KaryoTap Enables Aneuploidy Detection in Thousands of Single Human Cells

Joseph C. Mays[1], Sally Mei[1], Nazario Bosco[1,2], Xin Zhao[1], Joy J Bianchi[1], Gururaj Rao Kidiyoor[1], Liam J. Holt[1], Teresa Davoli[1]

[1] Institute for Systems Genetics and Department of Biochemistry and Molecular Pharmacology, NYU School of Medicine, New York, NY 10016, USA.
[2] Current address: Volastra Therapeutics, New York, NY 10027, USA.

Correspondence: teresa.davoli@nyulangone.org

Abstract

Investigating chromosomal instability and aneuploidy within tumors is essential for understanding their contribution to tumorigenesis and developing effective diagnostic and therapeutic strategies. Single-cell DNA sequencing (scDNA-seq) technologies have enabled such analyses, revealing aneuploidies specific to individual cells within the same tumor. However, scaling the throughput of these methods to identify aneuploidies occurring at low frequencies while maintaining high sensitivity has been difficult. To overcome this, we developed KaryoTap, a method combining custom targeted panels for the Tapestri scDNA-seq platform with a Gaussian mixture model analysis framework to enable detection of chromosome- and chromosome arm-scale aneuploidy in all human chromosomes across thousands of single cells simultaneously. This system will prove a powerful and flexible resource for the study of aneuploidy and chromosomal instability in tumors and normal tissues.

Introduction

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]. 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[2,3]. 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[4–6]. 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[7–11]. 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[6,9]. Methods for modeling aneuploidies of specific chromosomes in cell culture have also emerged as powerful tools for studying the effects of aneuploidy in cancer[12–14]. These methods have benefitted from scDNA-seq, as sequencing individual cells enables the evaluation of the specificity and accuracy of the engineered karyotypes[13,14].

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[15], MDA[16], and MALBAC[17] 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[18]. More recent methods vastly improve throughput by using microfluidic partitioning[19], droplet encapsulation[20], and liquid handling robots[21], allowing hundreds of thousands of cells to be sequenced at once. However, these methods suffer from uneven sequencing coverage and low depth across the genome, jeopardizing the confident detection of aneuploidy in individual cells[22]. Furthermore, the need for custom hardware, expensive liquid handlers, and complicated protocols makes these methods difficult to adopt for most laboratories[18].

To address the need for a high-throughput, high-sensitivity method for detecting chromosome-scale aneuploidy across the human genome, 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[23]. The platform is commonly used to detect tumor hotspot mutations in cancer driver genes and has not yet been utilized to identify aneuploidy across the genome. While targeted sequencing is typically used to identify mutations in single nucleotides[24], 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. 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.

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**). 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 gigabase 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[4,25], 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[26] 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[27]. 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 median of the cell-probe copy number values for all probes on a given chromosome. 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. Altogether, these data demonstrate that our Version 1 panel can resolve the copy number of populations of cells at the whole-chromosome level.

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)[28], 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 use 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 (see Methods for details).

To evaluate copy number calling performance, we determined the sensitivity of detection, calculated as the proportion of calls that were correct (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[4]; 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**). Sensitivity, or correctly identifying 2 copies of a chromosome in RPE1, varied between chromosomes and ranged from 95% for chr2 to 46% for chr22 with a mean of 79% (**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 sensitivity five times, each time reserving one of the sets from the model generation and using it only to calculate sensitivity. The mean absolute deviation of the 5 sensitivity measurements for each chromosome ranged from 0.61 to 4.75 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. Sensitivity for each chromosome correlated strongly with chromosome length (Pearson r = 0.72), and chromosome length itself correlated strongly with the number of probes targeting the chromosome (Pearson r = 0.93) (**Fig S2A**), suggesting that sensitivity is related to the number of probes targeting a chromosome. As expected, linear regression of sensitivity on the number of probes per chromosome indicated that the number of probes is predictive of copy number call sensitivity (R2 = 0.76; p = 1.18e-07) (**Fig 2D**). This suggests that the sensitivity 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 sensitivity 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 calling 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 82% and 3 copies had 62% (**Fig S2B**). The theoretical sensitivity for 2 copies strongly correlated with the empirical sensitivity for RPE1 calls (Pearson r = 0.96). As expected, theoretical sensitivity at all 5 copy number levels decreased for chromosomes with fewer probes as was the case with the empirical sensitivity (**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 sensitivity for correctly calling copy numbers in RPE1. The sensitivity ranged from 88% (chr2q) to 44% (chr19p) with a mean of 69%. The mean absolute deviation of sensitivity from 5-fold cross validation ranged from 0.82 to 6.2 percentage points. 16 out of 41 arms had sensitivity values of at least 75% (**Fig 3B**), demonstrating generally lower sensitivity compared to whole chromosomes, likely due to the smaller number of probes that typically target an arm compared to a whole chromosome. A positive relationship between the number of probes and sensitivity was again revealed by linear regression (R2 = 0.59; p = 5.1e-09) (**Fig 3C**). We calculated the theoretical sensitivity across the 5 copy number levels. Sensitivity was again highest for 1 copy with an average of 94%; 2 copies had an average of 72% and 3 copies had 52% (**Fig 3D; Fig S2C**). 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 Sensitivity

To confirm that copy number classification 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 (24 total probes) and 6 (18 total probes) and recalculated the classification sensitivity for the RPE1 cells. 50 samples of *n* probes were evaluated for each value of *n*. Consistent with our findings above, median sensitivity decreased from 93.9% (24 probes) to 66.8% (4 probes) for chr2 and from 89.0% (18 probes) to 65.2% (4 probes) for chr6, indicating that probe number, not chromosome size, affects classification sensitivity (**Fig 4A**). Furthermore, the interquartile range (IQR) of the sensitivity distributions increased from 1.7 percentage points (pp; 22 probes) to 10.3 pp (4 probes) for chr2 and from 2.0 pp (16 probes) to 12.5 pp (4 probes) for chr6 indicating that having fewer probes per chromosome increases the variability of classification sensitivity. Both the decrease in sensitivity and increase in sensitivity 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, 20 probes for 2 copies, 50 probes for 3 copies, and 90 probes for 4 copies. A maximum mean sensitivity of 87% 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, 20 probes for 2 copies, and 26 probes for ≥3 copies (**Fig S3B**). Furthermore, 99% sensitivity could be achieved at 22 probes for 1 copy, 66 probes for 2 copies, and 76 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.

Read Depth Does Not Affect Sensitivity

We reasoned that the sequencing depth of the scDNA-seq dataset could affect copy number call sensitivity since increasing depth should reduce the overall variance in read counts. To investigate this, we downsampled the read counts from our experiment to 80%, 60%, 40%, and 20% of average read depth (68, 51, 34, and 17 reads per cell per probe) from full depth at 85 reads per cell per probe, repeated for 50 trials at each downsampling level. Copy numbers for the RPE1 cells were called for each downsampled dataset. Mean theoretical sensitivity decreased as depth decreased, with an average decrease of 0.4 percentage points (pp) at 80% depth, 0.8 pp at

60%, 1.8 pp at 40%, and 4.4 pp at 20% depth across the 5 copy number levels (**Fig 4C**). Analysis of variance indicated that the effects of sequencing depth and the number of probes on sensitivity were independent (ANOVA interaction, p = 0.9-1 for all copy number levels). Considering the difference in sensitivity from full depth to 80%, 60%, and 40% depth is small (<2 percentage points), we would not expect that increasing the sequencing depth will result in meaningful gains in copy number call sensitivity for any copy number level.

KaryoTap Version 2 Panel Increases Sensitivity of Detecting Aneuploidy

As the sensitivity of our custom panel increases with the number of probes targeting a chromosome, we attempted to improve the sensitivity 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 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 sensitivity 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 fiv5e equally sized subsets and calculating sensitivity 5 times, each time reserving one of the sets from the model generation and using it only to calculate sensitivity. The mean absolute deviation of the 5 sensitivity measurements for each chromosome ranged from 0.49 to 4.74, suggesting that copy number calling performance would be maintained when classifying new data. The poorest performing chromosome in Version 1, chr22, had a sensitivity of 62% with Version 2 compared to 46% for Version 1 (**Fig 5A-B**). The mean sensitivity across all chromosomes was 85%, increased from 79% for Version 1. Sensitivity increased by 1.8 pp on average for each additional probe (**Fig 5C**). Similarly, for chromosome arms, the average sensitivity across arms increased from 69% with Version 1 to 76% with Version 2 (**Fig 5D-E**). 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 S4C-D**). 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 98% for losses, 88% for neutral states, and 85% for gains for whole chromosomes, and 96% for losses, 79% for neutral states, and 77% for gains for chromosome arms (**Fig S4D**). 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), 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 S4E**) 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 S4E**), 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 virtually all read counts from Probe AMP351 matched the number of reads matching the known barcode sequences in hPNE (**Fig S4E**), 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, a method we recently developed to induce chromosome-specific aneuploidy in cultured cells[12]. 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. 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 6; Table S2**; $p < 0.001$, Fisher's exact test). sgChr6-2 induced 30.8%

losses compared to 1.3% with sgNC, and 9.6% gains compared to 4.3% with sgNC (**Table S3**). sgChr7-1 induced 6.5% losses compared to 1% with sgNC, and 7.5% gains compared to 3.8% with sgNC. In the same experiment, we additionally tested 5 sgRNAs targeting chr20 that had not been previously evaluated. sgChr20-1, 20-2, 20-4, 20-5 and 20-6 did not induce changes in chr20 ($p = 0.583$-1.00). To our surprise, sgChr20-3 induced 9.2% losses and 18.5% gains in chr2 ($p < 0.001$), which we may not have observed if we had instead evaluated the gRNA using a chromosome-targeted method such as fluorescence in situ hybridization rather than a genome-wide method. The sgChr20-3 sequence matches 18 out of 20 base pairs for sites on the chr2 centromere[12], suggesting an explanation of the off-target effect. These data indicate that our panel can be used to efficiently screen CRISPR genome engineering methods like KayroCreate and successfully use gRNA detection to multiplex samples.

