Stony Brook University

High-Throughput Single-Cell Copy Number Profiling for Cancer Heterogeneity Analysis

A Dissertation Presented

By

Taimour Baslan

to

The Graduate School

In Partial Fulfillment of the

Requirements

for a Degree of

Doctor of Philosophy

in

Molecular and Cellular Biology

Stony Brook University

December 2014

Stony Brook University

The Graduate School

Taimour Baslan

We, the dissertation committee for the above candidate for the

Doctor of Philosophy degree, hereby recommend acceptance of this dissertation

James B. Hicks, Ph.D., Dissertation Advisor
Research Professor, Cancer Genomics
Cold Spring Harbor Laboratory

Kenneth R. Shroyer, M.D, Ph.D., Chairperson of Defense
The Marvin Kuschner Professor and Chair of Pathology
Stony Brook University

Jinfang Ju, Ph.D.
Associate Professor, Department of Pathology
Stony Brook University

Alexander Krasnitz, Ph.D.
Assistant Professor, Quantitative Biology
Cold Spring Harbor Laboratory

Kenneth Offit, M.P.H., M.D.
Chief, Clinical Genetics Service
Memorial Sloan Kettering Cancer Center

This dissertation is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Dissertation

High-Throughput Single-Cell Copy Number Profiling for Cancer Heterogeneity Analysis

by

Taimour Baslan

Doctor of Philosophy

in

Molecular & Cellular Biology

Stony Brook University

2014

Intra-tumoral genetic heterogeneity has long been recognized, yet remains poorly understood. This has primarily been due to the lack of sensitive technologies to measure it. Genome wide analysis at the level of single cells has recently emerged as a powerful tool to dissect cancer genome heterogeneity. However, to be truly transformative, single cell approaches must accommodate the analysis of large numbers of single cells. Here, using integrative informatics and molecular biology approaches this study presents a robust, low-cost, and high-throughput method to retrieve the genome-wide copy number landscape of hundreds of single cancer cells. Application of the method to human cancer cell lines and clinical cancer tissue illustrates the underlying genetic heterogeneity present in both and further reveals mosaicism of chromosomal amplifications in clinical cancer samples. The capacity of the method to facilitate the rapid profiling of hundreds and thousands of single cell genomes is bound to illuminate the biology of intra-tumoral heterogeneity.

I dedicate this thesis to my family for their unconditional love

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

CNV - Copy Number Variation
aCGH – array Comparative Genomic Hybridization
SNP – Single Nucleotide Polymorphism
ERBB2 – v-erb-b2 avian erthroblastic leukemia viral oncogene homolog 2
NGS – Next Generation Sequencing
WGA – Whole Genome Amplification
SNV – Single Nucleotide Variants
SV – Structural Variants
DOP-PCR – Degenerate Oligo-nucleotide Priming – Polymerase Chain Reaction
BSA – Bovine Serum Albumin
RNA-Seq – RNA Sequencing
Varbin – Variable Bin
CBS – Circular Binary Segmentation
CORE – Cores of Recurrent Events
C-DOP-L – Cleavable Degenerate Oligo-nucleotide Priming Ligation
MYC – v-myc avian myelocytomastosis viral oncogene neuroblastoma derived homolog
DCC – DCC netrin 1 receptor
ER – Estrogen Receptor
CCND1 – Cyclin D1
TOP2A – Topoisomerase (DNA) II alpha 170 KDa
SIX6 – SIX homeobox 6
PREX1 - Phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 1
EGFR – Epidermal Growth Factor Receptor
AR – Androgen Receptor
FA – Fanconi Anemia
DEB – Diepoxybutane
CTCs – Circulating Tumor Cells
SAGE – Serial Analysis of Gene Expression
PacBio – Pacific Biosciences
FFPE – Formalin Fixed Paraffin Embedded
OCT – Optimal Cutting Temperature Compound
DAPI – 4',6-diamidino-2-phenylindole
FACS – Fluorescence Activated Cell Sorting
ACDU – Automated Cell Deposition Unit
RSEM – RNA-SEQ by Expectation Maximization
TCGA – The Cancer Genome Atlas
HPLC – High Performance Liquid Chromatography
BWA – Burrows-Wheeler Aligner
EDTA – Ehylenediaminetetraacetic acid

