have different copy numbers. Accuracy for each chromosome correlated strongly with chromosome length (Pearson r = 0.73), and chromosome length itself correlated strongly with the number of probes targeting the chromosome (Pearson r = 0.93) (**Fig S2D**), suggesting that classifier accuracy is related to the number of probes targeting a chromosome. As expected, linear regression of accuracy on the number of probes per chromosome indicated that the number of probes is predictive of copy number call accuracy (R^2 = 0.81; p = 1.12e-08) (**Fig 2D**). This suggests that the classifier accuracy for poorly performing chromosomes, particularly the smaller chromosomes including 19, 21, and 22, could be improved by adding additional probes for those chromosomes to the panel.

The empirical accuracy for copy number calls in RPE1 only demonstrates the ability of our method to detect 2 copies of a chromosome. We can determine the theoretical or expected sensitivity for

detecting copy numbers of 1, 2, 3, 4, and 5 for each chromosome by calculating the proportion of each GMM component distribution that would be called correctly as belonging to that component. Overall, theoretical sensitivity was highest for 1 copy with an average of 97%, decreasing with each additional copy; 2 copies had an average sensitivity of 83% and 3 copies had 64% (**Table S2**). The theoretical sensitivity for 2 copies strongly correlated with the empirical accuracy for RPE1 calls (Pearson r = 0.97). As expected, theoretical sensitivity at all 5 copy number levels decreased for chromosomes with fewer probes, as was the case with the empirical accuracy (**Fig 2E**).

WGS indicated aneuploidy restricted to one arm of a chromosome in RPE1 (chr10q), LS513 (chr1p), and LoVo (chr5p) (**Fig S1A**). To determine if our Version 1 panel could also detect chromosome arm-level aneuploidy as well, we performed a similar analysis by smoothing cell-probe copy number scores across probes targeting each chromosome arm instead of across whole chromosomes (**Fig 3A**). Per-cell and average heatmap intensities indicate correctly that LS513 carries 3 copies of chr1q and 2 copies of chr1p, and LoVo carries 3 copies of chr5p. We called integer copy numbers using a GMM generated for each chromosome arm and evaluated the classifier accuracy for correctly calling copy numbers in RPE1 (**Fig S2E)**. Accuracy ranged from 91% (chr8q) to 50% (chr19p, chr22q) with a mean of 73%. The mean absolute deviation of accuracy from 5-fold cross validation ranged from 0.63 to 6.2 percentage points. Only 22 out of 41 arms had accuracy values of at least 75% (**Fig 3B**), demonstrating generally lower accuracy compared to whole chromosomes, likely because fewer probes typically target an arm than an entire chromosome. A positive relationship between the number of probes and accuracy was again revealed by linear regression (R^2 = 0.65; p = 5.1e-09) (**Fig 3C**). We calculated the theoretical sensitivity for detecting arm-level copy number across the 5 copy number levels. Sensitivity was again highest for 1 copy with an average of 95%; 2 copies had an average of 74% and 3 copies had 53% (**Fig 3D; Table S2**). Overall, we found that our system can accurately call

copy numbers for the majority of chromosomes and several chromosome arms, with less sensitivity for smaller chromosomes and arms.

Downsampling the Number of Probes Decreases Accuracy

To confirm that copy number classification accuracy/sensitivity is dependent on the number of probes targeted to a chromosome and not the size of the chromosome itself, we downsampled the probes targeting chromosomes 2 (23 total probes) and 6 (18 total probes) and recalculated the classification accuracy for the RPE1 cells. 50 samples of *n* probes were evaluated for each value of *n*. Consistent with our findings above, median accuracy decreased from 95.8% (24 probes) to 68.8% (4 probes) for chr2 and from 91.4% (18 probes) to 66.8% (4 probes) for chr6, indicating that probe number, not chromosome size, affects classification accuracy (**Fig 4A**). Furthermore, the interquartile range (IQR) of the accuracy distributions increased from 0.8 percentage points (pp; 22 probes) to 9.14 pp (4 probes) for chr2 and from 1.5 pp (16 probes) to 11.6 pp (4 probes) for chr6 indicating that having fewer probes per chromosome increases the variability of classification accuracy. Both the decrease in accuracy and increase in accuracy variance could be observed for the theoretical sensitivity at all 5 copy number levels as well (**Fig S3A**).

Additional Probes Increase Sensitivity

Since a greater number of probes correlates with higher copy number call sensitivity, we reasoned that we could further increase sensitivity for all chromosomes by increasing the number of probes targeting each chromosome. To determine the number of probes required to approach 100% sensitivity for all 5 copy number levels, we simulated a panel using all probes targeting chromosomes 1 through 6 (120 probes total) and smoothed their RPE1 cell-probe copy number scores as if they were measurements from a single hypothetical chromosome. This is possible because all 6 chromosomes have 2 copies in RPE1. For each trial, we constructed a new panel

probe-by-probe by sampling the 120 probes without replacement, recalculating RPE1 copy number and sensitivity at every step until all 120 probes were added, repeated for 50 trials. As expected, mean theoretical sensitivity increased with probe number for all copy number levels (**Fig 4B**). The simulation achieved at least 90% sensitivity on average at 4 probes per chromosome for 1 copy, 16 probes for 2 copies, 42 probes for 3 copies, and 78 probes for 4 copies. A maximum mean sensitivity of 91.5% was achieved for 5 copies at 120 probes. Again, the variability of the copy number call sensitivity decreased as the number of probes increased. In some cases, it may be sufficient for the user to detect either a gain or loss of an otherwise diploid chromosome rather than detect the specific copy number of the gained chromosome. In this circumstance, a GMM can be generated with only 3 components, representing states of loss (1 copy), neutral (2 copies), and gain (3 or more copies). We evaluated our simulation under this model, achieving at least 90% sensitivity at 4 probes for 1 copy, 16 probes for 2 copies, and 20 probes for ≥3 copies (**Fig S3B**). Furthermore, 99% sensitivity could be achieved at 17 probes for 1 copy, 55 probes for 2 copies, and 67 probes for ≥3 copies. These findings indicate that copy number call sensitivity can be increased for all copy number levels by adding probes to our panel.

KaryoTap Version 2 Panel Increases Accuracy of Aneuploidy Detection

As the accuracy of our custom panel increases with the number of probes targeting a chromosome, we attempted to improve the accuracy by adding probes to chromosomes with lower coverage. To balance the cost of producing a larger custom panel with meaningful sensitivity gains, we removed the 61 least efficient probes (by total read counts) and added 82 probes such that each chromosome was targeted by at least 12 probes (**Table S1**). We also included 4 probes targeting chrY, which was not covered by Version 1, to enable the detection of all 24 chromosomes (**Fig S4A**). The new panel, Version 2 (v2; CO610) comprises 352 total probes, 270 of which are shared with Version 1 (**Fig 1; Additional File 2**).

To evaluate performance for the Version 2 panel, we performed a Tapestri experiment using RPE1 cells (**Fig S4B**) and determined the empirical accuracy as the proportion of cells with correct copy number calls based on known copy number from karyotyping. We again performed 5-fold cross validation by partitioning the RPE1 cells randomly into 5 equally sized subsets and calculating accuracy 5 times, each time reserving one of the sets from the model generation and using it only to calculate accuracy. The mean absolute deviation of the 5 accuracy measurements for each chromosome ranged from 0.92 to 3.89, suggesting that copy number calling performance would be maintained when classifying new data. The poorest performing chromosome in Version 1, chr22, had an accuracy of 70% with Version 2 compared to 49% for Version 1 (**Fig 5A-B, Fig S4C**). The mean accuracy across all chromosomes was 89%, increased from 82% for Version 1. Accuracy increased by 2.2 pp on average for each additional probe (**Fig 5C**). Similarly, for chromosome arms, the average accuracy across arms increased from 73% with Version 1 to 80% with Version 2 (**Fig 5D-E, Fig S4D**). We also calculated the theoretical sensitivity for copy number values of 1, 2, 3, 4, and 5 and saw increased average sensitivity compared to Version 1 (**Fig S4E, Table S3**). Furthermore, we calculated theoretical sensitivity for a simpler 3-component model representing chromosome loss, neutral, and gain states, which may be a more practical choice for certain users. The 3-component model had a mean sensitivity of 99% for losses, 90% for neutral states, and 87% for gains for whole chromosomes; in addition it showed a sensitivity of and 97% for losses, 81% for neutral states, and 78% for gains for chromosome arms (**Table S3**). These data provide strong evidence that increasing the number of probes targeting each chromosome improves the sensitivity of the panel in calling copy numbers in individual cells.

Detection of Lentiviral Barcodes and gRNAs

To extend the capabilities of our system, we added two probes to the Version 2 panel that target and amplify either a DNA barcode sequence or CRISPR guide RNA (gRNA) sequence integrated

into a cell's genome by lentiviral transduction. DNA barcoding of cells can be used in situations where several samples from the same cell line or individual are sequenced in one experiment and are therefore unable to be distinguished by genotype. Similarly, CRISPR gRNAs can be used both for functional studies and as barcodes themselves, indicating the gRNA treatment received by an individual cell.

As a proof-of-concept, we transduced RPE1 cells and human colorectal epithelial cells (hCECs) each with distinct gRNA constructs (gRNA1 and gRNA2, respectively; **Table S4**), and human Pancreatic Nestin-Expressing cells (hPNEs) with a mix of two DNA barcode constructs that drive expression of BFP. We used distinct cell lines for each construct so that the three populations could be distinguished by genotype without assuming successful barcoding. To enable panel Version 2 to detect gRNAs, we designed a probe, Probe AMP350, to target the region surrounding and including the gRNA sequence in the lentiviral vector. To enable the detection of DNA barcodes from the BFP-expressing vector, we similarly designed a probe, AMP351, to target the region surrounding and including the barcode sequence (**Fig 5F**).

gRNAs were each transduced into target cells with a multiplicity of infection (MOI) of 1-1.5 followed by puromycin selection to ensure that each cell had an average of ~1 integration and at least 1 integration. The BFP barcodes were transduced at a higher MOI and cells were enriched for BFP expression by FACS; a high MOI was used for the barcoding sequences to increase the chances of detection. The three cell populations were pooled and analyzed in a single Tapestri experiment using panel Version 2. The populations were distinguished by PCA, UMAP, and clustering of VAFs (**Fig S4B**) as done previously. To determine if gRNA1 could be detected in RPE1 cells, we took the aligned reads from Probe AMP350 associated with RPE1 cells and searched for the sequence of gRNA1. RPE1 cells had an average of 34 gRNA1 reads per cell, while hCECs and hPNEs had 0. Similarly, hCECs had an average of 30 gRNA2 reads per cell,

while RPE1 and hPNEs had 0 (**Fig 5G**). To determine if the BFP barcodes could be detected in hPNEs, we similarly took aligned reads from Probe AMP351 and searched for the sequence of the barcodes. hPNEs had an average of 81 barcode reads per cell, while RPE1 and hCECs had 0. Altogether, these data indicate that gRNA sequences and specific DNA sequences can be recovered from transduced cells using panel Version 2.

To determine the limit of detection for a gRNA in transduced cells, we analyzed the proportion of cells with 0 reads matching the appropriate gRNA sequence. 21% of RPE1 and 31% of hCECs had 0 counts per cell for gRNA1 and gRNA2, respectively (**Fig 5G**). We compared the read counts per cell for Probe AMP350 with the number of reads matching the appropriate gRNA sequence in both RPE1 and hCECs and found that virtually all of the counted reads from Probe AMP350 matched the number of gRNA1 sequence reads for RPE1 and the gRNA2 sequence reads for hCECs (**Fig S4F**), indicating no contamination from other sequences. Altogether, these data indicate that Probe AMP350 in Panel Version 2 can detect at least one gRNA sequence in ~70-80% of cells, though the rate of detection may be improved by transduction at a higher MOI.