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 cells in one experiment, greatly improving the throughput of existing plate-based whole genome amplification methods[15–17] while retaining the sensitivity to confidently detect chromosomal gains and losses in each individual cell. 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 S5**).

While single-cell aneuploidy detection is not unique to KaryoTap, our design has several advantages, the most critical being the ability to scale up throughput to thousands of cells in one experimental run, thus significantly reducing hands-on time, reagent cost, and sequencing cost per cell. 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[24]. Furthermore, the total number of sequencing reads needed to obtain high-sensitivity 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 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[20].

As evidenced by the analysis of RPE1, our method will generate a range of smooth copy number measurements for chromosomes with the same 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[29]. 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 (i.e., sensitivity). As the number of probes increases, the variance of the model components decreases, resulting in an increase in sensitivity 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.

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[20]. 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, further reducing experimental costs. 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, 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[30]. LOH has been demonstrated to promote tumorigenesis by inactivating tumor suppressor genes through chromosomal loss[31,32]. 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[30].

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[20]. 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[33].

Our system in its current form is limited to calling whole chromosome aneuploidy, and, with less confidence, chromosome arm-level aneuploidy. 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 sensitivity.

Coupled with the Tapestri platform, KaryoTap shows promise as an easily adoptable, flexible, and highly scalable method for detecting aneuploidy in thousands of single 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.

Figure & 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 sensitivity for RPE1 cells by chromosome. Chr10 is omitted. Dot indicates mean sensitivity and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average (avg) sensitivity across chromosomes. D) Linear regression of RPE1 copy number call sensitivity 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.

Figure 3

Copy number calling for chromosome arms using custom KaryoTap panel Version 1. A) Heatmap of copy number scores for each cell, smoothed across chromosome arms, for five 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). B) 5-fold cross validation of copy number call sensitivity for RPE1 cells by chromosome arm. Dot indicates mean sensitivity and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average sensitivity across chromosome arms. C) Linear regression of RPE1 copy number call sensitivity 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 sensitivity on probe sampling and sequencing depth simulations. A) Box plots and inter quartile range (IQR) of sensitivity 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. C) Sensitivity from read depth downsampling simulations for RPE1 cells. Each dot is the mean theoretical sensitivity for 50 trials. Lines represent linear regression for each downsampling level.

Figure 5

Copy number calling and lentiviral barcoding for custom KaryoTap panel Version 2. A) 5-fold cross validation of copy number call sensitivity for RPE1 cells by chromosome using panel Version 2. Chr10 is omitted. Dot indicates

mean sensitivity and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average (avg) sensitivity across chromosomes. B) Change in copy number call sensitivity 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 sensitivity 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 sensitivity for RPE1 cells by chromosome arms using custom Tapestri panel Version 2. Dot indicates mean sensitivity and lines indicate ± mean absolute deviation. Horizontal dotted line indicates average across chromosome arms. E) Change in copy number call sensitivity 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.

Figure 6

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 copy number gains and losses in each chromosome in each sample to the corresponding chromosome in the sgNC negative control sample are shown where $p < 0.2$. Some p-values > 0.2 are reported where relevant. All p-values are reported in Table S2. Bars with negative values and $p > 0.2$ have reduced opacity for clarity.

Supplemental Figure 1

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) Chromosome length (in 100 megabases) vs. sensitivity of RPE1 panel Version 1 copy number calls chromosome. Chr10 is omitted. Trendline fit by linear regression. B) Theoretical sensitivity of panel Version 1 copy number calls for each chromosome at copy number values of 1, 2, 3, 4 and 5, calculated from GMMs. Average (avg) sensitivity across chromosomes is indicated for each copy number level. C) Theoretical sensitivity of panel Version 1 copy number calls for each chromosome arm at copy number values of 1, 2, 3, 4 and 5, calculated from GMMs. Average sensitivity across chromosome arms is indicated for each copy number level.

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) Mean theoretical sensitivity of panel Version 2 copy number calls for each chromosome at copy number values of 1, 2, 3, 4 and 5, compared to those from panel Version 1. Dotted red line indicates $x = y$. D) Theoretical sensitivity of panel Version 2 copy number calls for each chromosome and chromosome arm for both 3- and 5- component GMMs. E) 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

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

Supplemental Table 1

Number of probes per chromosome arm for custom Tapestri panels.

Supplemental Table 2

Results of Fisher's exact test for chromosomal gains and losses induced by KaryoCreate. p-values from Fisher's Exact test comparing the proportion of copy number gains and losses in each chromosome in each sample to the corresponding chromosome in the sgNC negative control sample are given.

Supplemental Table 3

Percentage of chromosomal gains and losses Induced by KaryoCreate. Percentage indicates the proportion of cells in each KayroCreate-treated sample with either 1, 2, or 3+ copies of each chromosome.

Supplemental Table 4

Sequences of barcodes and gRNAs.

Methods

Data and Code Availability

The current stable version of the karyotapR package is available on CRAN at https://CRAN.R-project.org/package=karyotapR, and the development version and source code are available on Github at http://github.com/joeymays/karyotapR. Documentation and analysis tutorials are available at https://joeymays.xyz/karyotapR. 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. All analysis scripts are available at https://github.com/joeymays/karyotap-publication; the script versions used for this study are archived on Zenodo under DOI https://doi.org/10.5281/zenodo.8329277. Raw sequencing data and aligned reads are available at SRA under NCBI BioProject accession PRJNA950110.

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) and SW48 cells were incubated in DMEM, supplemented with 10% FBS. LoVo cells were incubated in Ham's F12-K media with 10% FBS. LS513 cells and hTERT human pancreatic nestin-expressing cells (hPNEs) were incubated in RPMI media with 10% FBS. CL11 cells were incubated in DMEM:F12 and 20% FBS. hTERT p53-/- human colonic epithelial cells (hCECs) 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.

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[34,35] (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 (https://portal.missionbio.com) 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 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 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[36].

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)[23]. 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[37], aligned to the hg38 genome using bwa-mem[38], and analyzed for copy number variants using the CopywriteR[39] 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[40]. 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[41]. 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[42]. 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[12]. 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[12]. To compare conditions, Fisher's exact test was performed in R using the `fisher.test()` function. For multiple comparisons, Bonferroni correction 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[43], with tolerance for up to 2 base mismatches. BAM files were manipulated using the Rsamtools R package[44].

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[26]. 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[45]. The vcfR[46] R package 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.

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

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

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

Smooth copy number scores for each chromosome in each cell (cell-chromosome scores) were generated by taking the median of the probe-specific copy number values for probes targeting a common chromosome (Equation 4). This was also modified to calculate cell-chromosome-arm scores for probes common to a chromosome arm.

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

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[47]. 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.

For the read depth simulation, the read count matrix was randomly downsampled to 80%, 60%, 40%, and 20% of the total read counts using the downsampleMatrix function from the scuttle R package[48]. 50 downsampled matrices were generated for each downsampling level. 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 5-component model. Least squares linear regression was performed at each copy number level in R with the formula 'sensitivity ~ log(n.probes) + depth` where depth was a categorical variable.

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, Aleah Goldberg, Pan Cheng, Lisa Katsnelson, and Maria Trifas for their helpful suggestions and insights in the preparation of this work. T.D. was supported by the Cancer Research UK Grand Challenge and the Mark Foundation for Cancer Research (C5470/A27144), R37 R37CA248631, the MRA Young Investigator Award.

Supplemental Tables

- Table S1: Number of probes per chromosome arm for custom Tapestri panels
- Table S2: Results of Fisher's exact test for Karyocreate samples
- Table S3: Percentage of chromosomal gains and losses induced by KaryoCreate
- Table S4: Sequences of barcodes and gRNAs

Additional Files

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

References

1. Navin, N. *et al.* Inferring tumor progression from genomic heterogeneity. *Genome Res* **20**, 68–80 (2010).
2. Bakhoum, S. F. *et al.* Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* **553**, 467–472 (2018).
3. Bakhoum, S. F. & Cantley, L. C. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. *Cell* **174**, 1347–1360 (2018).
4. Ippolito, M. R. *et al.* Gene copy-number changes and chromosomal instability induced by aneuploidy confer resistance to chemotherapy. *Developmental Cell* **56**, 2440-2454.e6 (2021).
5. Lee, A. J. X. *et al.* Chromosomal Instability Confers Intrinsic Multidrug Resistance. *Cancer Research* **71**, 1858–1870 (2011).
6. Anagnostou, V. *et al.* Evolution of Neoantigen Landscape during Immune Checkpoint Blockade in Non–Small Cell Lung Cancer. *Cancer Discovery* **7**, 264–276 (2017).
7. Lynch, A. R. *et al.* A survey of CIN measures across mechanistic models. 2023.06.15.544840 Preprint at https://doi.org/10.1101/2023.06.15.544840 (2023).
8. Bakker, B. *et al.* Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biology* **17**, 115 (2016).
9. Baslan, T. & Hicks, J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. *Nat Rev Cancer* **17**, 557–569 (2017).
10. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–94 (2011).
11. Wang, Y. *et al.* Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* **512**, 155–160 (2014).
12. Bosco, N. *et al.* KaryoCreate: A CRISPR-based technology to study chromosome-specific aneuploidy by targeting human centromeres. *Cell* (2023) doi:10.1016/j.cell.2023.03.029.
13. Truong, M. A. *et al.* A kinesin-based approach for inducing chromosome-specific mis-segregation in human cells. *The EMBO Journal* **42**, e111559 (2023).
14. Tovini, L. *et al.* Targeted assembly of ectopic kinetochores to induce chromosome-specific segmental aneuploidies. *The EMBO Journal* **42**, e111587 (2023).

15. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).

16. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proceedings of the National Academy of Sciences* **99**, 5261–5266 (2002).

17. Zong, C., Lu, S., Chapman, A. R. & Xie, X. S. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science* **338**, 1622–1626 (2012).

18. Mallory, X. F., Edrisi, M., Navin, N. & Nakhleh, L. Methods for copy number aberration detection from single-cell DNA-sequencing data. *Genome Biology* **21**, 208 (2020).

19. Zahn, H. *et al.* Scalable whole-genome single-cell library preparation without preamplification. *Nat Methods* **14**, 167–173 (2017).

20. Yin, Y. *et al.* High-Throughput Single-Cell Sequencing with Linear Amplification. *Molecular Cell* (2019) doi:10.1016/j.molcel.2019.08.002.

21. Laks, E. *et al.* Resource: Scalable whole genome sequencing of 40,000 single cells identifies stochastic aneuploidies, genome replication states and clonal repertoires. 411058 Preprint at https://doi.org/10.1101/411058 (2018).

22. Evrony, G. D., Hinch, A. G. & Luo, C. Applications of Single-Cell DNA Sequencing. *Annu Rev Genomics Hum Genet* **22**, 171–197 (2021).

23. Mission Bio. Tapestri: The Precision Genomics Platform. *Mission Bio Tapestri* https://missionbio.com/tapestri/ (2020).

24. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* **17**, 175–188 (2016).

25. Crasta, K. *et al.* DNA breaks and chromosome pulverization from errors in mitosis. *Nature* **482**, 53–58 (2012).

26. Reijns, M. A. M. *et al.* Signatures of TOP1 transcription-associated mutagenesis in cancer and germline. *Nature* **602**, 623–631 (2022).

27. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology* **12**, R18 (2011).

28. Jothilakshmi, S. & Gudivada, V. N. Chapter 10 - Large Scale Data Enabled Evolution of Spoken Language Research and Applications. in *Handbook of Statistics* (eds. Gudivada, V. N., Raghavan, V. V., Govindaraju, V. & Rao, C. R.) vol. 35 301–340 (Elsevier, 2016).
29. 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* **11**, (2020).
30. Ryland, G. L. *et al.* Loss of heterozygosity: what is it good for? *BMC Medical Genomics* **8**, 45 (2015).
31. Knudson, A. G. Mutation and Cancer: Statistical Study of Retinoblastoma. *Proceedings of the National Academy of Sciences* **68**, 820–823 (1971).
32. Shah, J. B. *et al.* Analysis of matched primary and recurrent BRCA1/2 mutation-associated tumors identifies recurrence-specific drivers. *Nat Commun* **13**, 6728 (2022).
33. Shalem, O., Sanjana, N. E. & Zhang, F. High-throughput functional genomics using CRISPR–Cas9. *Nat Rev Genet* **16**, 299–311 (2015).
34. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).
35. Nassar, L. R. *et al.* The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res* **51**, D1188–D1195 (2023).
36. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).
37. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
38. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).
39. Kuilman, T. *CopywriteR: Copy number information from targeted sequencing using off-target reads*. https://github.com/PeeperLab/CopywriteR (2022).
40. Chen, B. *et al.* Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155**, 1479–1491 (2013).
41. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281–2308 (2013).
42. Sack, L. M. *et al.* Profound Tissue Specificity in Proliferation Control Underlies Cancer Drivers and Aneuploidy Patterns. *Cell* **173**, 499-514.e23 (2018).

43. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. *Biostrings: Efficient manipulation of biological strings*. https://bioconductor.org/packages/Biostrings (2022).

44. Morgan, M., Pagès, H., Obenchain, V. & Hayden, N. *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. https://bioconductor.org/packages/Rsamtools (2022).

45. GATK. https://gatk.broadinstitute.org/hc/en-us.

46. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources* **17**, 44–53 (2017).

47. Delignette-Muller, M. L. & Dutang, C. fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software 64, 1–34 (2015).

48. McCarthy, D. J., Campbell, K. R., Lun, A. T. L. & Wills, Q. F. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* **33**, 1179–1186 (2017).

Figure 1

Custom Tapestri Panel Probes
Version 2 (CO610)
Version 1 (CO261)