MET – MET proto-oncogene, receptor tyrosine kinase
PC – Principle Components
PCA – Principle Component Analysis
FBS – Fetal Bovine Serum
RE – Restriction Enzyme
LumA – Luminal A
LumB – Luminal B
FNA – Fine Needle Aspirate
NSCLC – Non Small Cell Lung Cancer

ACKNOWLEDGEMENTS

I would like to extend my thanks to a long list of people without whom the work presented here would not be possible. I am very grateful for my advisor, James Hicks, for giving me endless portions of freedom, supplanted with guidance and direction. I have learned much about science as well as life in general from him. I would like to thank my thesis committee, Dr. Jingfang Ju, Dr. Alex Krasnitz, Dr. Kenneth Offit, and Dr. Kenneth Shroyer for guiding my voyage and believing in my work. I would like to thank both graduate student directors, Dr. Rolf Sternglanz and Dr. Wali Karzai, for their guidance as well. I would also like to extend my thanks to Dr. Michael Wigler for his support and guidance.

I would like to thank all my laboratory colleagues at Cold Spring Harbor Laboratory who assisted me greatly in my work, specifically, Jude Kendall, Linda Rodgers, and Hilary Cox. None of this would be possible without the contribution of every single one of them. I would like to further thank the community, both scientific and administrative, at Cold Spring Harbor Laboratory and Stony Brook University. It takes a close community to raise a scientist, and both places are exactly that. I would also like to thank my collaborators from the academia and well as the industry. Our work together has taught me a great deal.

I wish to thank my close friends at Cold Spring Harbor and Stony Brook who supported me in my journey in many ways. Specifically, Assaf Gordon whom I have learned from a tremendous amount.

Finally, I would like to thank my family. My father, for supporting my decisions and believing in me. My brother and sister for unconditional love. My aunt, for being my best friend and confidant. My wife for fueling my passion for science and life. And lastly, my mum for giving me wings and teaching me the power of dreams.

1. INTRODUCTION

1.1. COPY NUMBER VARIATION IN ORGANISMAL GENETICS AND BIOLOGY

Copy number variation (CNV) is an important source of genetic variation in organismal biology and is known to influences phenotypic traits[1]. From simple eukaryotic organisms such as yeast to more complex ones such as cattle, copy number variants have been linked to variable phenotypes. For example, in yeast, copy number gains of the FLO11D locus have been shown to confer adaptation to environmental conditions such as osmostress[2]. In maize, genome wide surveys of cohort lines has demonstrated extensive copy number variation in inbred populations[3] with more detailed analysis linking copy number variants to traits such as aluminum tolerance[4]. In cattle, copy number variants have been associated with breeding specific traits involving health and reproduction[5]. Importantly in humans, ever since the initial description of large-scale variation in copy number polymorphisms[6,7], human genome copy number investigations have proliferated at an fast rate. Many studies have found associations between copy number variants and normal phenotypes such as human olfactory receptors and smell[8] and the amylase gene and diet[9]. In addition, copy number variation has also been linked with a wide range of deleterious phenotypes and disorders, ranging from the congenital to the developmental. Copy number variants have been associated with congenital heart disease[10-12] as well as congenital kidney disease[13,14]. In developmental disorders such as

obesity and autism, where initial copy number variation studies provided compelling evidence for association[15,16], large scale investigations have resulted in catalogues of associated mutations illuminating the underlying genetics of these disorders[17-20]. Thus, copy number variation plays an important part in the underlying functional genetics of genomes and contributes significantly to phenotypic expressivity.

1.2. CANCER GENETICS AND COPY NUMBER VARIATION

Cancer is a genetic disease of an evolutionary nature where the interplay between germline and somatic mutations selects for an unrestrained proliferative phenotype[21]. Of the myriad of genetic alterations cancer genomes carry, copy number variants in the form of chromosomal deletions and duplications/amplifications, both in somatic and germline contexts, occupy a central role. Initially observed and studied using chromosome banding techniques and cytogenetics at the gross scale and later thru microsatellite analysis, copy number alterations have been found to be non-random and recurrent across many different tumor types[22-26]. Subsequently, technological improvements gave way for Array Comparative Genomic Hybridization (aCGH)[27,28] and Single Nucleotide Polymorphism (SNP) arrays (SNP-arrays)[29,30] (based on differential labeling of tumor DNA samples with fluorophores, hybridization to arrays containing oligonucleotide probes, and analysis of fluorometric signal ratios) which allowed for the analysis of copy number variation at a genome-wide level and at higher resolution. The availability of these technologies and the realization of the commonality of copy number alterations in tumor genomes led to intense investigations of genome-wide copy number profiles across many cancer types