We repeated a similar analysis for Probe AMP351 to determine the limit of detection for a DNA barcode in transduced cells. 6% of hPNEs had 0 barcode sequence counts per cell (**Fig 5G**) and >99% of the reads from Probe AMP351 matched the known barcode sequences in hPNEs (**Fig S4F**), indicating no contamination. Altogether, these data indicate that Probe AMP351 in Panel Version 2 can detect a DNA barcode in 94% of cells, which may be improved by increasing depth or MOI.

Evaluation of Aneuploidy Induction by KaryoCreate

To demonstrate the combined copy number detection and multiplexing capabilities of our system, we tested it on samples treated with KaryoCreate (Karyotype CRISPR Engineered Aneuploidy

Technology), a method we recently developed to induce chromosome-specific aneuploidy in cultured cells(14). KaryoCreate uses CRISPR gRNAs to target a mutant KNL1-dCas9 fusion protein to the centromere of a specific chromosome, causing missegregation in ~20% of cells. KaryoTap represents an ideal method to evaluate the efficiency of aneuploidy induction and chromosome-specificity of KaryoCreate. To do this, we performed a Tapestri experiment on hCECs that had been treated with 1 of 3 gRNAs previously tested using KaryoCreate: sgNC does not have a target and is used as a negative control, sgChr6-2 targets chr6 and sgChr7-1 targets chr7 (**Table S4)**. The gRNA sequences amplified by AMP350 were used to identify the gRNA that each cell received. sgChr6-2 and sgChr7-1 induced gains and losses specifically in the intended chromosomes, but not others, compared to sgNC (**Fig 6A, Table S5**; $p < 0.01$, Fisher's exact test). sgChr6-2 induced 26.6% losses of chr6 compared to 0.5% with sgNC, and 8.3% gains compared to 2.0% with sgNC (**Table S6**). sgChr7-1 induced 6.1% losses of chr7 compared to 0.5% with sgNC, and 5.1% gains compared to 4.3% with sgNC.

In the same experiment, we also performed a small screen to address a current limitation of KaryoCreate in which we were unable to engineer aneuploidy of certain chromosomes, such as chromosome 20, one of the most frequently gained chromosomes in human cancer (2). In fact, while we could design gRNAs that are specific to the centromere of chromosome 20, it was not possible to visualize centromeric foci through the co-transduction of cells with gRNAs and fluorescently-tagged dCas9 by imaging, possibly due to the low (~700) number of gRNA binding sites (14). Furthermore, given the small size of chromosome 20, the single-cell RNA sequencing-based approach used in Bosco et al. does not have sufficient sensitivity to confidently assess gains and losses of this chromosome. Thus, using KaryoTap, we screened 5 sgRNAs targeting chromosome 20 (sgChr20-2, 20-3, 20-4, 20-6 and 20-7) that were previously described but not validated by imaging (14). sgChr20-2, 20-4, 20-6 and 20-7 did not induce changes in chr20 ($p = 0.67$-0.94). sgChr20-3 was able to induce 7.4% gains in chr20 compared to 3.0% with the sgNC

control, and 4.7% losses compared to 1.8% with sgNC ($p = 0.006$). We also note that sgChr20-3 induced 9.4% losses and 17.8% gains in chr2 ($p < 0.001$), which we might not have observed if we had instead evaluated the effect of the gRNA using a chromosome-targeted method such as fluorescence in situ hybridization rather a method which covers all chromosomes. The sgChr20-3 sequence (G**GC**AGCTTTGAGGATTTCGT) matches 18 out of 20 base pairs for loci on the chr2 centromere (G**AT**AGCTTTGAGGATTTCGT) (14), suggesting an explanation for the off-target effect. These data indicate that KaryoTap successfully enables simultaneous detection of aneuploidy and gRNA/barcodes in the same cells and thus can be used to perform CRISPR-based (i.e., gRNA-based) or ORF-based (barcode-based) functional screens.

Detection of Copy Number Neutral Loss of Heterozygosity (CNN-LOH)

Gains and losses of diploid chromosomes result in a shift in VAF for their heterozygous SNPs from 50% in the direction of 100% or 0% depending on which parental chromosome copy (i.e., haplotype) experienced a copy number change. In addition, a shift in VAF can also be observed in the absence of copy number changes in copy number neutral loss of heterozygosity (CNN-LOH), which has been observed in cancer as well as normal tissues(33,34) . Because each probe is sequenced at a high depth, KaryoTap should be able to detect this shift, allowing us to determine which of the two parental chromosomes/haplotypes was gained or lost. This is especially important for detecting loss of heterozygosity (LOH), a common event in cancer whereby a heterozygous-to-homozygous shift by chromosomal loss can inactivate tumor suppressor genes(35,36). To determine if KaryoTap could detect allele frequency shifts following chromosomal gains and losses, we examined the cells from the KaryoCreate experiment (**Fig 6A**) that had lost a copy of chromosome 6 after treatment with the sgChr6-2 gRNA. We identified 9 heterozygous variants on chr6 called by the Tapestri Pipeline by identifying variants with a mean allele frequency between 20-80% in the sgNC control population. We then calculated a relative (i.e., haplotype-agnostic) allele frequency for sgChr6-2 treated cells with 1 or 2 copies of chr6

(called by GMM) by calculating the absolute difference between raw allele frequency and 50% such that 0 corresponded to heterozygous alleles and 50 corresponded to fully homozygous alleles (**Fig 6B**). 4 distinct clusters of cells emerged using K-means clustering on relative AFs. The cluster comprising cells with a copy number call of 2 for chr6 shows that the 9 variants are heterozygous in diploid cells as expected. The cluster with 1 copy of chr6 shows a shift across the variants from heterozygous to homozygous (i.e. a loss of heterozygosity), supporting the loss of one copy of each allele in these cells. There are also two smaller clusters representing the loss of either chromosome arm but not the other, supported by both the loss of heterozygosity in the variants on the affected arm and the copy number call of 1 for that arm. This indicates that KaryoTap can be used to detect loss of heterozygosity in single cells at the population level.

It is possible that the loss of a chromosome can be followed by a duplication of the remaining chromosome, such that the copy number of the chromosome remains the same, but one allele is lost, i.e., a CNN-LOH. To determine if KaryoCreate can cause CNN-LOH in the targeted chromosome, we took the relative allele frequencies for sgChr6-2 treated cells with 1 or 2 copies of chr6 calculated above and averaged them such that each cell had one mean relative AF value for chr6. We also repeated this calculation for the cells treated with the sgNC control gRNA. If we consider relative AF between 40% and 50% to indicate homozygosity of chr6 alleles and 0% to 40% to indicate heterozygosity, all sgNC-treated cells with a chr6 copy number of 2 were heterozygous (**Fig 6C**)**.** Cells treated with sgChr6-2 that lost a copy were detected to be homozygous. 85% of cells treated with sgChr6-2 that had 2 copies of chr6 detected were heterozygous, while 15% were detected as homozygous, indicating a loss of heterozygosity with no change in the net copy number (2 copies of chr6). This indicates that KaryoCreate can induce CNN-LOH and KaryoTap can detect CNN-LOH events.

KaryoTap Version 3 Panel Further Improves Accuracy of Aneuploidy Detection

To further improve the accuracy of our system for detection of aneuploidy, especially for chromosome arms, we modified panel Version 2 to create Version 3 (v3, CO810; **Fig 1; Additional File 3**). Version 3 comprises 399 total probes, 309 of which are shared with Version 2. We removed 43 less-efficient probes from Version 2 and added 90 new probes, prioritizing chromosome arms with less coverage. The barcode and gRNA detecting probes described above were also included in the new design.

To evaluate the performance of panel Version 3, we performed a Tapestri experiment using RPE1 cells, made copy number calls using the GMM strategy as described above, and calculated accuracy as the proportion of cells with correct copy number calls based on known copy number. We again performed 5-fold cross validation by partitioning the RPE1 cells randomly into 5 equally sized subsets and calculating accuracy 5 times, each time reserving one of the sets from the model generation and using it only to calculate accuracy. The mean absolute deviation of the 5 accuracy measurements for each chromosome ranged from 0.60 to 3.31. The poorest performing chromosome in Version 2, chr22, had an accuracy of 78% with Version 3 compared to 70% for Version 2 (**Fig 7A-B**). The mean accuracy across all chromosomes was 91%, increased from 89% for Version 2. Similarly, for chromosome arms, the average accuracy across arms increased from 80% with Version 2 to 83% with Version 3 (**Fig 7C-D**). We also calculated the theoretical sensitivity for copy number values of 1, 2, 3, 4, and 5 and saw increased average sensitivity compared to Version 2 (**Fig S5A-B, Table S7**) for both whole chromosomes and chromosome arms. Furthermore, we calculated theoretical sensitivity for a simpler 3-component model representing chromosome loss, neutral, and gain states. The 3-component model had a mean sensitivity of 99% for losses, 91% for neutral states, and 88% for gains for whole chromosomes; it addition it showed a sensitivity of 97% for losses, 83% for neutral states, and 80% for gains for chromosome arms (**Table S7**). When compared qualitatively, the general degree of noisiness in

heatmaps of the GMM copy number calls for RPE1 decreases across chromosome arms between KaryoTap panels Version 1, 2 and 3, supporting improvement in the accuracy of copy number calling afforded by panel Version 3 (**Fig 7F**). Noisiness also decreases in heatmaps of copy number calls for the LoVo and LS513 cell lines between panels Version 1 and Version 3 (**Fig 7G-H**), supporting the improvement of copy number variant detection sensitivity for individual cells using our system. Altogether, these data indicate that panel Version 3 can deliver accurate copy number calls in thousands of single cells.

Discussion

We designed and improved two custom panels for the Tapestri platform that enable targeted scDNA-seq of 330-352 specific loci across all 24 human autosomes and sex chromosomes. This, coupled with a GMM-based copy number calling analysis pipeline, allowed us to identify chromosome- and chromosome arm-scale aneuploidy in thousands of individual cells in a single experiment with high accuracy and at greatly reduced sequencing depth compared to single-cell WGS methods. To increase the ease-of-use, we compiled the computational scripts used to analyze these data into an R package, *karyotapR,* which automates all steps for calling copy number and for basic visualization of the results (**Fig S6**).

While single-cell aneuploidy detection is not unique to KaryoTap, our design has several advantages, the most critical being the leveraging of Tapestri's throughput thus significantly reducing hands-on time, reagent cost, and sequencing cost per cell compared to low-throughput WGA methods. Additionally, targeting of specific loci allows us to forgo the typical technical difficulties of conventional WGS analysis, including correcting for mappability bias and GC bias, and the use of segmentation algorithms to make copy number calls(27). Furthermore, the total

number of sequencing reads needed to obtain high-accuracy copy number calls for KaryoTap is greatly reduced compared to WGS-based scDNA-seq methods and the targeted nature of the assay spreads the reads more evenly across the genome, preventing the biasing of detection sensitivity toward larger chromosomes that is seen with WGS. We expect this will be particularly important for assessing chromosomal instability as the smallest of chromosomes will be more equally represented in the data relative to the larger chromosomes. Finally, the commercial availability of the Tapestri system allows for easier adoption compared to non-commercialized "homebrew" methods that need to be established and optimized in each lab from scratch(23).