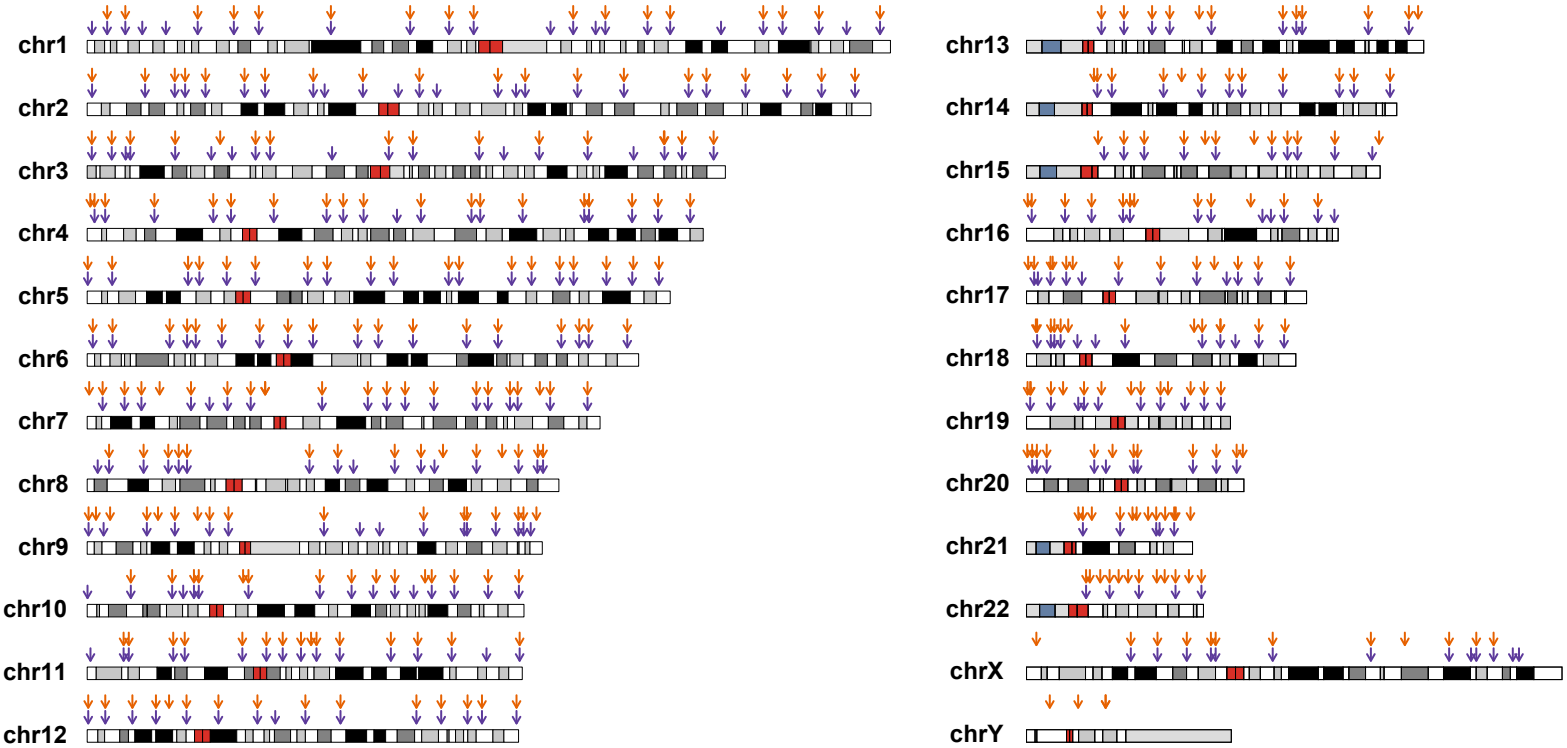


Figure 2

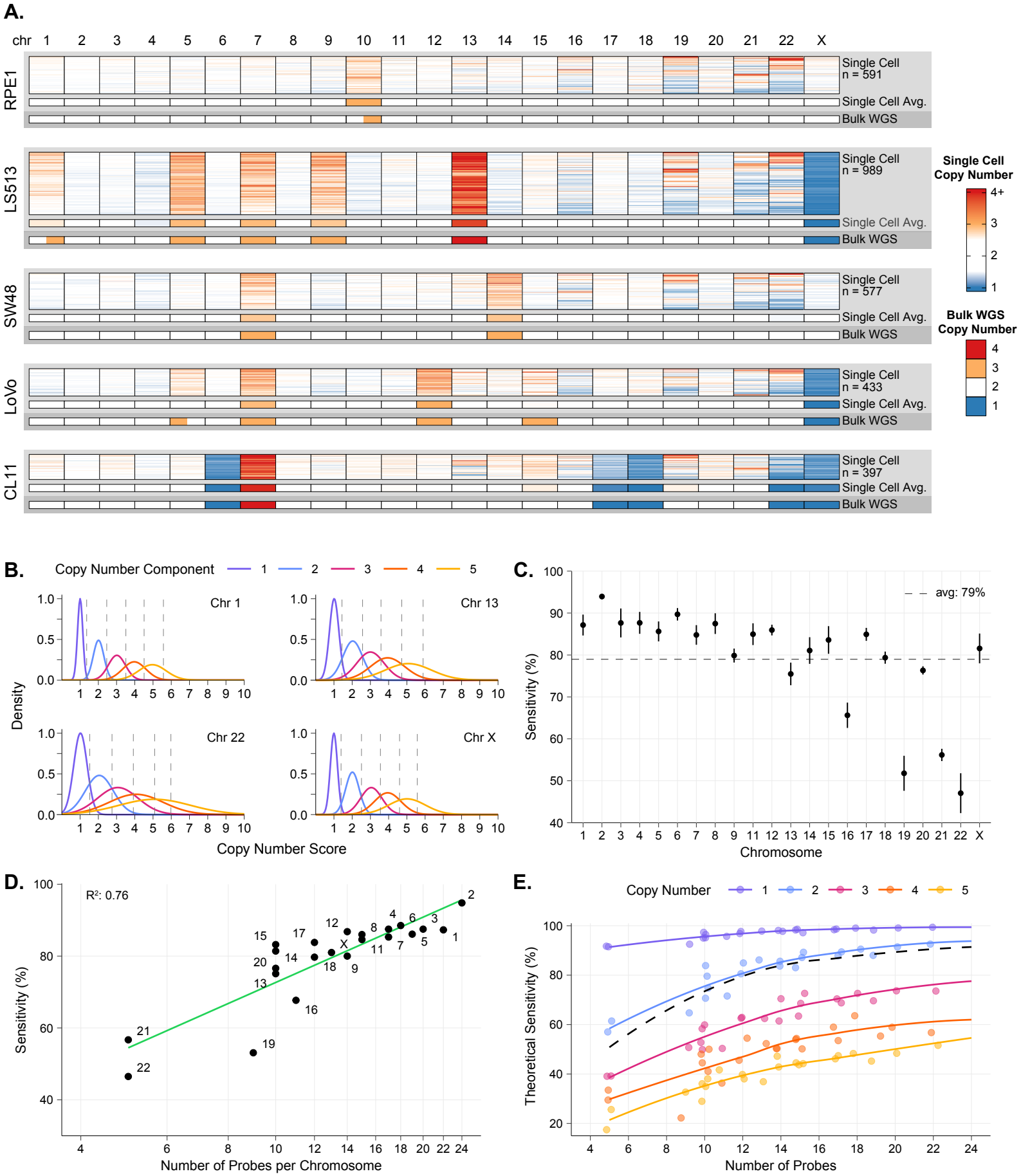

A.
chr 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X
RPE1
LS513
SW48
LoVo
CL11
Single Cell n = 591
Single Cell n = 989
Single Cell n = 577
Single Cell n = 433
Single Cell n = 397
Single Cell Avg.
Bulk WGS
Single Cell Copy Number
4+
3
2
1
Bulk WGS Copy Number
4
3
2
1
B.
Copy Number Component 1 2 3 4 5
Chr 1
Chr 13
Chr 22
Chr X
Density
Copy Number Score
C.
avg: 79%
Sensitivity (%)
Chromosome
D.
R²: 0.76
Sensitivity (%)
Number of Probes per Chromosome
E.
Copy Number 1 2 3 4 5
Theoretical Sensitivity (%)
Number of Probes

Figure 3

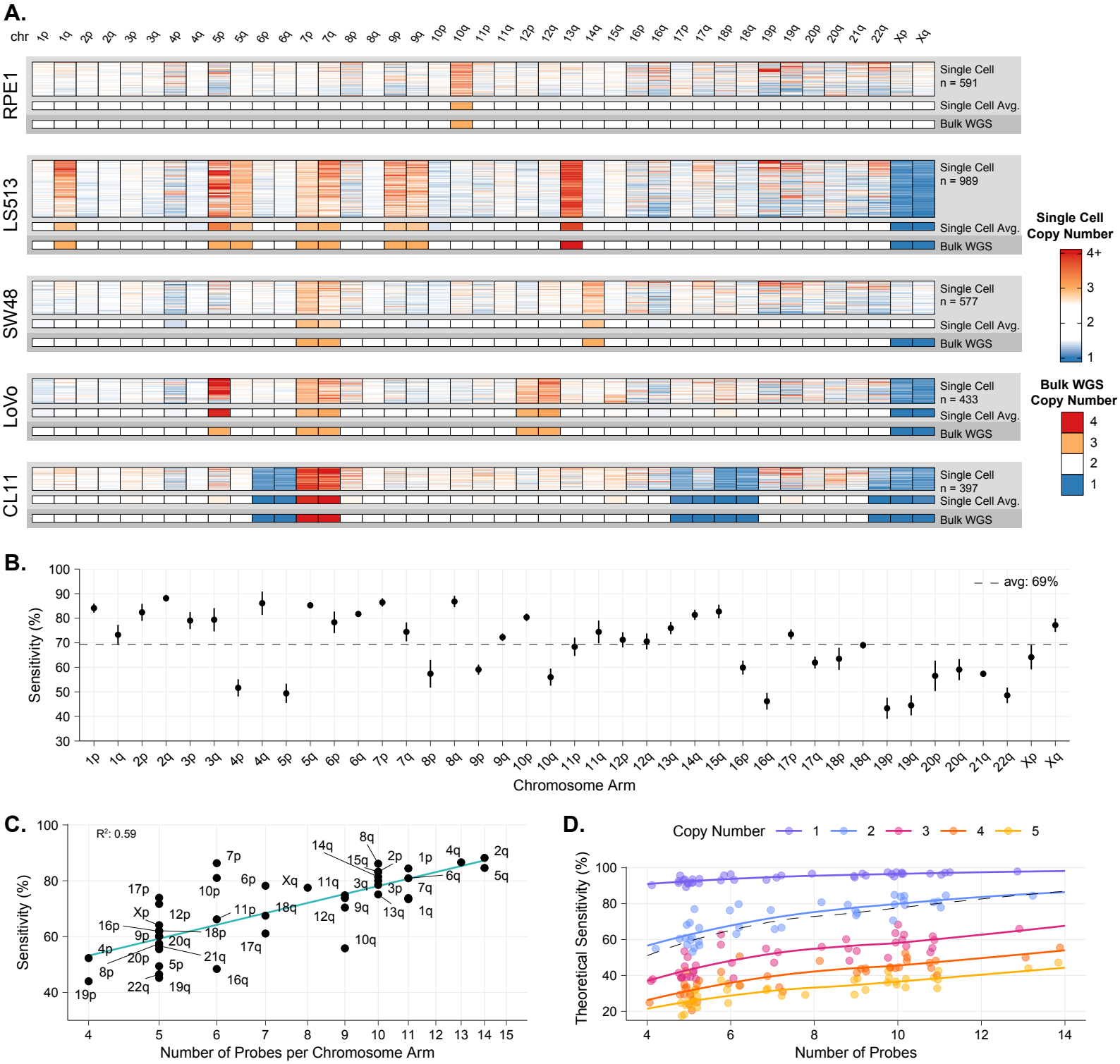

A.
chr
RPE1
LS513
SW48
LoVo
CL11
Single Cell
n = 591
n = 989
n = 577
n = 433
n = 397
Single Cell Avg.
Bulk WGS
Single Cell Copy Number
4+
3
2
1
Bulk WGS Copy Number
4
3
2
1
B.
avg: 69%
Sensitivity (%)
Chromosome Arm
C.
R²: 0.59
Sensitivity (%)
Number of Probes per Chromosome Arm
D.
Copy Number
1
2
3
4
5
Theoretical Sensitivity (%)
Number of Probes