As evidenced by the analysis of RPE1, our method will generate a range of smooth copy number measurements for chromosomes with the same discrete copy number, indicating some level of technical error. To account for this error and convert the continuous smooth copy number scores to discrete copy number values, we used a Gaussian mixture model (GMM) classification strategy, which has been previously used for copy number analysis of single-cell whole genome sequencing data(37). This allows each smooth copy number score to be associated with a set of (posterior) probabilities of being measured from a chromosome of a given range of copy numbers (e.g., 1, 2, 3, 4, or 5). While we assign each smooth score to the discrete copy number value for which its posterior probability of belonging is highest, the probabilities of belonging to the other copy number components of the model indicate the confidence the investigator can have that the call is accurate. As the number of probes increases, the variance of the model components decreases, resulting in an increase in classifier accuracy that we observe between our Version 1 and Version 2 panel designs. While we use a near-diploid cell line as our ground truth, the GMM strategy also allows us to calculate the expected (theoretical) sensitivity that a chromosome with copy number 1, 2, 3, 4, or 5 would be correctly called using our system by calculating the proportion of overlap between the copy number components of the model. We used this in our panel simulation to extrapolate an optimal number of probes for detecting 1 copy (loss), 2 copies

(neutral) or 3 copies (gain) of a chromosome and determined that at least 90% sensitivity could be achieved at 4 probes for 1 copy, 20 probes for 2 copies, and 26 probes for ≥3 copies, and 99% sensitivity could be achieved at 22 probes for 1 copy, 66 probes for 2 copies, and 76 probes for ≥3 copies. Further improving the panel by increasing the number of probes and thus reducing the technical variation will allow us to more confidently observe the karyotype heterogeneity in these samples as well as in tumors and other tissues. It is important to note here that our method requires a baseline sample with which to compare the other cells in the experiment. However, while we set the baseline copy number using near-diploid RPE1 to scale the read counts of each probe relative to 2 copies, it is not strictly necessary to spike a diploid control cell line into the sample preparation. Any distinguishable, largely homogenous subset of cells in an experiment can be used to set the baseline as long as the average copy number for each chromosome in that subset is known.

Deep sequencing (~80-100X on average per cell) of each target region allows for robust single nucleotide variant (SNV) calling that is not possible at the lower genomic coverage afforded by other high-throughput methods(23). This enabled us to resolve and identify 5 multiplexed cell lines in a single experiment by clustering cells by variant allele frequencies. Since our panels specifically target loci known to harbor SNPs across the human population, we can extend sample multiplexing to clinical samples (e.g., tumor tissue) from different individuals without the need for barcodes. While we demonstrated sample identification using a clustering approach here, sample identities for each cell can be determined directly from known SNPs that occur at sequenced loci. Additionally, while our panels were designed for copy number analysis, additional probes could be added that cover tumor suppressor genes and oncogenes of interest, thus revealing consequential point mutations alongside chromosomal copy number. Mission Bio offers several ready-made panels covering mutational hotspots and genes relevant to a broad range of tumor types, allowing for a great degree of customizability. Furthermore, we demonstrated that SNV-

associated allele frequency shifts detected using KaryoTap could be used to infer loss of heterozygosity (LOH), a common event in cancer where the germline heterozygous state of a chromosome changes to a homozygous state in tumor cells(38). LOH has been demonstrated to promote tumorigenesis by inactivating tumor suppressor genes through chromosomal loss(35,36). In the context of chromosomal instability, the chromosome remaining after the loss of its homologue can be duplicated, resulting in LOH with a net-neutral copy number change (copy number neutral (CNN)-LOH). The deep sequencing depth and copy number detection enabled by KaryoTap allow for discovery of CNN-LOH events, which would otherwise be difficult to detect with the shallow coverage typical of other scDNA-seq methods(38).

To enable experimental design flexibility when using our custom panels, we added a set of probes that can detect DNA barcode and CRISPR gRNA sequences integrated into the genome. DNA barcodes can be used to multiplex and resolve cells belonging to different samples in a single Tapestri experiment that otherwise could not be distinguished by genotype. Through barcoding, users can compare samples from the same individual or compare experimental and control conditions in the same cell lines while minimizing batch effects. Regardless of design, combining several samples into one experimental run greatly reduces the per-sample reagent and sequencing costs in addition to the hands-on time required to process the samples. Detecting barcodes in thousands of cells is made possible by exploiting the targeted nature of the sequencing assay. Single-cell DNA sequencing methods with comparable throughput rely on inefficient and random transposon insertion, which would only detect a randomly inserted barcode in about 20% of cells(23). By specifically targeting the barcoded insert, we can reliably recover the barcode sequence in over 90% of cells. Including a probe that targets inserted CRISPR gRNA sequences allows for an additional layer of experimental design flexibility where the gRNA-mediated treatment each cell receives can be identified by the gRNA sequence itself. Here, we demonstrated the gRNA detection and multiplexing capabilities of our system by evaluating the

efficiency and specificity of KaryoCreate, our method for inducing chromosome specific aneuploidy. Since the gRNAs can be detected in 70-80% of cells when transduced at low MOI, this system could also be used for CRISPR screen applications where cells are randomly treated with one gRNA from a library of hundreds of possible gRNAs and thus require high detection sensitivity(39).

Our system in its current form is limited to calling whole chromosome aneuploidy, and, with less confidence, chromosome arm-level aneuploidy. Further optimization of the panels will be required to achieve sufficient confidence in copy number detection for some chromosome arms. The sensitivities of either measurement vary according to the number of probes used, and thus confidently calling aneuploidy in increasingly smaller regions becomes challenging. Sub-arm (i.e., focal) aneuploidy could potentially be detected with a greater density of probes, though manufacturing increasingly larger panels also increases the cost of the panels. Fortunately, this cost can be offset by lowering sequencing depth, as we demonstrated that sequencing at as low as ~35 average reads per cell per probe does not significantly affect aneuploidy call accuracy.

Coupled with the Tapestri platform, KaryoTap shows considerable promise as an easily adoptable, flexible, and highly scalable method for detecting chromosome- and chromosome arm-scale aneuploidy in thousands of single cells. Here we demonstrated population-level copy number detection in several cell lines and, most significantly, highly accurate copy number classification in individual cells using a Gaussian mixture model framework, which is otherwise unattainable using currently available methods. We identified the number of PCR probes per chromosome as a dominant factor affecting copy number classification performance and calculated the number of probes necessary to sufficiently improve detection sensitivity for various applications. Finally, we applied our method to the aneuploidy-engineering tool KaryoCreate to demonstrate sample multiplexing capabilities and the ability to detect gRNAs in transduced cells.

We believe this system lays the groundwork for a new class of tools for studying aneuploidy and chromosomal instability in healthy and diseased tissues and tumors.

Methods

Cell Culture

All cells were grown at 37°C with 5% CO2 levels. All cell media was supplemented with 1X pen-strep, and 1X L-glutamine. hTERT human retinal pigment epithelial cells (RPE-1; ATCC CRL-4000) and SW48 cells (ATCC CCL-231) were incubated in DMEM, supplemented with 10% FBS. LoVo cells (ATCC CCL-229) were incubated in Ham's F12-K media with 10% FBS. LS513 cells (ATCC CRL-2134) and hTERT human pancreatic nestin-expressing cells (hPNEs; ATCC CRL-4023) were incubated in RPMI media with 10% FBS. CL11 cells (Cellosaurus CVCL_1978) were incubated in DMEM:F12 and 20% FBS. hTERT p53-/- human colonic epithelial cells (hCECs; Ly et al.(40)) were cultured in a 4:1 mix of DMEM:Medium 199, supplemented with 2% FBS, 5 ng/mL EGF, 1 µg/mL hydrocortisone, 10 µg/mL insulin, 2 µg/mL transferrin, 5 nM sodium selenite, pen-strep, and L-glutamine. For long-term storage, cells were cryopreserved at −80°C in 70% cell medium, 20% FBS, and 10% DMSO. All cell lines were tested for mycoplasma.

Custom Tapestri Panel Design

Panel Version 1 (CO261) comprises 330 probes across the 22 human autosomes and the X chromosome. To identify candidate target regions for the panel, we used the Common SNP files downloaded from UCSC(41,42) (snp151Common, hg19), and considered only synonymous variant SNPs with a major allele frequency at >0.5 and <0.6. For cytobands with more than 4 synonymous variants, we split the cytoband into 4 subregions based on the percentile of the

cytoband coordinates (0-25th percentile, 25-50th percentile, 50-75th percentile and 75-100th percentile). From each subregion, we randomly selected 1 SNP as a representative candidate. In cases where there were less than 5 synonymous variant SNPs, all SNPs were used. We submitted all candidate SNPs to the Tapestri Panel Designer to generate a panel design and ensured that the designed probes targeted the candidate SNPs and had similar GC contents. Next, randomly selected probes such that each chromosome had a probe density of ~1 per 10MB. Panel Version 2 (CO610) comprises 352 probes across all 24 human chromosomes. This panel was generated using Panel v1 as a base: first, we removed 61 probes that had low PCR amplification efficiency based on total read counts per probe. Then we added 82 probes such that each chromosome was targeted by at least 12 probes and included 4 probes targeting chrY. To enable the detection of lentiviral-delivered gRNAs, we added one probe targeting the region of the construct containing the gRNA sequence and one probe targeting a region upstream as a vector control. Similarly for the detection of lentiviral-delivered DNA barcodes, we added one probe targeting the region of the construct surrounding the barcode sequence, and one probe targeting a region downstream as a vector control. Support for the custom panel design and synthesis of the panel was provided by Mission Bio (San Francisco, CA, USA). Panel maps were created using the karyoploteR R package(43).

Tapestri Single Cell DNA Sequencing

Cell lines were trypsinized for 2-3 minutes, washed in room temperature Mg^{2+}/Ca^{2+}-free DPBS, centrifuged at 300g for 5 minutes, and resuspended in DPBS at a concentration of 3K cells/uL. For the experiment using the RPE1, SW48, LS513, LoVo, and CL11 cell lines, 600K cells from each cell line were combined, centrifuged at 300g for 5 minutes, and resuspended in Tapestri Cell Buffer at a concentration of 3.5K cells/uL. For the experiment using the RPE1, hPNE, and hCEC cell lines, 45K cells from each cell line were combined, centrifuged at 300 x g for 5 minutes,

and resuspended in Tapestri Cell Buffer at a concentration of 4K cells/uL. For the KaryoCreate experiment, ~100K cells from each condition were combined, centrifuged at 300 x g for 5 minutes, and resuspended in Tapestri Cell Buffer at a concentration of 3.4K cells/uL. Cell droplet encapsulation, barcoding, and sequencing library preparation were performed using the Tapestri instrument according to the manufacturer's instructions (Mission Bio, San Francisco, CA, USA). Sequencing was performed using an Illumina NovaSeq 6000 or NextSeq 500 in 2x150bp paired-end format. After sequencing, deconvolution of barcodes, read counting, and variant calling were handled by the online Tapestri Pipeline (v2.0.2)(26). The pipeline outputs both read counts per probe for each cell and variant allele frequencies for called variants for each cell.

Low Pass Whole Genome Sequencing & Karyotyping

Genomic DNA was extracted from cell pellets using 0.3 µg/µL Proteinase K (QIAGEN #19131) in 10mM Tris pH 8.0 for 1 hour at 55°C, following heat inactivation at 70°C for 10 minutes. DNA was digested using NEBNext dsDNA Fragmentase (NEB #M0348S) for 25 minutes at 37°C followed by magnetic DNA bead cleanup with 2X Sera-Mag Select Beads (Cytiva #29343045). Library prep was performed using NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB #E7103) according to the manufacturer’s instructions, generating DNA libraries with an average library size of 320 bp. Quantification was performed using a Qubit 2.0 fluorometer (Invitrogen #Q32866) and the Qubit dsDNA HS kit (Invitrogen #Q32854). Libraries were sequenced on an Illumina NextSeq 500 at a target depth of 4-8 million reads. Reads were trimmed using trimmomatic(44), aligned to the hg38 genome using bwa-mem(45), and analyzed for copy number variants using the CopywriteR(46) R package. G-banded karyotyping of 100 RPE-1 cells was performed by WiCell Research Institute, Inc. (Madison, WI).

Cloning of sgRNAs

We modified the scaffold sequence of pLentiGuide-Puro (Addgene #52963) by Gibson assembly to contain the A-U flip (F) and hairpin extension (E) described by Chen et al(47). for improved sgRNA-dCas9 assembly, obtaining pLentiGuide-Puro-FE. sgRNAs were designed and cloned into this pLentiGuide-Puro-FE vector according to the Zhang Lab General Cloning Protocol(48). To be suitable for cloning into *BbsI*-digested vectors, sense oligos were designed with a CACC 5' overhang and antisense oligos were designed with an AAAC 5' overhang. The sense and antisense oligos were annealed, phosphorylated, and ligated into *BbsI*-digested pLentiGuide-Puro-FE for KaryoCreate purposes. Sequences were confirmed by Sanger sequencing.

Gateway Recombination Cloning for Generation of Barcode Library

pHAGE-CMV-DEST-PGKpuro-C-BC was a library of lentiviral vectors containing 24-bp random barcodes that was built as described in Sack & Davoli et al., 2018(49). Destination vector pHAGE-CMV-DEST-PGKpuro-C-BC and entry vector pDONR223_BFP (Addgene: 25891) are recombined following the manufacturer's protocol. Briefly, 50 ng of entry vector and 100 ng of destination vector are mixed with LR Clonase™ enzyme and incubated overnight at room temperature. The next day, the reaction mixture is incubated with Proteinase K at 37°C for 10 minutes, followed by inactivation at 75°C for 15 minutes. The reaction is then transformed into stbl3 bacterial competent cells, plated onto LB agar plates, and incubated overnight at 37°C. Individual clones are collected into 96 well plates and expanded. Plasmid is extracted from the bacterial culture using a 96-well mini-prep kit (Zymo kit, Zippy 96 plasmid kit). All clones are sequenced by Sanger sequencing at the site of the barcode using primer ACTTGTGTAGCGCCAAGTGC. Duplicates are eliminated, and unique barcodes are retained in

the final library. BFP expression and Puromycin selection are validated by transfecting randomly selected clones into HEK293T cells.

Lentivirus Production and Nucleofection

For transduction of cells, lentivirus was generated as follows: 1 million 293T cells were seeded in a 6-well plate 24 hours before transfection. The cells were transfected with a mixture of gene transfer plasmid (2 µg) and packaging plasmids including 0.6 µg ENV (VSV-G; addgene #8454), 1 µg Packaging (pMDLg/pRRE; addgene #12251), and 0.5 µg pRSV-REV (addgene #12253) along with $CaCl_2$ and 2× HBS or using Lipofectamine 3000 (Thermo #L3000075).The medium was changed 6 hours later and virus was collected 48 hours after transfection by filtering the medium through a 0.45-µm filter. Polybrene (1:1000) was added to the filtered medium before infection.

KaryoCreate Experiments

KaryoCreate experiments were performed as described in Bosco et al., 2023(14). Briefly, p53-/- hCEC were first lentivirally transduced with pHAGE-KNL1Mut-dCas9 and selected with blasticidin. The cells were then lentivirally transduced with the indicated sgRNAs and selected with puromycin. scDNA seq was performed ~10 days after transduction with the gRNAs. The sequences of the gRNAs targeting the centromeres of specific chromosomes are listed in Table S4 and were designed as described in Bosco et al., 2023(14). To compare conditions, Fisher's exact test was performed in R using the `fisher.test()` function, comparing the proportion of cells for each chromosome and sample that are diploid (copy number = 2) and aneuploid (copy number = {1, 3, or 4} between the sgNC control and sample and the given experiment sample. The

Benjamini-Hochberg correction for multiple comparisons was applied to p-values using `p.adjust()`.

Parsing and Counting of Barcoded Reads

To detect specific gRNA or DNA barcode sequences, we searched for the known sequences against the cell-associated aligned reads (cells.bam file) generated from the Tapestri Pipeline. Search queries were conducted vcountPattern() using the Biostrings R package(50), with tolerance for up to 2 base mismatches. BAM files were manipulated using the Rsamtools R package(51).

Cell Line Demultiplexing and Identification

To demultiplex cells from different cell lines in the initial scDNA-seq experiment, we use the allele frequency (AF) of variants that are called by GATC as part of the Tapestri Pipeline. Variants were filtered by selecting those with standard deviations of AF >20 to select variants whose allele frequencies vary the most across all cell lines. PCA was used to reduce the dimensions of the remaining variants. The top 4 principal components were embedded in two dimensions by UMAP and then clustered using the dbscan method. The 5 clusters with the greatest number of cells were kept, corresponding to the 5 expected cell lines. The remaining clusters, likely representing cell doublets, were discarded from further analyses. This method was repeated for subsequent Tapestri experiments, adjusting for the expected number of cell populations.

The cluster containing RPE1 cells in each experiment was identified by clustering with published deep WGS of RPE1. Published RPE1 WGS data was obtained from SRA Accession ERR7477340(29,52). Reads were aligned to the hg19 genome using bwa. MarkDuplicatesSpark

and HaplotypeCaller from GATK were used to mark duplicate reads and get AFs from called variants(53). The vcfR R package(54) was used to extract the AFs for called variants common to the published data and our dataset. The mean AF for each variant was calculated for each of the 5 cell lines. PCA was used to cluster our mean AF dataset with the RPE1 AFs. The cell line that clustered most closely with the published RPE1 data was labeled as RPE1 cells.

Copy Number Calling

Copy number scores for each probe in each cell (cell-probe scores) were calculated relative to RPE1, for which we know the copy number of each chromosome: The raw count matrix was normalized by scaling each cell's mean to 1 (Equation 1) and then each probe's median to 1 (Equation 2). The normalized counts were then scaled such that the value of the median normalized RPE1 counts for each probe was set to 2 for all probes except those targeting chr10q, which were set to 3 (Equation 3). The identities of the remaining 4 populations of cells were identified by comparing their overall copy number profile with matched bulk WGS data.

$$RC_{intermediate} = {RC_{cell,probe}}\Big/{mean(RC_{cell}) + 1} \qquad (1)$$

$$RC_{norm} = {2 \cdot RC_{intermediate}}\Big/{median(RC_{intermediate,probe}) + 0.05} \qquad (2)$$

$$CN_{cell,probe} = {CN_{WGS,probe} \cdot RC_{norm}}\Big/{median(RC_{norm,RPE1})} \qquad (3)$$

Smooth copy number scores for each chromosome in each cell (cell-chromosome scores) were generated by taking the weighted median of the probe-specific copy number values for probes

targeting a common chromosome (Equation 4). Weights for each probe were calculated as the proportion of RPE1 cell-probe scores that fell within ± 0.5 of the known copy number (3 for chr10q, 2 for all others). This was also modified to calculate cell-chromosome-arm scores for probes common to a chromosome arm.

$$CN_{smooth,cell,chr} = weightedMedian\left(CN_{cell,probe}\right), \text{ for all probes on chromosome } chr \quad (4)$$

Integer copy number values for each cell-chromosome were classified using Gaussian mixture models (GMMs) with either five components representing possible copy number values of 1, 2, 3, 4, and 5, or three components representing copy number values 1, 2, and 3. To generate the GMMs, the normalized counts for each probe for the RPE1 cells were fitted to Weibull distributions using the fitdistrplus R package(55). These Weibull parameters represented parameters for copy number = 2 for all probes except those targeting chr10q, which has 3 copies in RPE1. The scale parameters were then scaled for possible copy number values 1 through 6, relative to the RPE1 copy number: probes with RPE1 copy number = 2 were scaled by 50%, 100%, 150%, 200%, 250%, and 300% for copy number = 1, 2, 3, 4, 5 and 6; probes with RPE1 copy number = 3 were scaled by 33%, 67%, 100%, 133%, 167%, and 200% for copy number = 1, 2, 3, 4, 5 and 6. 500 Weibull-distributed values are drawn using each of the six parameter sets for each probe to simulate six matrices of 500 simulated cells. For each cell, the values were smoothed across the probes belonging to each chromosome to simulate cell-chromosome copy number values. The distribution of the scores for each chromosome was then fit to Gaussian (normal) distributions, separately for each copy number level. The result is a set of normal parameters (mean μ and standard deviation σ) for each chromosome for each value of copy number k = {1, 2, 3, 4, 5, 6}. The six copy number Gaussian components for each chromosome were combined into a GMM, representing the probability densities for each copy number value for that chromosome (Equation

5). Using Bayes rule and assuming equal priors, the posterior probability of a cell-chromosome copy number score being generated under each component k is given by Equation 6.

$$pdf\left(x = CN_{smooth,cell,chr}\right) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad (5)$$

$$P(k|x) = {pdf_k(x)}\Big/{\sum_{k=1}^{6} pdf_k(x)} \qquad (6)$$

Decision boundaries for the GMMs are calculated by finding the transitions between components, i.e., the point x where the PDFs of the components are equal. We evaluated copy numbers using GMMs including copy number components 1-5, throughout the study and 1-3 where indicated. Upper boundaries for component 5 were calculated using components 1-6. Theoretical sensitivity for each copy number component was calculated as the proportion of the component PDF that falls within its decision boundaries (i.e., true positive rate). 5-fold cross validation was performed by partitioning RPE1 cells into 5 equally sized groups and using each group once to evaluate a model generated using the remaining 4 groups. R scripts for copy number calling were compiled into an R package, karyotapR. karyotapR version 0.1 was used for analyses in this study.

Panel Simulations

For the probe downsampling simulation of chromosome 2, 50 samples each of *n* probes from the set of 24 probes targeting chr2 were generated where *n* = {4, 6, 8 … 20, 22}. Copy numbers were called for each set of probes for each cell. The sensitivity of the copy number calls was recalculated for the RPE1 cells as well as the theoretical sensitivity for all GMM components. This analysis was repeated for the set of probes targeting chr6 where *n* = {4, 6, 8 … 14, 16}.

For the simulation of an expanded custom panel, the set of 120 probes targeting chromosomes 1, 2, 3, 4, 5, and 6 were sampled 50 times to produce 50 sets of the 120 probes in unique orders. Starting with the first 4 probes of each set, copy numbers were called for RPE1 cells using those 4 probes, and again as each additional probe was added to the set until all 120 probes were used for the calculation. This procedure was repeated for each of the 50 sets. The sensitivity of copy number cells was recalculated for the RPE1 cells at each step as well as the theoretical sensitivity for all GMM components using a model with 5 components (with the upper boundary of the 5th component being calculated using a 6-component model), and with 3 components where noted.