Figure 4

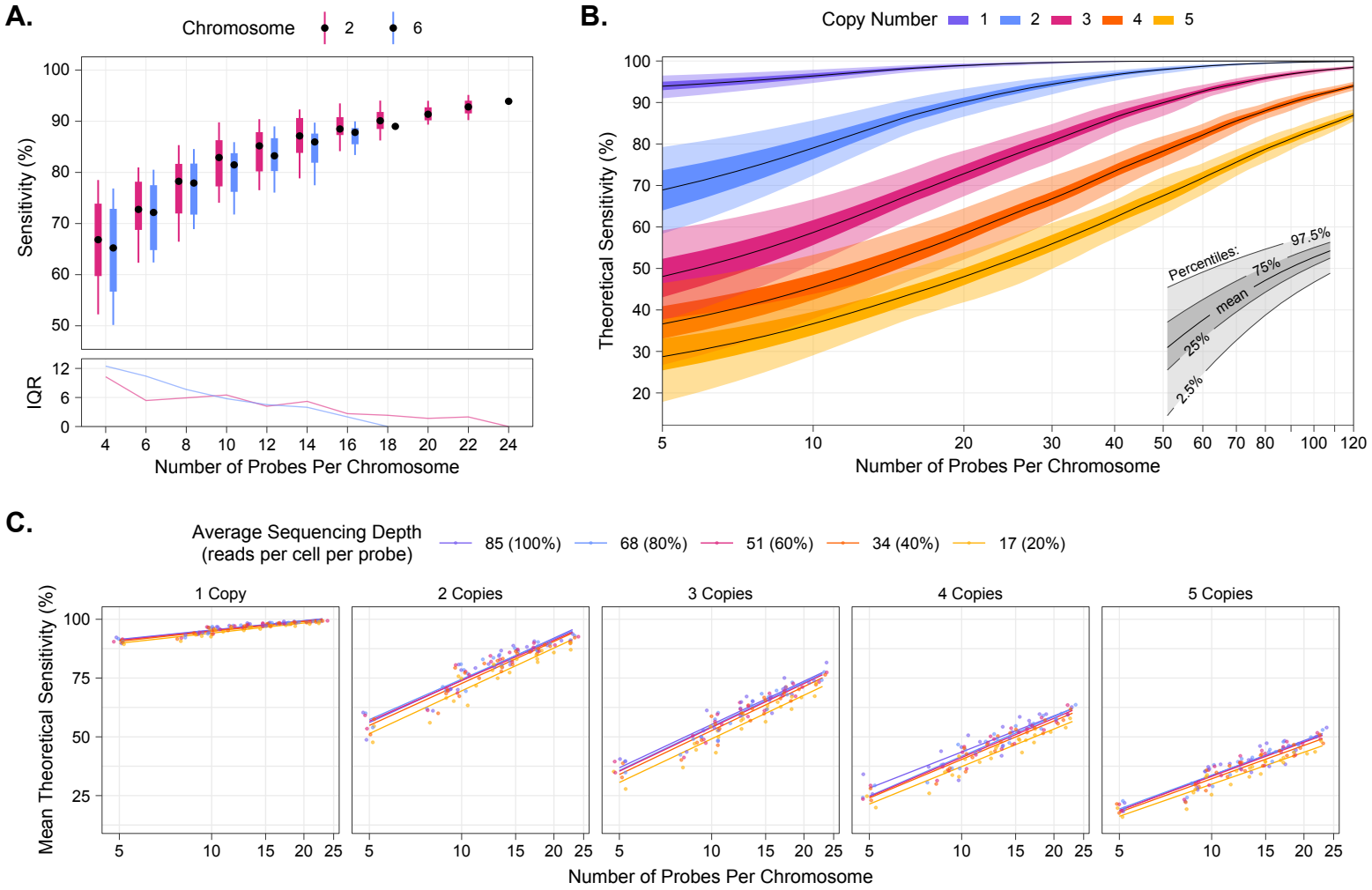

A.
Chromosome 2 6
Sensitivity (%)
IQR
Number of Probes Per Chromosome
B.
Copy Number 1 2 3 4 5
Theoretical Sensitivity (%)
Percentiles:
97.5%
75%
mean
25%
2.5%
Number of Probes Per Chromosome
C.
Average Sequencing Depth
(reads per cell per probe)
85 (100%)
68 (80%)
51 (60%)
34 (40%)
17 (20%)
1 Copy
2 Copies
3 Copies
4 Copies
5 Copies
Mean Theoretical Sensitivity (%)
Number of Probes Per Chromosome

Figure 5

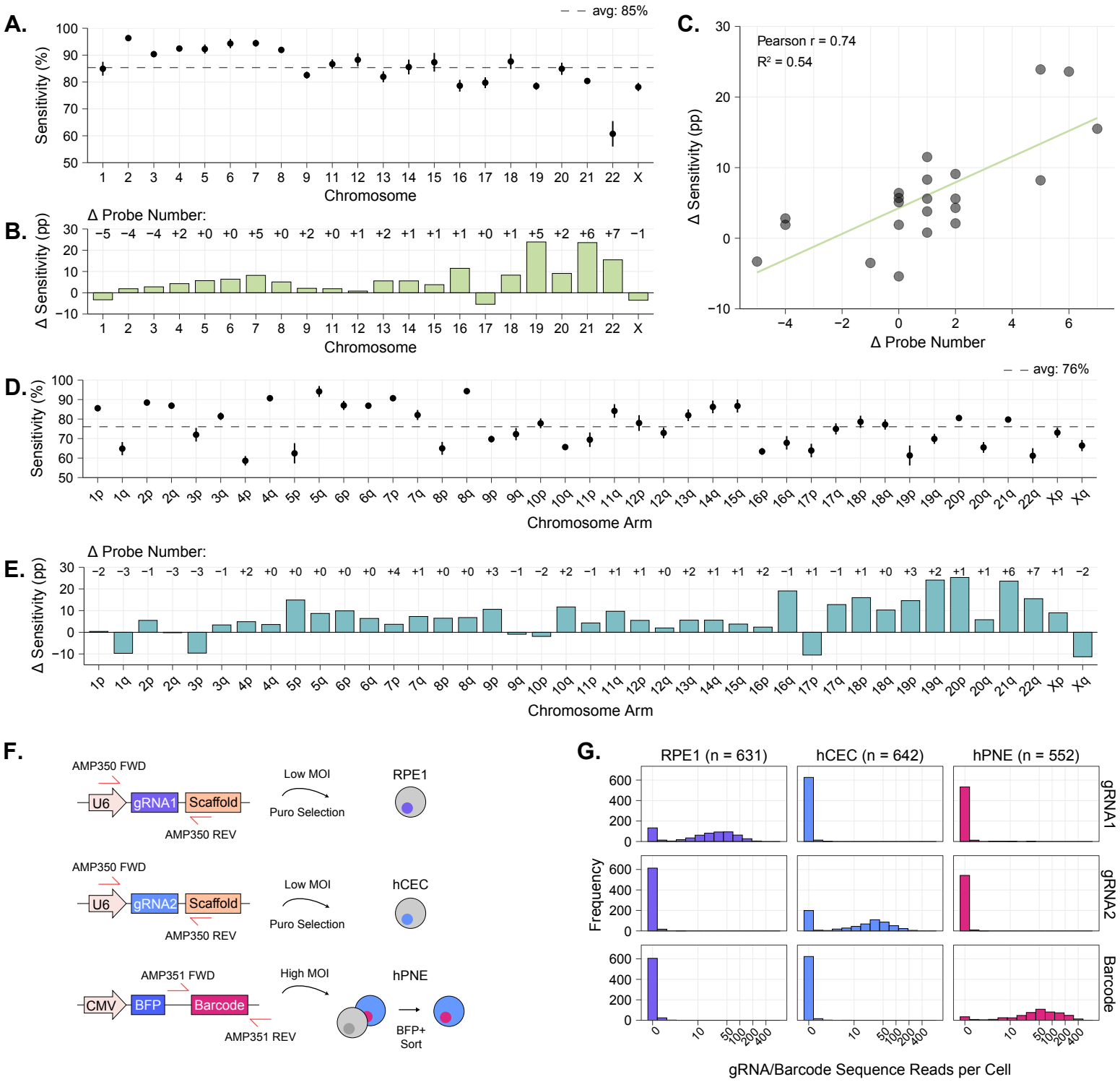
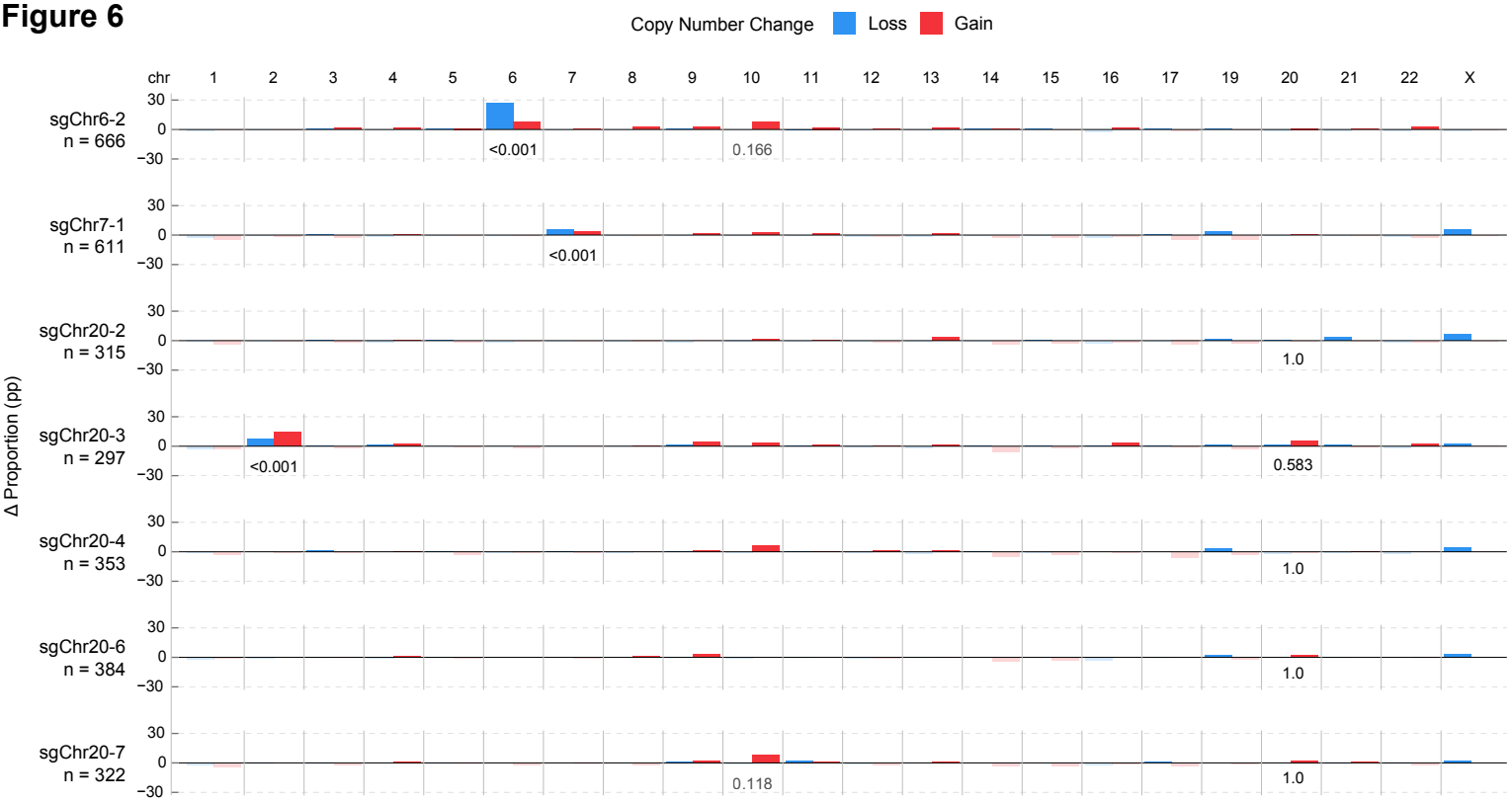
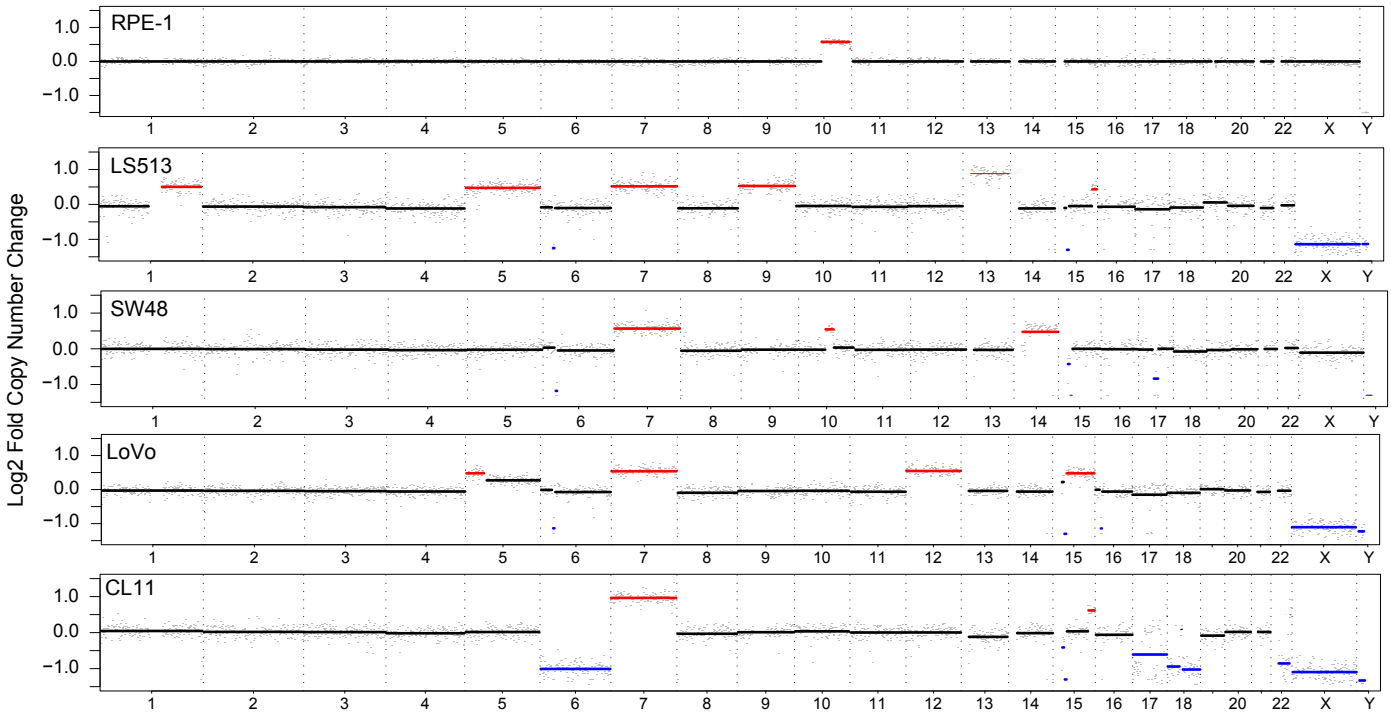

A.
avg: 85%
Sensitivity (%)
Chromosome
B.
Δ Probe Number:
−5 −4 −4 +2 +0 +0 +5 +0 +2 +0 +1 +2 +1 +1 +1 +0 +1 +5 +2 +6 +7 −1
Δ Sensitivity (pp)
Chromosome
C.
Pearson r = 0.74
R² = 0.54
Δ Sensitivity (pp)
Δ Probe Number
D.
avg: 76%
Sensitivity (%)
Chromosome Arm
E.
Δ Probe Number:
−2 −3 −1 −3 −3 −1 +2 +0 +0 +0 +0 +0 +4 +1 +0 +0 +3 −1 −2 +2 −1 +1 +1 +0 +2 +1 +1 +2 −1 +1 −1 +1 +0 +3 +2 +1 +1 +6 +7 +1 −2
Δ Sensitivity (pp)
Chromosome Arm
F.
AMP350 FWD
U6
gRNA1
Scaffold
AMP350 REV
Low MOI
Puro Selection
RPE1
AMP350 FWD
U6
gRNA2
Scaffold
AMP350 REV
Low MOI
Puro Selection
hCEC
AMP351 FWD
CMV
BFP
Barcode
AMP351 REV
High MOI
hPNE
BFP+
Sort
G.
RPE1 (n = 631)
hCEC (n = 642)
hPNE (n = 552)
gRNA1
gRNA2
Barcode
Frequency
gRNA/Barcode Sequence Reads per Cell

Figure 6


Copy Number Change
Loss
Gain
chr
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
19
20
21
22
X
Δ Proportion (pp)
sgChr6-2
n = 666
<0.001
0.166
sgChr7-1
n = 611
<0.001
sgChr20-2
n = 315
1.0
sgChr20-3
n = 297
<0.001
0.583
sgChr20-4
n = 353
1.0
sgChr20-6
n = 384
1.0
sgChr20-7
n = 322
0.118
1.0
30
0
−30

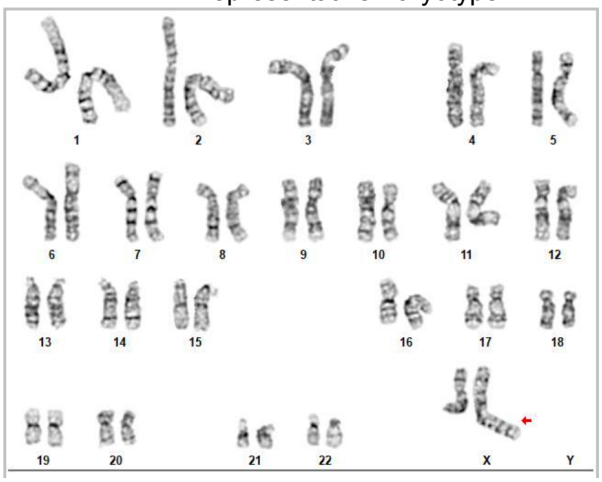
Supplemental Figure 1

A.



B.