Figure and Table Legends

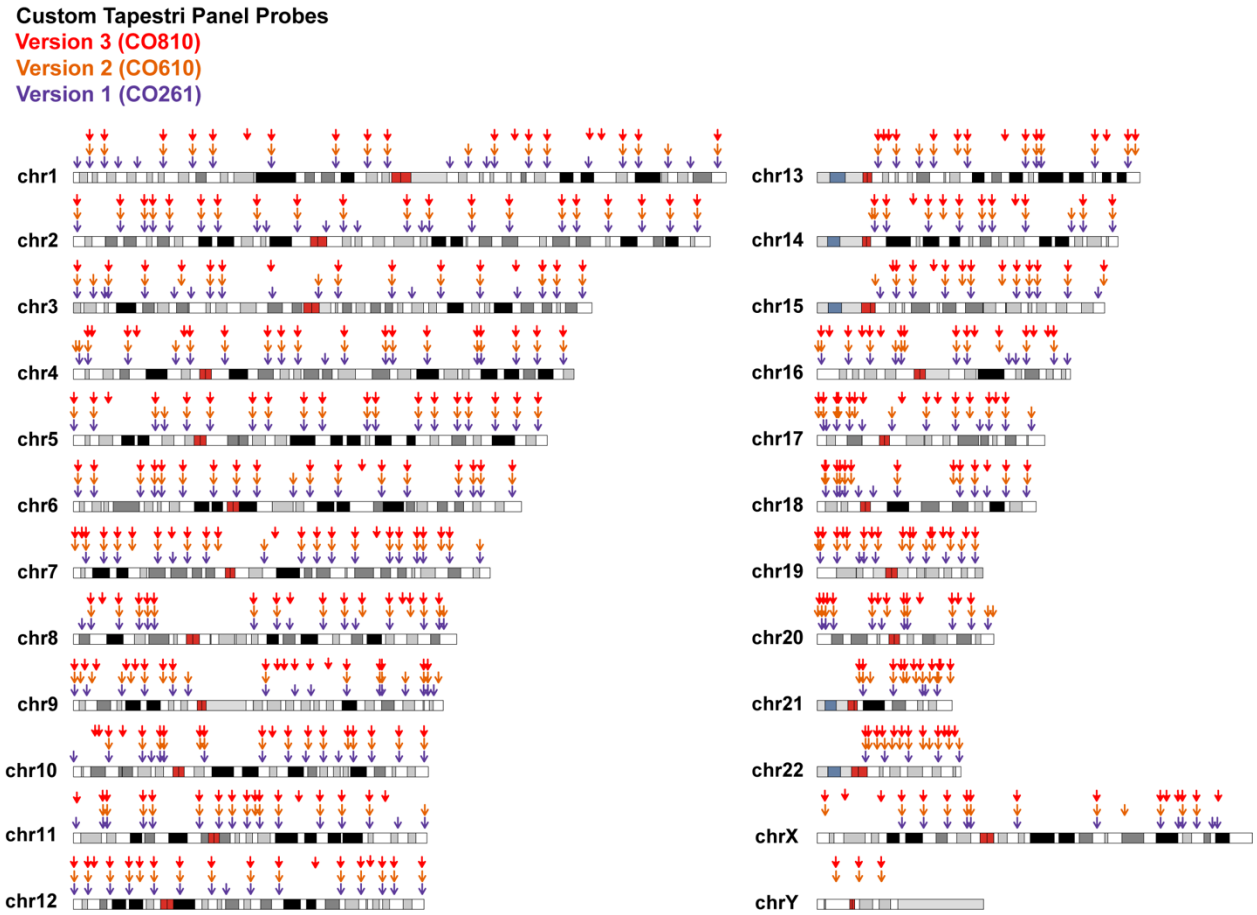


Figure 1

Map of PCR probe locations (arrows) for custom Tapestri panels for KaryoTap Version 1 (CO261) and Version 2 (CO610) on human genome hg19. Red blocks indicate centromeres, grayscale blocks indicate the G-band intensity of cytobands, and blue blocks indicate acrocentric chromosome arms.

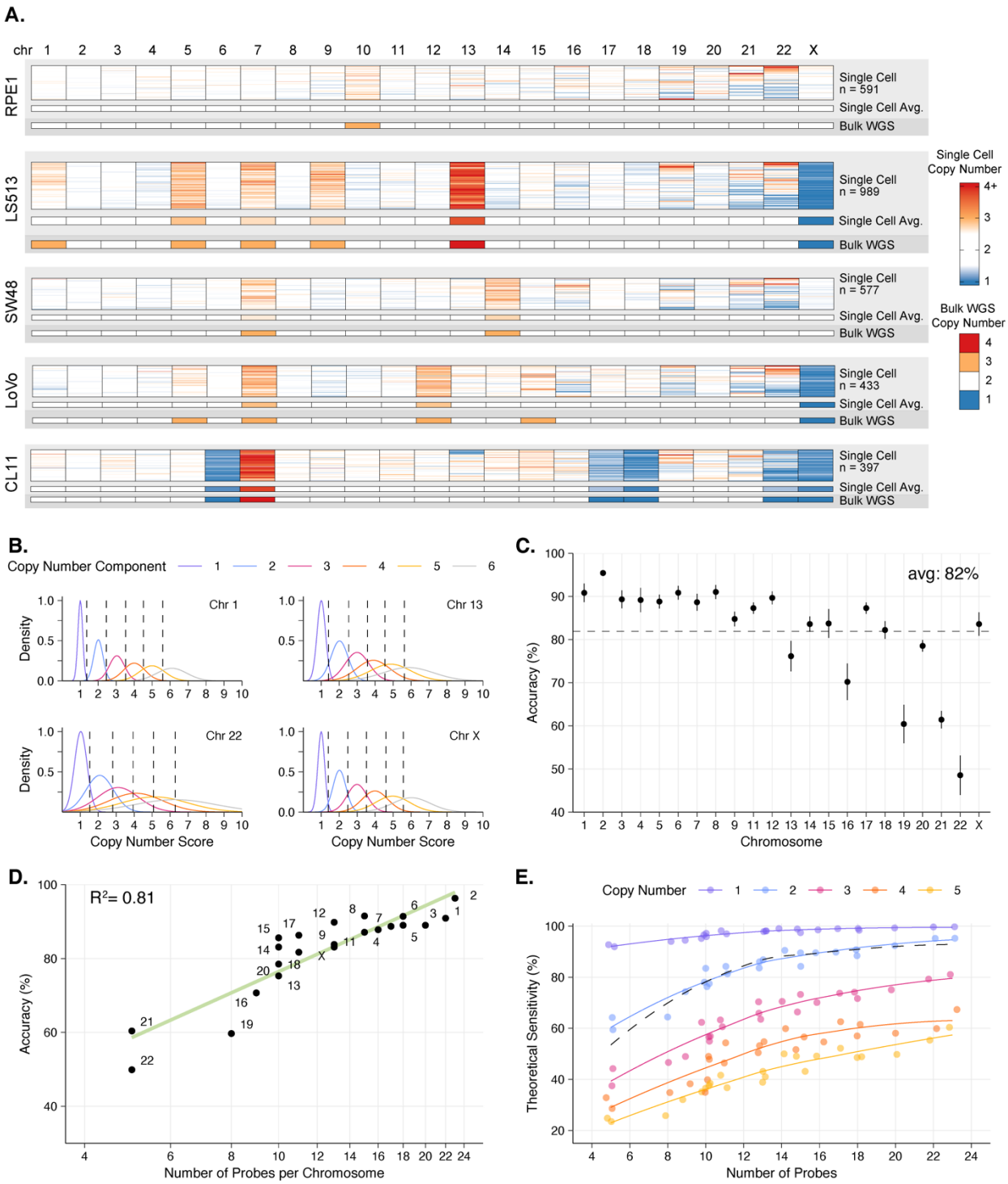
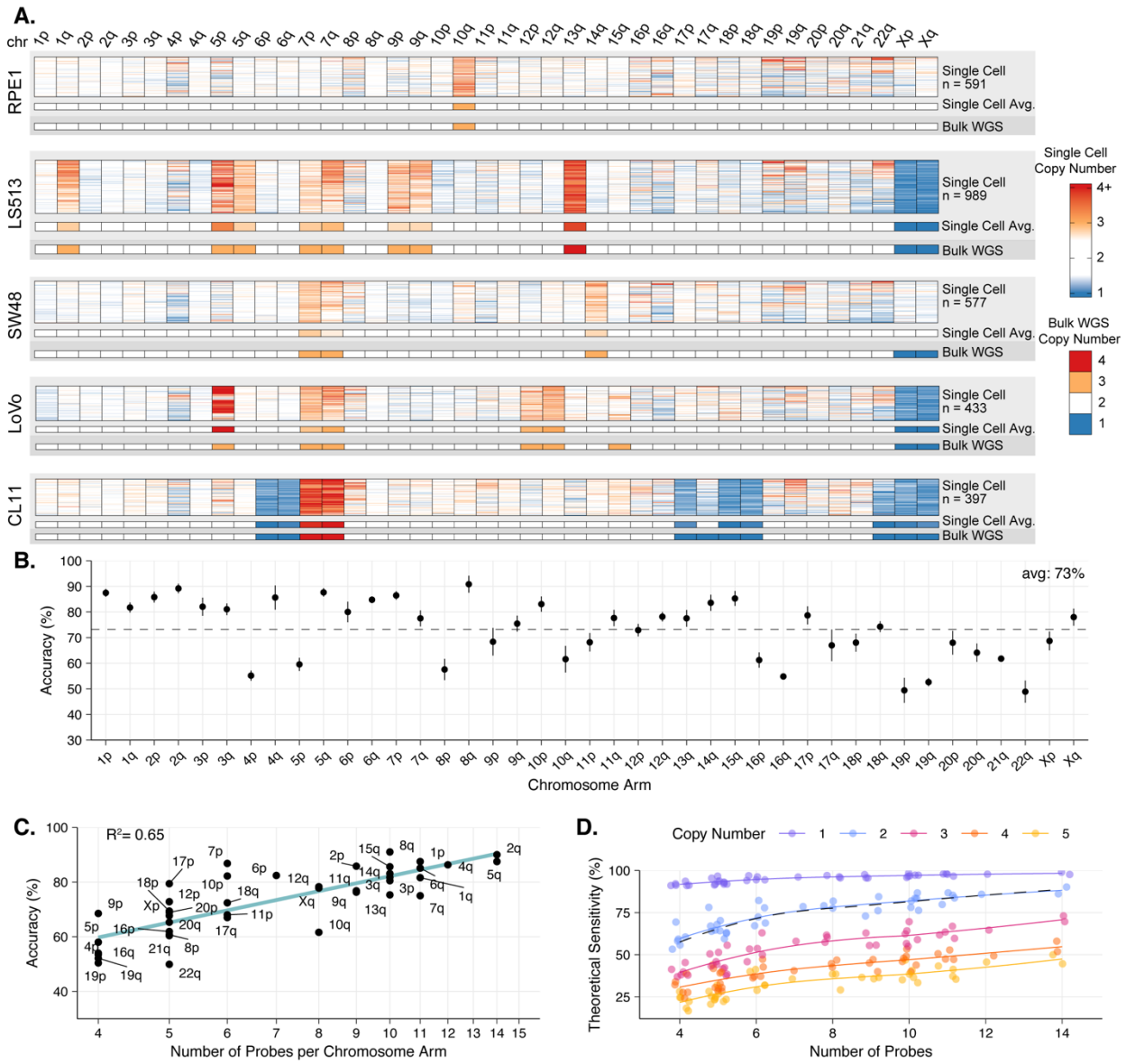


Figure 2

Copy number calling for whole chromosomes using KaryoTap panel Version 1. A) Heatmap of copy number scores for each cell-chromosome unit for 5 cell lines using custom Tapestri panel

Version 1. Upper blocks indicate copy number scores for each cell, middle blocks indicate average intensity of upper blocks, and lower blocks indicate copy number from bulk WGS (see Fig S1A). Half-filled lower blocks indicate chromosome arm-level aneuploidy. Number of cells included in single-cell blocks is indicated. B) Probability density functions of Gaussian mixture models (GMM) fit for chromosomes 1, 13, 22, and X using RPE1 cells. Dotted lines indicate decision boundaries between GMM components. C) 5-fold cross validation of copy number call accuracy for RPE1 cells by chromosome. Chr10 is omitted. Dot indicates mean accuracy and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average (avg) accuracy across chromosomes. D) Linear regression of RPE1 copy number call accuracy for each chromosome on number of probes per chromosome. X-axis is log-scaled to reflect log transformation of number of probes in regression. E) Theoretical copy number call sensitivity for each chromosome and copy number level calculated from GMMs. Points are slightly jittered horizontally to decrease overlapping.



921

922

923

924 Figure 3

925 Copy number calling for chromosome arms using custom KaryoTap panel Version 1. A) Heatmap

926 of copy number scores for each cell, smoothed across chromosome arms, for five cell lines using

927 custom Tapestri panel Version 1. Upper blocks indicate copy number scores for each cell, middle

928 blocks indicate average intensity of upper blocks, and lower blocks indicate copy number from

bulk WGS (see Fig S1A). B) 5-fold cross validation of copy number call accuracy for RPE1 cells by chromosome arm. Dot indicates mean accuracy and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average accuracy across chromosome arms. C) Linear regression of RPE1 copy number call accuracy for each chromosome arm on number of probes per chromosome arm. X-axis is log-scaled to reflect log transformation of number of probes in regression. D) Theoretical copy number call sensitivity for each chromosome arm and copy number level calculated from GMMs. Points are slightly jittered horizontally to decrease overlapping.