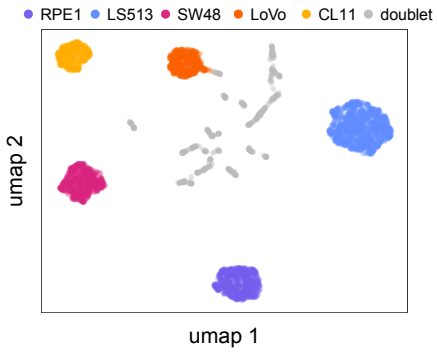
RPE-1 Representative Karyotype



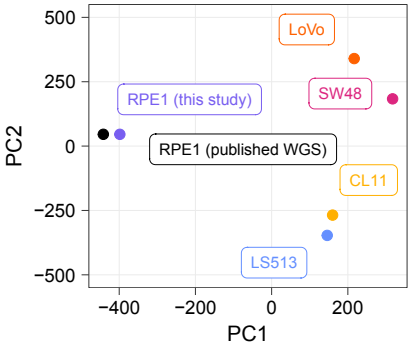
100/100: 3 copies 10q (translocated to X)

- 3/100: 3 copies 12
- 1/100: 3 copies 9, 3 copies 18
- 1/100: 3 copies 13

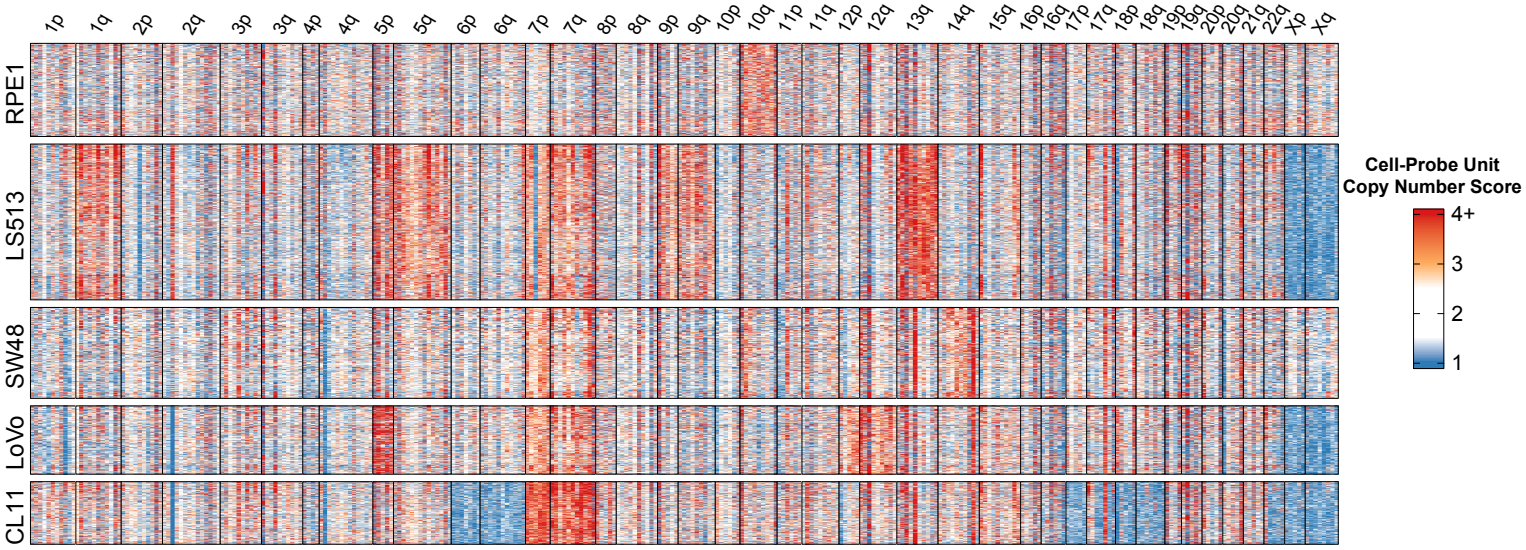
C.



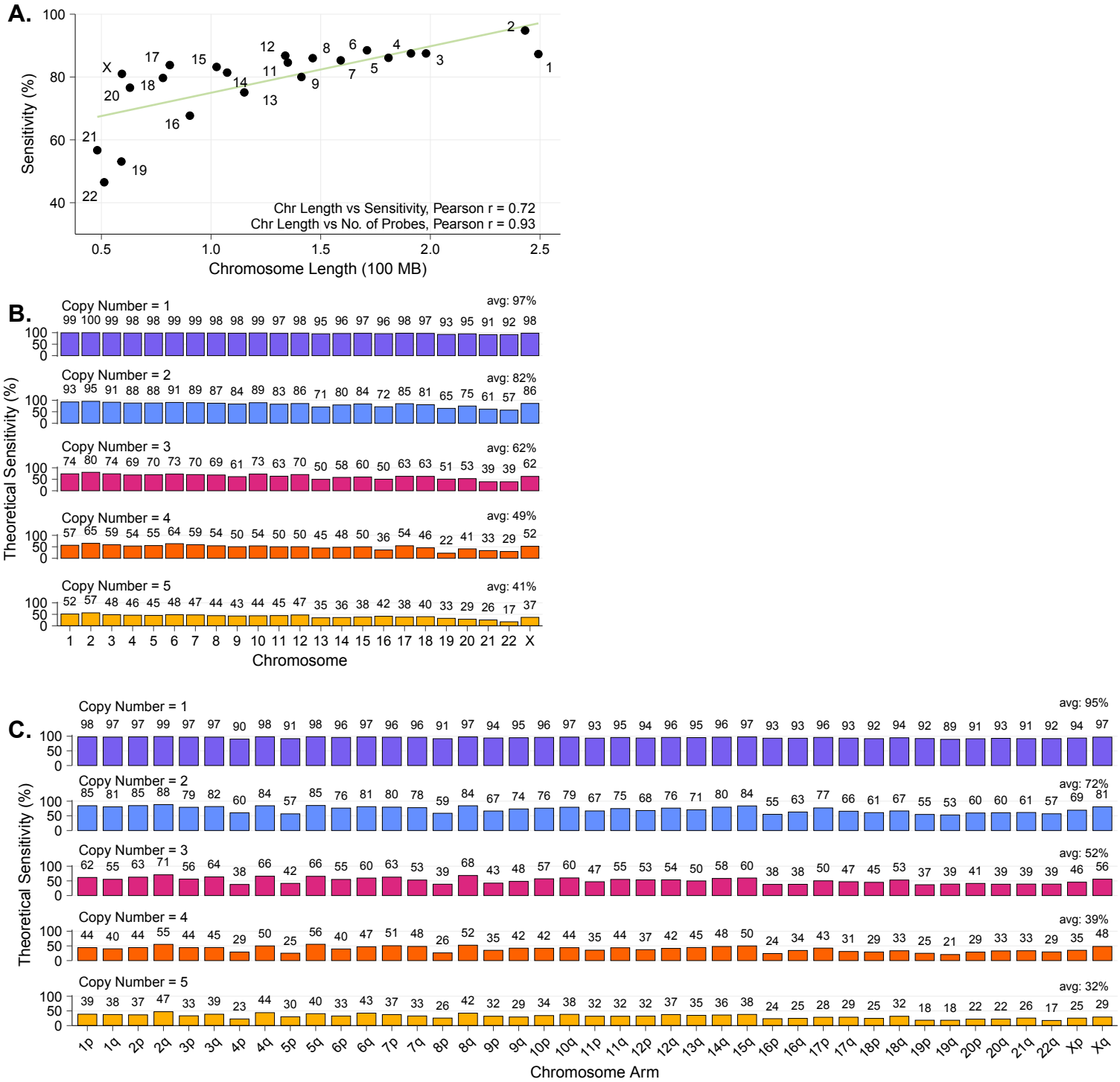
D.



E.