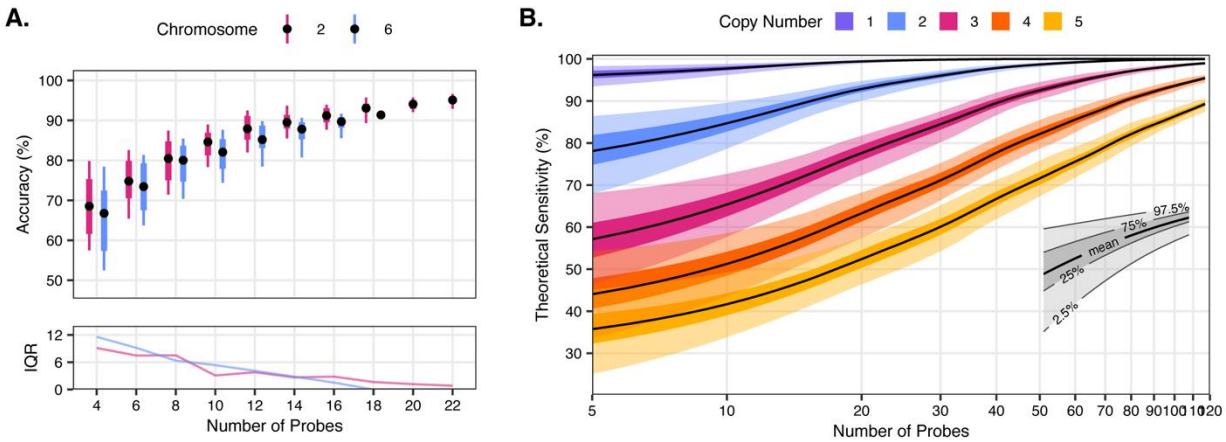


Figure 4

Effects on copy number call accuracy on probe sampling simulations. A) Box plots and inter quartile range (IQR) of accuracy from 50 probe downsampling simulations for chr2 and chr6. Boxes encompass middle 50%, whiskers encompass middle 95%, dot indicates median. B) Theoretical sensitivity of 50 panel simulations. Values for each copy number level were smoothed by Loess regression. The black line represents the mean, the darker inner shading indicates the middle 50% of the data, and the lighter outer shading represents the middle 95% of the data.

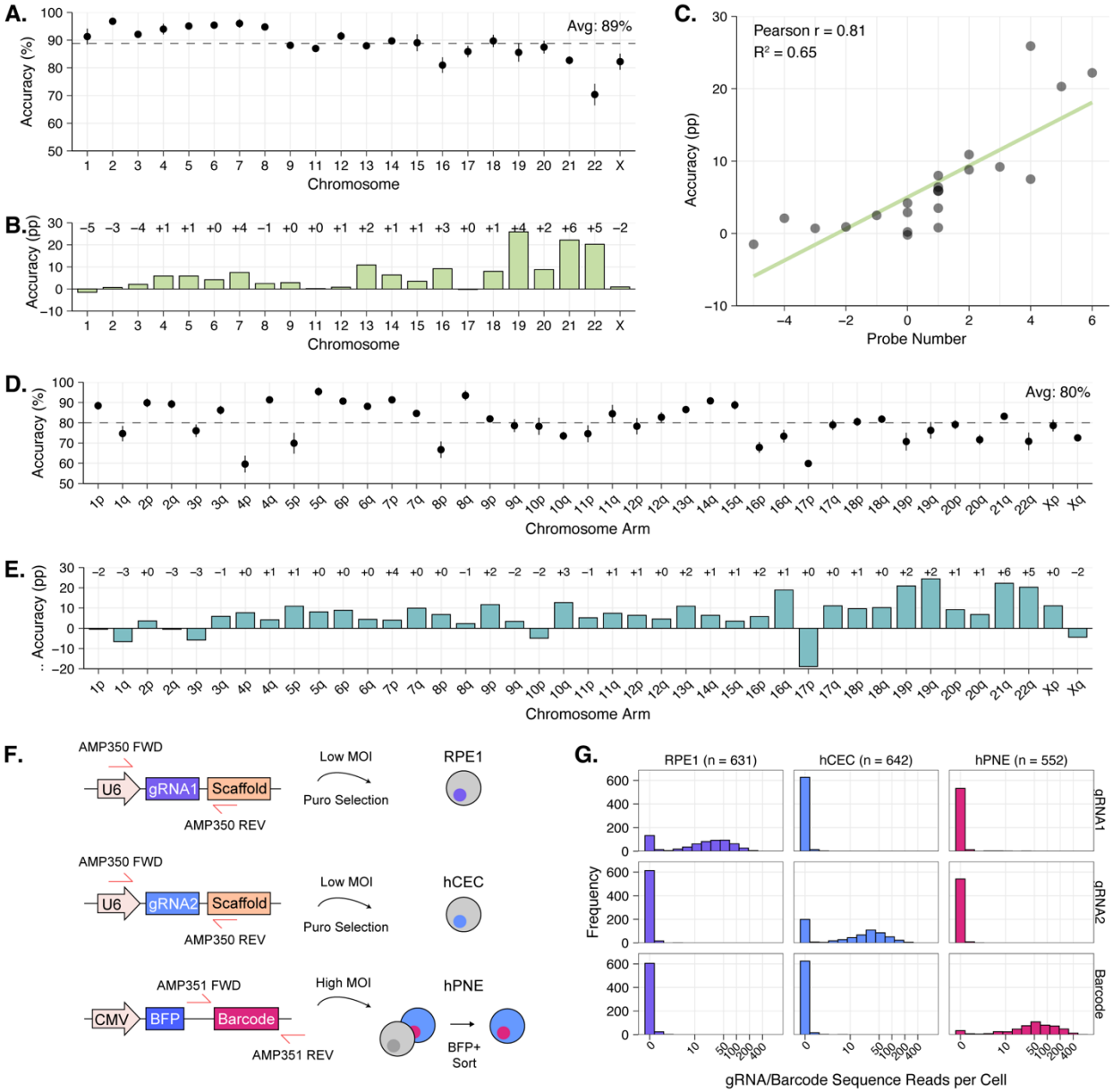


Figure 5

Copy number calling and lentiviral barcoding for custom KaryoTap panel Version 2. A) 5-fold cross validation of copy number call accuracy for n=631 RPE1 cells by chromosome using panel Version 2. Chr10 is omitted. Dot indicates mean accuracy and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average (avg) accuracy across chromosomes. B) Change in copy number call accuracy by chromosome for RPE1 cells between Version 2 and

Version 1 panels. Δ Probe Number is the difference in number of probes targeting a given chromosome between Version 2 and Version 1. pp: percentage points. C) Linear regression of the change in copy number call accuracy for RPE1 cells between Version 2 and Version 1 panels on the change in probe number targeting each chromosome. D) 5-fold cross validation of copy number call accuracy for RPE1 cells by chromosome arms using custom Tapestri panel Version 2. Dot indicates mean accuracy and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average across chromosome arms. E) Change in copy number call accuracy by chromosome arm for RPE1 cells between Version 2 and Version 1 panels. F) Plasmid constructs for lentiviral transduction of RPE1, hCEC D29, and hPNE cell lines. Probe pair AMP350 amplifies a 253 bp region including a CRISPR gRNA sequence and part of the U6 promoter and F+E scaffold. Probe pair AMP351 amplifies a 237 bp region including a barcode sequence. G) Number of reads in each cell that match the expected sequence of gRNA1, gRNA2, or barcodes in 3 transduced cell lines. X-axis is log-scaled. Number of cells of each cell line is indicated.

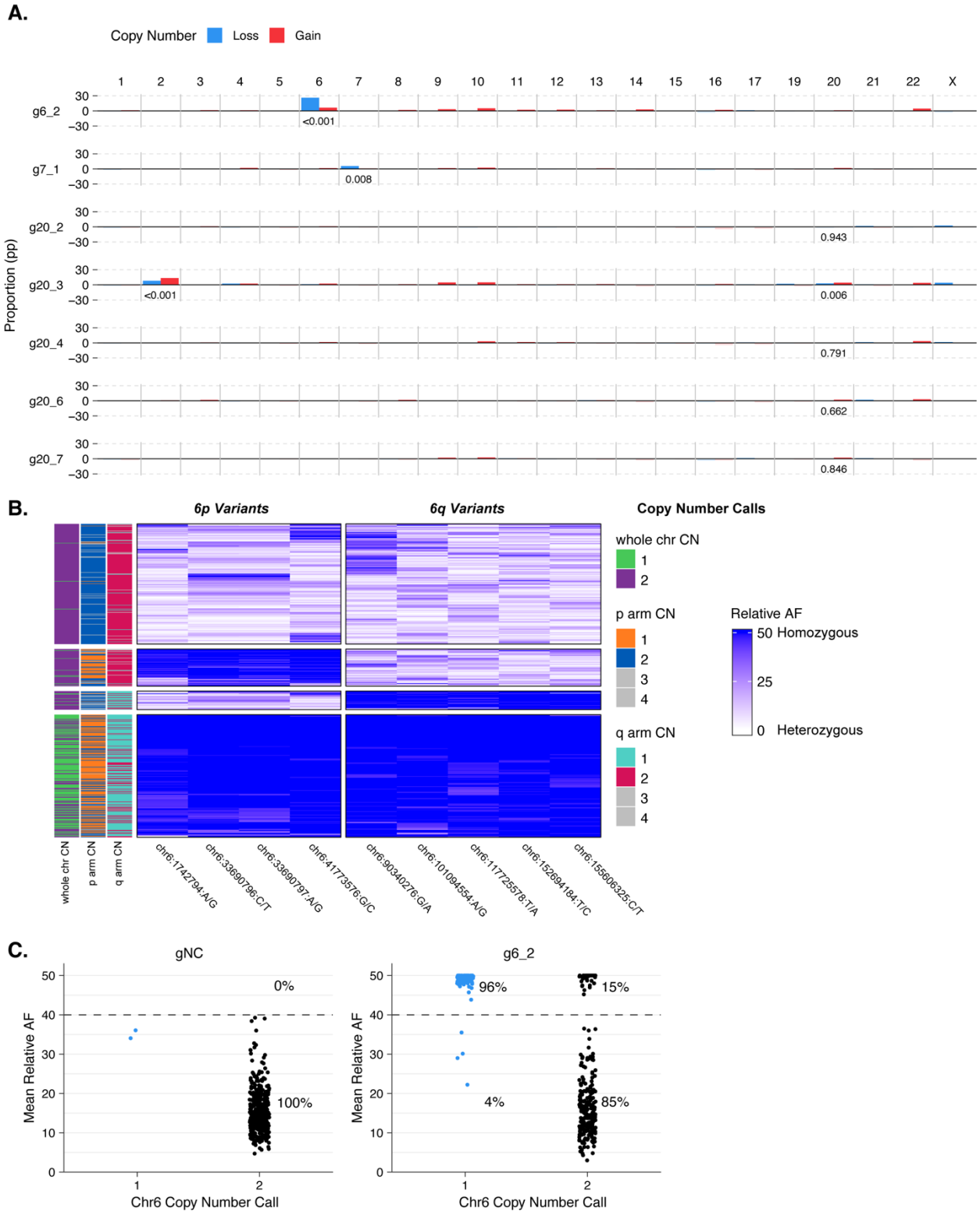

A.
Copy Number
Loss
Gain
Proportion (pp)
g6_2
g7_1
g20_2
g20_3
g20_4
g20_6
g20_7
<0.001
0.008
0.943
<0.001
0.006
0.791
0.662
0.846
B.
6p Variants
6q Variants
Copy Number Calls
whole chr CN
p arm CN
q arm CN
Relative AF
50 Homozygous
25
0 Heterozygous
chr6:1742794:A/G
chr6:33690796:C/T
chr6:33690797:A/G
chr6:41773576:G/C
chr6:90340276:G/A
chr6:101094554:A/G
chr6:117725578:T/A
chr6:152694184:T/C
chr6:155606325:C/T
C.
gNC
g6_2
Mean Relative AF
0%
100%
96%
4%
15%
85%
Chr6 Copy Number Call

Figure 6

Evaluation of KaryoCreate and Loss of Heterozygosity. A) Evaluation of KaryoCreate technology by KaryoTap panel Version 2. Bars represent the percentage point (pp) change (Δ) in the proportion of chromosomal losses (1 copy) and gains (3+ copies), compared to sgNC negative control (n = 398 cells). Chr18 was omitted due to additional copies of the chromosome in the cell line. p-values from Fisher's Exact test comparing the proportion of cells with copy number = 2 to copy number = {1, 3, or 4} (i.e., diploid vs. aneuploid) in each chromosome in each sample to the corresponding chromosome in the sgNC negative control sample are shown where $p < 0.1$. Additional p-values > 0.1 are reported where relevant. All p-values are reported in Table S2. Bars with negative values and $p > 0.1$ have reduced opacity for clarity. B) Heatmap of relative VAFs for sgChr6-2 treated cells with 1 or 2 copies of chr6 as called by GMM, for 9 originally heterozygous variants. Relative allele frequencies calculated as the absolute difference between raw allele frequency and 50%. 0 corresponds to balanced heterozygous alleles and 50 corresponds to fully homozygous alleles. Heatmap rows split by k-means clustering where k=4 and sorted by hierarchical clustering. CN: copy number. C) Mean relative VAFs for sgNC and sgChr6-2 treated cells with 1 or 2 copies of chr6 as called by GMM. 0%-40% AF indicates heterozygous haplotype, 40%-50% AF indicates homozygous haplotype.

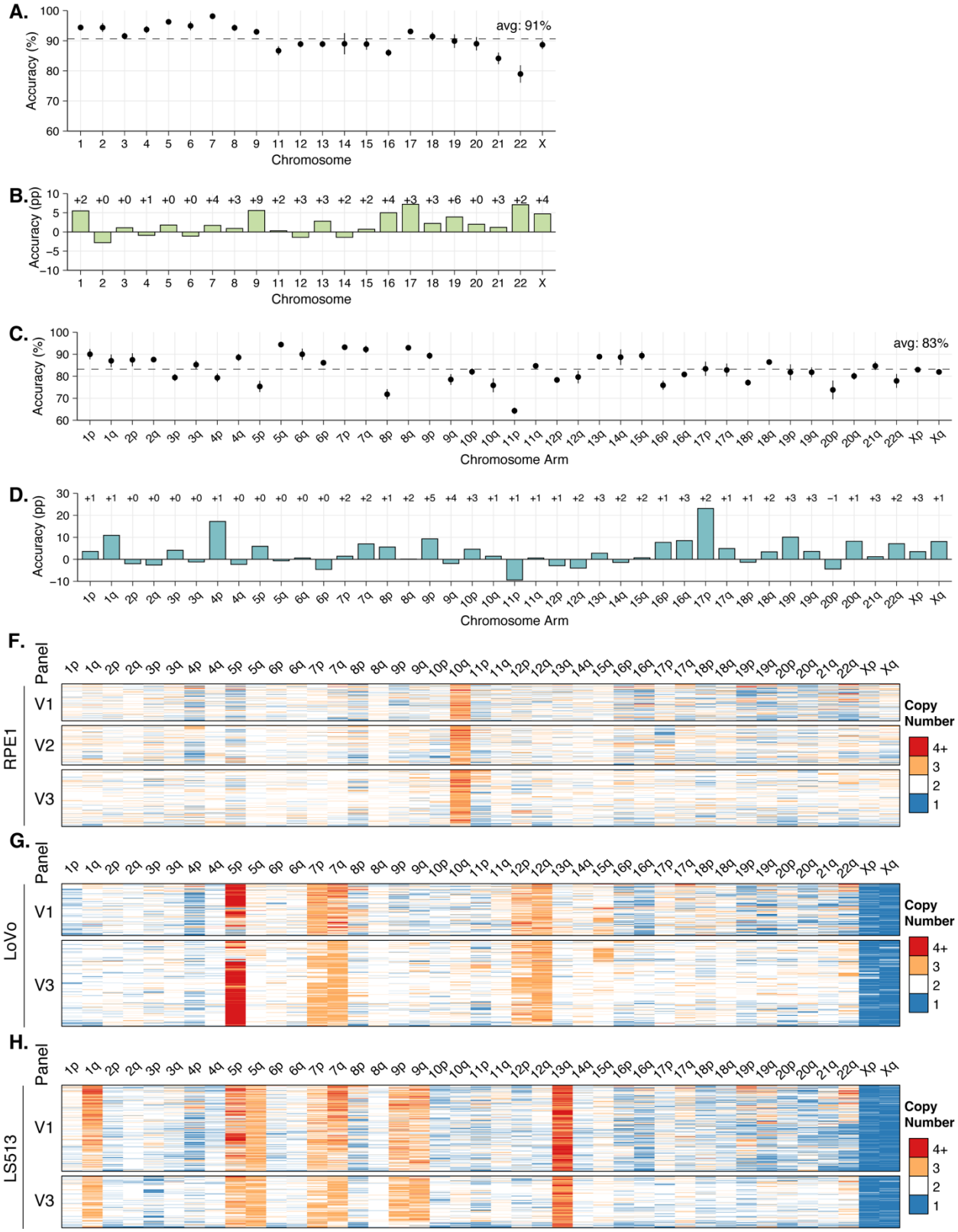

A.
Accuracy (%)
avg: 91%
Chromosome
B.
Accuracy (pp)
Chromosome
C.
Accuracy (%)
avg: 83%
Chromosome Arm
D.
Accuracy (pp)
Chromosome Arm
F.
Panel
RPE1
V1
V2
V3
Copy Number
4+
3
2
1
G.
Panel
LoVo
V1
V3
Copy Number
H.
Panel
LS513
V1
V3
Copy Number

Figure 7

Copy number calling and lentiviral barcoding for custom KaryoTap panel Version 3. A) 5-fold cross validation of copy number call accuracy for n=908 RPE1 cells by chromosome using panel Version 3. Chr10 is omitted. Dot indicates mean accuracy and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average (avg) accuracy across chromosomes. B) Change in copy number call accuracy by chromosome for RPE1 cells between Version 3 and Version 2 panels. Δ Probe Number is the difference in number of probes targeting a given chromosome between Version 3 and Version 2. pp: percentage points. C) 5-fold cross validation of copy number call accuracy for RPE1 cells by chromosome arms using custom Tapestri panel Version 3. Dot indicates mean accuracy and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average across chromosome arms. D) Change in copy number call accuracy by chromosome arm for RPE1 cells between Version 3 and Version 2 panels. E-G: Heatmaps of GMM copy number calls for chromosome arms using panels Version 1, 2, and 3 for cell lines RPE1, LoVo, and LS513.

<u>Supplemental Figure 1</u>

A) Bulk low-pass whole genome sequencing of RPE1, LS513, SW48, LoVo, and CL11 cell lines. Red highlight indicates amplification of at least one copy of highlighted segment; blue similarly indicates deletion. B) Representative g-banded karyogram of RPE1, indicating additional copy of chr10q translocated to the X chromosome (red arrow). C) UMAP projection of top 4 principal components of allele frequencies for N=2,986 cells representing 5 cell lines. Clustering was performed using the dbscan method. Cells were considered doublets if they were not members of the 5 largest clusters. D) PCA plot of first two principal components of mean allele frequencies for previously published deep sequencing of RPE1 and the 5 cell lines analyzed by scDNA-seq. E) Heatmap of cell-probe copy number values for five cell lines using custom Tapestri panel Version 1. Probes are organized by chromosome arm in genomic order.

Supplemental Figure 2

A) k-means clustering of arm-level copy number scores for n=2,986 using Panel Version 1. Color annotation indicates which cell line each row belongs to as determined by cell line SNPs. B) k-means clustering of arm-level copy number scores for LoVo cells (n=433) using Panel Version 1, at k = 2. C) Heatmap of GMM calls for whole chromosomes. D) Chromosome length (in 100 megabases) vs. accuracy of RPE1 panel Version 1 copy number calls chromosome. Chr10 is omitted. Trendline fit by linear regression. E) Heatmap of GMM calls for chromosome arms. using Panel Version 1 across 5 cells lines.

Supplemental Figure 3

A) Theoretical copy number call sensitivity from 50 probe downsampling simulations for chr2 and chr6 across five copy number levels. Boxes encompass middle 50%, whiskers encompass middle 95%, dot indicates median. B) Theoretical copy number call sensitivity of 50 panel simulations, using a 3 component GMM. Values for each copy number level were smoothed by Loess regression. The black line represents the mean, the darker inner shading indicates the middle 50% of the data, and the lighter outer shading represents the middle 95% of the data.

Supplemental Figure 4

A) Read counts per cell of four probes targeting chrY across 3 cell lines. B) UMAP projection of top 2 principal components of allele frequencies for N=2,347 cells representing 3 cell lines. Clustering was performed using the dbscan method. Cells were considered doublets if they were not members of the 3 largest clusters. C) Heatmap of GMM calls for whole chromosomes across 3 cell lines using panel Version 2. D) Heatmap of GMM calls for chromosome arms across 3 cell lines using panel Version 2. E) Mean theoretical accuracy of panel Version 2 copy number calls for each chromosome (left panel) or arm (right panel) at copy number values of 1, 2, 3, 4 and 5,

compared to those from panel Version 1. Dotted red line indicates $x = y$. F) Relationship between read counts per cell for either the gRNA or DNA Barcode probes (x-axis) and the number of reads per cell that match the specific gRNA or DNA Barcode sequences (y-axis). Dotted red line indicates $x = y$.

Supplemental Figure 5

A) Mean theoretical accuracy of panel Version 3 copy number calls for each chromosome at copy number values of 1, 2, 3, 4 and 5, compared to those from panel Version 2. Dotted red line indicates $x = y$. B) Mean theoretical accuracy of panel Version 3 copy number calls for each chromosome arm at copy number values of 1, 2, 3, 4 and 5, compared to those from panel Version 2.

Supplemental Figure 6

Overview of workflow for copy number analysis using karyotapR R package.

List of Abbreviations

BFP Blue florescent protein

CIN Chromosomal instability

CNV Copy number variant

DLP Direct library preparation

DOP-PCR Degenerate-oligonucleotide-primed PCR

FACS Fluorescence associated cell sorting

GMM Gaussian mixture model

gRNA guide RNA

hCECs Human colorectal epithelial cells

hPNEs Human pancreatic nestin-expressing cells

MALBAC Multiple annealing and looping-based amplification cycles

MDA Multiple displacement amplification

MOI Multiplicity of infection

ORF Open reading frame

PCA Principal components analysis

PCR Polymerase chain reaction

RPE1 Retinal pigment epithelial (cells) 1

scDNA-seq Single-cell DNA sequencing

scRNA-seq Single-cell RNA sequencing

sgRNA Single guide RNA

SNP Single-nucleotide polymorphism

SNV Single nucleotide variant

VAF Variant allele frequency

WGA Whole genome amplification

WGS Whole genome sequencing

Declarations

Ethics Approval and Consent to Participate

Not applicable

Consent for Publication

Not applicable

Availability of Data and Materials

Sequencing data for Tapestri experiments are available in the SRA repository under NCBI BioProject accession PRJNA950110, https://www.ncbi.nlm.nih.gov/bioproject/950110 (56). The

karyotapR package is available on GitHub at http://github.com/joeymays/karyotapR (57).The source code for karyptapR version 0.1 used for this study is archived on Zenodo under DOI https://doi.org/10.5281/zenodo.8305561(58). All data analysis scripts used in this study are available at https://github.com/joeymays/karyotap-publication (59) and are archived on Zenodo under DOI https://doi.org/10.5281/zenodo.8329277. Tapestri Pipeline output files used in this study are available on Zenodo under DOI https://doi.org/10.5281/zenodo.8305841(60).

Competing Interests

TD is on the Scientific Advisory Board of io9 and founder of KaryoVerse Therapeutics.

Funding

LJH and GRK were supported by NIH R37 CA240765. TD and members of her lab were supported by the Cancer Research UK Grand Challenge and the Mark Foundation for Cancer Research (C5470/A27144), NIH R37 R37CA248631, R01 R01HG012590, R01 R01DK135089, the MRA Young Investigator Award, the Breast Cancer Alliance Young Investigator Award and a grant from The National Foundation for Cancer Research.