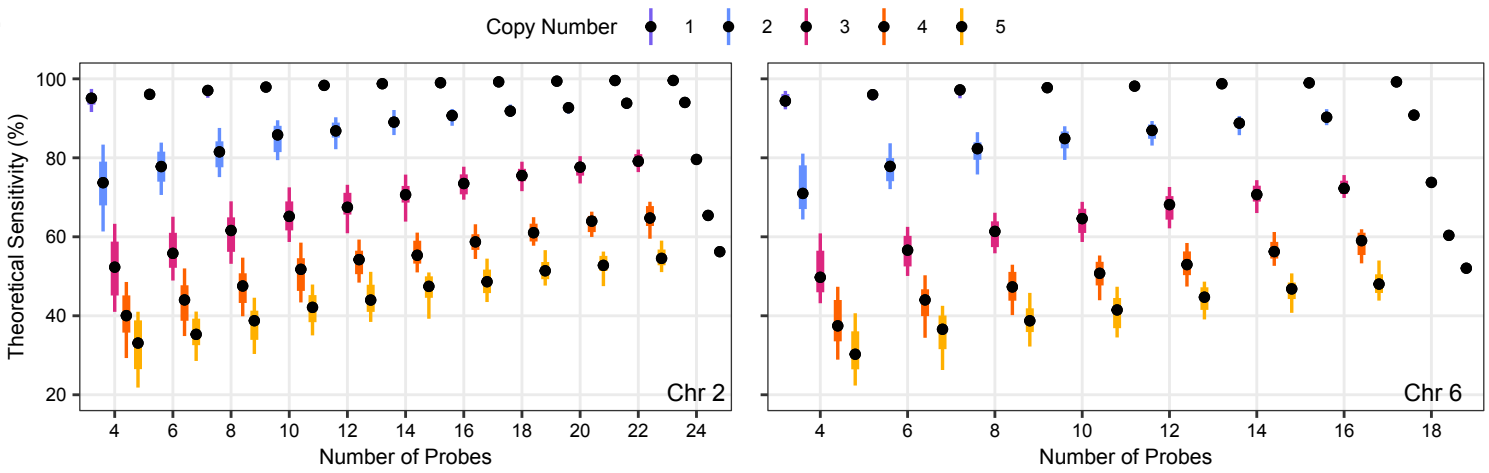


Supplemental Figure 2

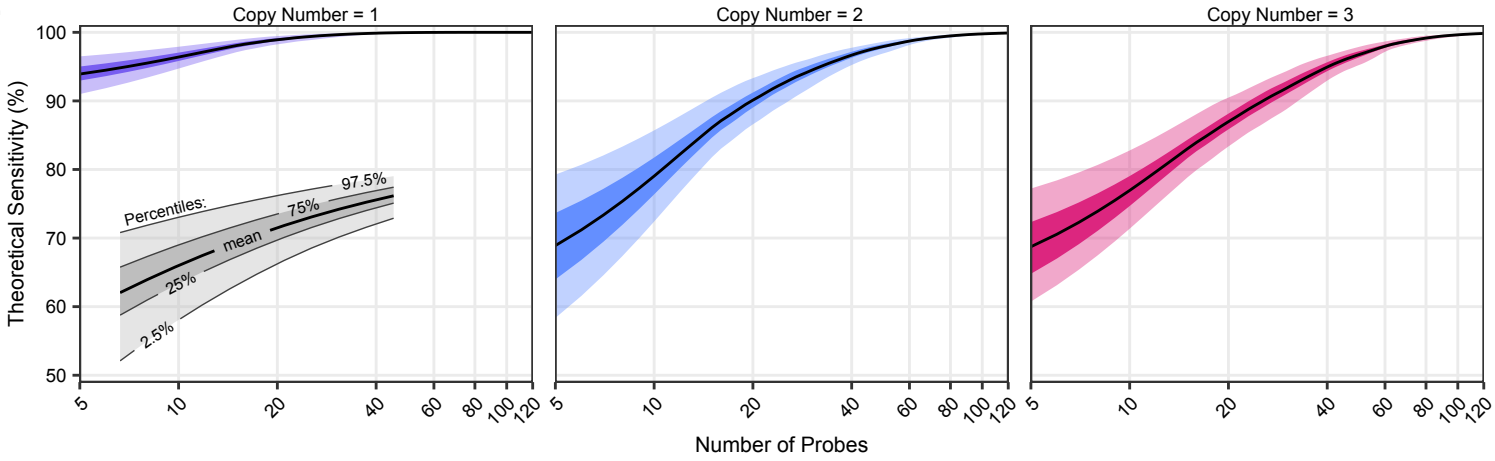


Supplemental Figure 3

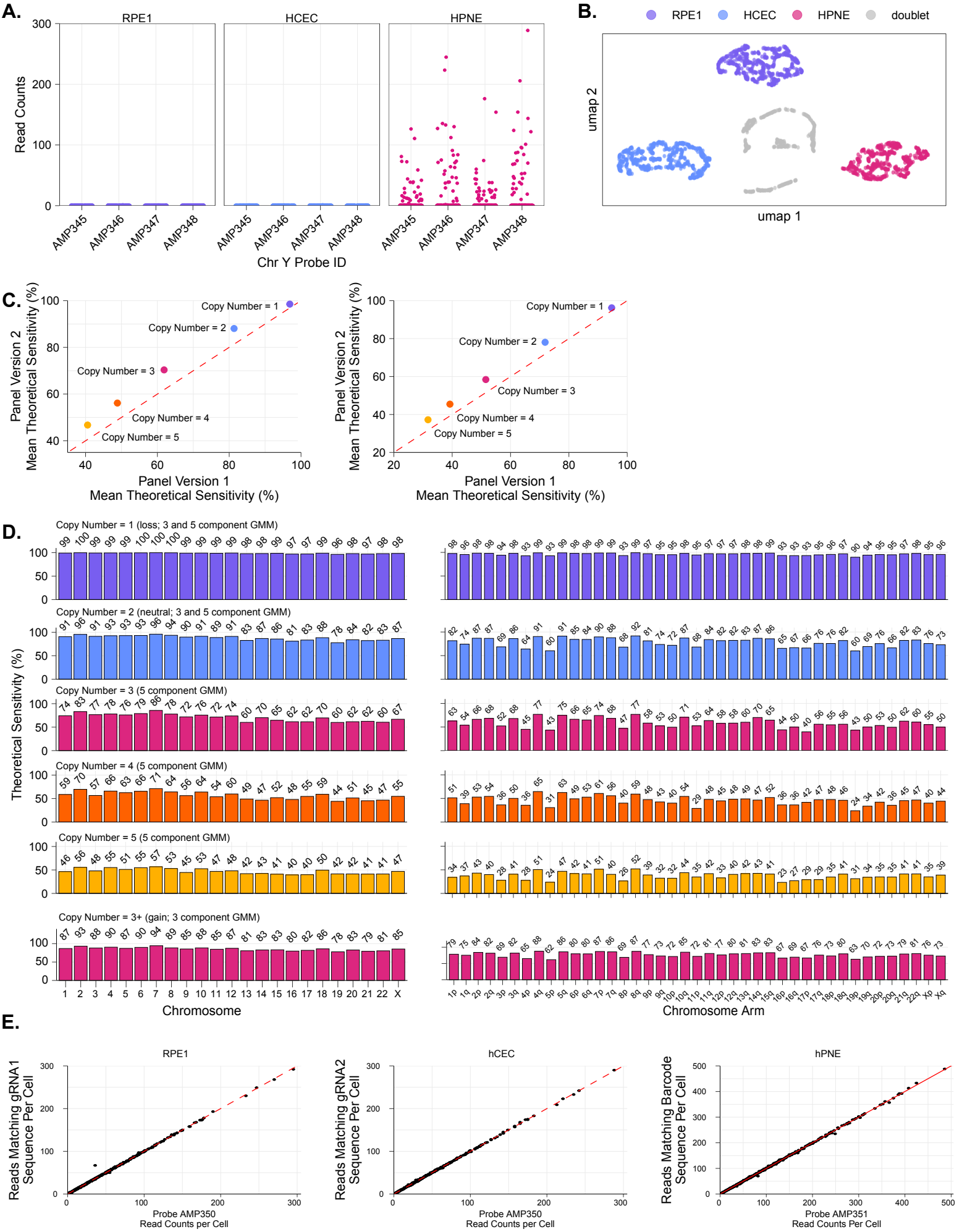
A.

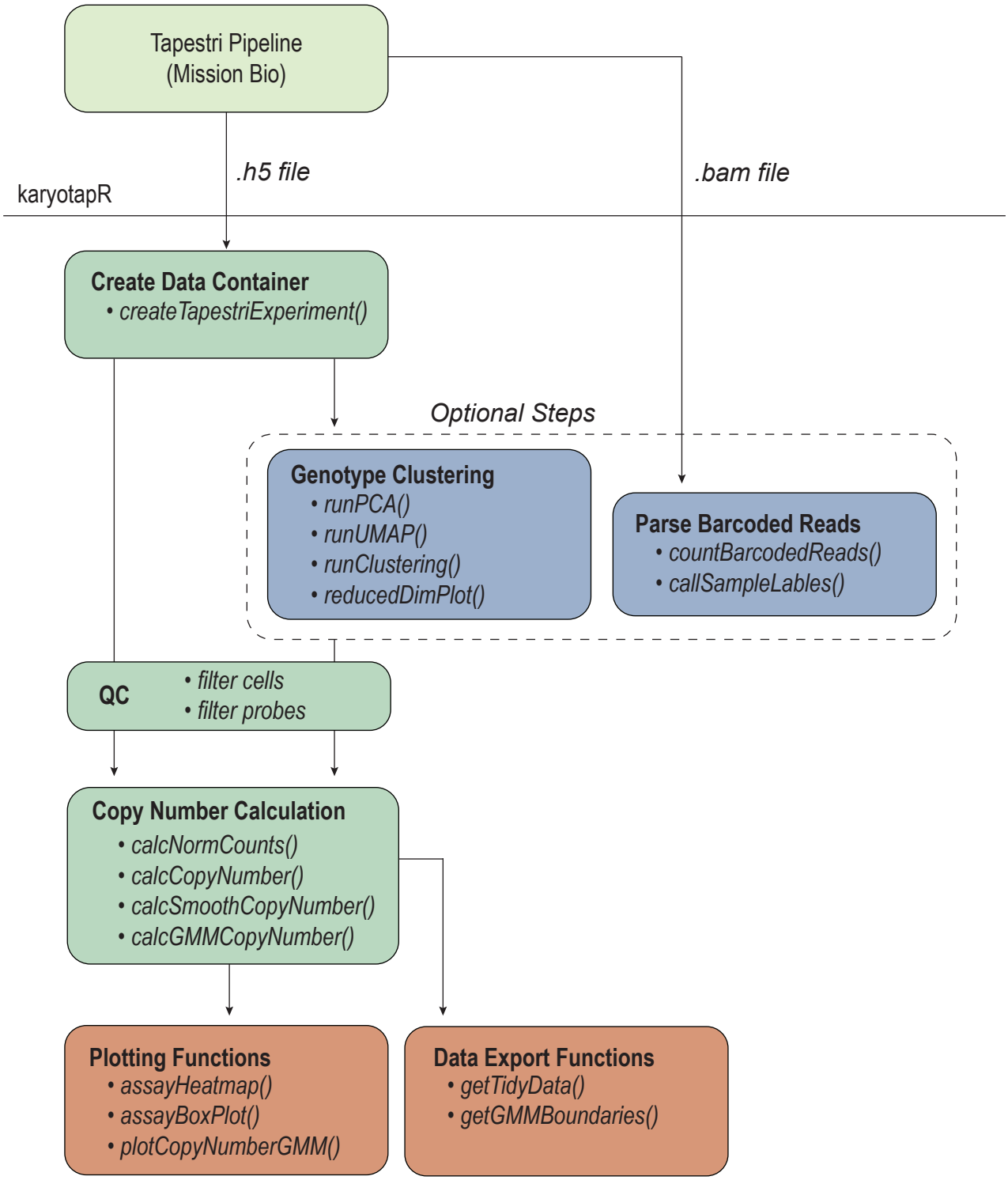


B.



Supplemental Figure 4



Supplemental Figure 5

karyotapR Workflow