Authors' contributions

JCM designed the study, performed wet lab experiments with help from others, performed bioinformatics analyses, developed the karyotapR software package, and wrote the manuscript. SM performed lentiviral preparation and transductions. NB and AG performed wet lab experiments. XZ designed Tapestri panels and performed bioinformatics analyses. JJB performed and supported cell culture. GRK prepared barcode vectors. LJH and DF supervised research. TD designed the study, edited the manuscript, and supervised research. MK and HMQ edited the manuscript. All authors read and approved the final manuscript.

Acknowledgments

NYU Langone's Genome Technology Center (RRID: SCR_017929) is supported by the Cancer Center Support Grant P30CA016087 at the Laura and Isaac Perlmutter Cancer Center. The computational requirements for this work were supported in part by the NYU Langone High Performance Computing (HPC) Core's resources and personnel. We thank Paolo Mita, Pan Cheng, Lizabeth Katznelson, and Maria Trifas for their helpful suggestions and insights in the preparation of this work.

Additional Files

- AdditionalFile01.csv: Probe Design for Panel Version 1 (CO216)
- AdditionalFile02.csv: Probe Design for Panel Version 2 (CO610)
- AdditionalFile03.csv: Probe Design for Panel Version 3 (CO810)

References

1. Navin N, Krasnitz A, Rodgers L, Cook K, Meth J, Kendall J, et al. Inferring tumor progression from genomic heterogeneity. Genome Res. 2010 Jan;20(1):68–80.

2. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. Nature Reviews Genetics. 2020 Jan;21(1):44–62.

3. Weaver BAA, Cleveland DW. Aneuploidy: Instigator and Inhibitor of Tumorigenesis. Cancer Research. 2007 Nov 1;67(21):10103–5.

4. Bakhoum SF, Ngo B, Laughney AM, Cavallo JA, Murphy CJ, Ly P, et al. Chromosomal instability drives metastasis through a cytosolic DNA response. Nature. 2018 Jan;553(7689):467–72.

5. Bakhoum SF, Cantley LC. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. Cell. 2018 Sep 6;174(6):1347–60.

6. Ippolito MR, Martis V, Martin S, Tijhuis AE, Hong C, Wardenaar R, et al. Gene copy-number changes and chromosomal instability induced by aneuploidy confer resistance to chemotherapy. Developmental Cell. 2021 Sep 13;56(17):2440-2454.e6.

7. Lee AJX, Endesfelder D, Rowan AJ, Walther A, Birkbak NJ, Futreal PA, et al. Chromosomal Instability Confers Intrinsic Multidrug Resistance. Cancer Research. 2011 Mar 1;71(5):1858–70.

8. Anagnostou V, Smith KN, Forde PM, Niknafs N, Bhattacharya R, White J, et al. Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non–Small Cell Lung Cancer. Cancer Discovery. 2017 Mar 5;7(3):264–76.

9. Lynch A, Bradford S, Burkard ME. The reckoning of chromosomal instability: past, present, future. Chromosome Res. 2024 Feb 17;32(1):2.

10. Bakker B, Taudt A, Belderbos ME, Porubsky D, Spierings DCJ, de Jong TV, et al. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. Genome Biology. 2016 May 31;17(1):115.

11. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer. 2017 Sep;17(9):557–69.

12. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature. 2011 Apr;472(7341):90–4.

13. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. Nature. 2014 Aug;512(7513):155–60.

14. Bosco N, Goldberg A, Zhao X, Mays JC, Cheng P, Johnson AF, et al. KaryoCreate: A CRISPR-based technology to study chromosome-specific aneuploidy by targeting human centromeres. Cell [Internet]. 2023 Apr 18 [cited 2023 Apr 20]; Available from: https://www.sciencedirect.com/science/article/pii/S0092867423003264

15. Truong MA, Cané-Gasull P, de Vries SG, Nijenhuis W, Wardenaar R, Kapitein LC, et al. A kinesin-based approach for inducing chromosome-specific mis-segregation in human cells. The EMBO Journal. 2023 May 15;42(10):e111559.

16. Tovini L, Johnson SC, Guscott MA, Andersen AM, Spierings DCJ, Wardenaar R, et al. Targeted assembly of ectopic kinetochores to induce chromosome-specific segmental aneuploidies. The EMBO Journal. 2023 May 15;42(10):e111587.

17. Telenius H, Carter NP, Bebb CE, Nordenskjo¨ld M, Ponder BAJ, Tunnacliffe A. Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. Genomics. 1992 Jul 1;13(3):718–25.

18. Arneson N, Hughes S, Houlston R, Done S. Whole-Genome Amplification by Degenerate Oligonucleotide Primed PCR (DOP-PCR). Cold Spring Harb Protoc. 2008 Jan 1;2008(1):pdb.prot4919.

19. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proceedings of the National Academy of Sciences. 2002 Apr 16;99(8):5261–6.

20. Zong C, Lu S, Chapman AR, Xie XS. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. Science. 2012 Dec 21;338(6114):1622–6.

21. Mallory XF, Edrisi M, Navin N, Nakhleh L. Methods for copy number aberration detection from single-cell DNA-sequencing data. Genome Biology. 2020 Aug 17;21(1):208.

22. Zahn H, Steif A, Laks E, Eirew P, VanInsberghe M, Shah SP, et al. Scalable whole-genome single-cell library preparation without preamplification. Nat Methods. 2017 Feb;14(2):167–73.

23. Yin Y, Jiang Y, Lam KWG, Berletch JB, Disteche CM, Noble WS, et al. High-Throughput Single-Cell Sequencing with Linear Amplification. Molecular Cell [Internet]. 2019 Sep 5 [cited 2019 Sep 6]; Available from: http://www.sciencedirect.com/science/article/pii/S1097276519306185

24. Laks E, McPherson A, Zahn H, Lai D, Steif A, Brimhall J, et al. Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. Cell. 2019 Nov 14;179(5):1207-1221.e22.

25. Evrony GD, Hinch AG, Luo C. Applications of Single-Cell DNA Sequencing. Annu Rev Genomics Hum Genet. 2021 Aug 31;22:171–97.

26. Mission Bio. Mission Bio Tapestri. 2020 [cited 2020 May 7]. Tapestri: The Precision Genomics Platform. Available from: https://missionbio.com/tapestri/

27. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016 Mar;17(3):175–88.

28. Crasta K, Ganem NJ, Dagher R, Lantermann AB, Ivanova EV, Pan Y, et al. DNA breaks and chromosome pulverization from errors in mitosis. Nature. 2012 Feb;482(7383):53–8.

29. Reijns MAM, Parry DA, Williams TC, Nadeu F, Hindshaw RL, Rios Szwed DO, et al. Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. Nature. 2022 Feb;602(7898):623–31.

30. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology. 2011 Feb 21;12(2):R18.

31. Wang K, Kumar T, Wang J, Minussi DC, Sei E, Li J, et al. Archival single-cell genomics reveals persistent subclones during DCIS progression. Cell. 2023 Aug 31;186(18):3968-3982.e15.

32. Jothilakshmi S, Gudivada VN. Chapter 10 - Large Scale Data Enabled Evolution of Spoken Language Research and Applications. In: Gudivada VN, Raghavan VV, Govindaraju V, Rao CR, editors. Handbook of Statistics [Internet]. Elsevier; 2016 [cited 2023 Aug 22]. p. 301–40. (Cognitive Computing: Theory and Applications; vol. 35). Available from: https://www.sciencedirect.com/science/article/pii/S0169716116300463

33. Loh PR, Genovese G, Handsaker RE, Finucane HK, Reshef YA, Palamara PF, et al. Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. Nature. 2018 Jul;559(7714):350–5.

34. Melcher R, Hartmann E, Zopf W, Herterich S, Wilke P, Müller L, et al. LOH and copy neutral LOH (cnLOH) act as alternative mechanism in sporadic colorectal cancers with chromosomal and microsatellite instability. Carcinogenesis. 2011 Apr 1;32(4):636–42.

35. Knudson AG. Mutation and Cancer: Statistical Study of Retinoblastoma. Proceedings of the National Academy of Sciences. 1971 Apr;68(4):820–3.

36. Shah JB, Pueschl D, Wubbenhorst B, Fan M, Pluta J, D'Andrea K, et al. Analysis of matched primary and recurrent BRCA1/2 mutation-associated tumors identifies recurrence-specific drivers. Nat Commun. 2022 Nov 7;13(1):6728.

37. Dong X, Zhang L, Hao X, Wang T, Vijg J. SCCNV: A Software Tool for Identifying Copy Number Variation From Single-Cell Whole-Genome Sequencing. Frontiers in Genetics [Internet]. 2020 [cited 2023 Jun 2];11. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2020.505441

38. Ryland GL, Doyle MA, Goode D, Boyle SE, Choong DYH, Rowley SM, et al. Loss of heterozygosity: what is it good for? BMC Medical Genomics. 2015 Aug 1;8(1):45.

39. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR–Cas9. Nat Rev Genet. 2015 May;16(5):299–311.

40. Ly P, Eskiocak U, Kim SB, Roig AI, Hight SK, Lulla DR, et al. Characterization of Aneuploid Populations with Trisomy 7 and 20 Derived from Diploid Human Colonic Epithelial Cells. Neoplasia. 2011 Apr 1;13(4):348-IN17.

41. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308–11.

42. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, et al. The UCSC Genome Browser database: 2023 update. Nucleic Acids Res. 2023 Jan 6;51(D1):D1188–95.

43. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017 Oct 1;33(19):3088–90.

44. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

45. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.

46. Kuilman T, Velds A, Kemper K, Ranzani M, Bombardelli L, Hoogstraat M, et al. CopywriteR: DNA copy number detection from off-target sequence data. Genome Biol. 2015 Feb 27;16:49.

47. Chen B, Gilbert LA, Cimini BA, Schnitzbauer J, Zhang W, Li GW, et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. Cell. 2013 Dec 19;155(7):1479–91.

48. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. Genome engineering using the CRISPR-Cas9 system. Nat Protoc. 2013 Nov;8(11):2281–308.

49. Sack LM, Davoli T, Li MZ, Li Y, Xu Q, Naxerova K, et al. Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns. Cell. 2018 Apr 5;173(2):499-514.e23.

50. Pagès H, Aboyoun P, Gentleman R, DebRoy S. Biostrings: Efficient manipulation of biological strings [Internet]. 2022. Available from: https://bioconductor.org/packages/Biostrings

51. Morgan M, Pagès H, Obenchain V, Hayden N. Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import [Internet]. 2022. Available from: https://bioconductor.org/packages/Rsamtools

52. Reijns MAM. WT1_ancestral_culture [Internet]. SRA; 2021 [cited 2023 Jan 18]. Available from: https://www.ncbi.nlm.nih.gov/sra/?term=ERR7477340

53. GATK [Internet]. [cited 2023 Mar 21]. Available from: https://gatk.broadinstitute.org/hc/en-us

54. Knaus BJ, Grünwald NJ. vcfr: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources. 2017;17(1):44–53.

55. Delignette-Muller ML, Dutang C. fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software. 2015 Mar 20;64:1–34.

56. Mays J, Davoli T. ID 950110 - BioProject - NCBI [Internet]. 2023 [cited 2023 Sep 12]. Available from: https://www.ncbi.nlm.nih.gov/bioproject/950110

57. Mays JC. karyotapR: CNV Analysis for Tapestri [Internet]. 2023 [cited 2023 Sep 8]. Available from: https://github.com/joeymays/karyotapR

58. Mays J. joeymays/karyotapR: Version 0.1 [Internet]. Zenodo; 2023 [cited 2023 Sep 8]. Available from: https://zenodo.org/record/8305561

59. Mays J. KaryoTap Publication Codebook [Internet]. 2023 [cited 2023 Sep 8]. Available from: https://github.com/joeymays/karyotap-publication

60. Mays J, Davoli T. KaryoTap Enables Aneuploidy Detection in Thousands of Single Human Cells [Internet]. Zenodo; 2023 [cited 2023 Sep 8]. Available from: https://zenodo.org/record/8305841