and in some cases thousands of tumor samples[31-35]. These analysis have (1) linked copy number variants to genetic predisposition to cancer development across many tumor types[36-38], (2) led to the identification of cancer driver oncogenes and tumor suppressor genes[39-42], (3) guided therapeutic decisions (for example Herceptin in ERBB2 amplified breast cancers)[43,44], (4) helped predict drug sensitivity of tumors[45,46], and (5) assisted in the prognostication of cancer patients[47-51]. Together, these studies offered strong evidence for the importance of copy number variation in cancer biology and the need to further the understanding of its occurrence.

1.3. INTRA-TUMORAL GENETIC HETEROGENEITY IN CANCER AND NEXT GENERATION SEQUENCING ANALYSIS

Implicit in the description of cancer as a genetic disease subject to the principles of natural selection is the argument that along the tumor's evolutionary trajectory, different genetically distinct sub-populations are likely to evolve and dynamically interact with each other[52-55]. Indeed, the occurrence of intra-tumoral genome heterogeneity has long been hypothesized[56] and researchers did embark upon its characterization[57-60]. However, most studies were limited in scope, primarily due to the inadequacies of the existing technologies. This, however, has changed with advent of Next Generation Sequencing (NGS) technologies[61,62]. The highly quantitative and qualitative nature of the sequencing data, along with the ever increasing output of next generation sequencing machines have led to the adoption of sequencing technologies in all facets of genomic research, from the identification of single nucleotide polymorphisms[63,64] to the

delineation of copy number variants[65,66], at a genome-wide scale. Importantly, the depth of sequencing data that has be generated in some studies has also facilitated the identification of sub-clonal variants[67,68], and consequently re-kindled the cancer community's interest in intra-tumoral heterogeneity. Many reports based on deep sequencing of whole genomes or targeted regions such as protein coding genes (i.e. exome) have provided quantitative descriptions of genetic intra-tumoral heterogeneity[69-71] and in many cases have linked it to disease progression[72,73], metastasis[74,75], as well as therapeutic resistance to targeted therapies[76-78]. Nonetheless, our knowledge of cancer genome heterogeneity is still lacking and therefore, new technologies are urgently needed to facilitate the dissection of intra-tumoral heterogeneity.

1.4. SINGLE CELL SEQUENCING APPROACHES FOR THE STUDY OF INTRA-TUMORAL GENETIC HETEROGENEITY

Recently, by coupling the power of Next Generation Sequencing (NGS) technologies to Whole Genome Amplification (WGA) approaches, single cell genomic analysis have emerged as a powerful approach to dissect cancer genetic heterogeneity[79,80]. Single cell sequencing approaches have been developed to query, at a genome-wide level, single nucleotide variants (SNVs)[81,82], structural variants (SVs)[83], epigenetics states[84] as well as copy number variation[79,80]. Investigation utilizing single cell sequencing methods have begun to illuminate valuable and novel aspects of cancer biology and promise to deliver more[85-87]. However, to realize the potential of single cell sequencing in understanding the biology of heterogeneity, methods are needed that allow

the investigation of hundreds of single cell genomes at reasonable cost in time, effort and reagents. Sequencing hundreds of single cells to the nucleotide level is simply not affordable even with the remarkable NGS platforms that are available. Fortunately, copy number analysis requires only sparse sequence coverage, yet it can distinguish subpopulations and provides deep insights into genetic heterogeneity. Thus, in theory, coupling sparse sequencing with molecular barcoding approaches offers a mean to profile many cells together.

1.5. CHALLENGES IN HIGH LEVEL MULTIPLEXING OF SINGLE CELLS

The feasibility of multiplex single cell analysis has been demonstrated by combining up to eight barcoded single cells on a single sequencing lane[88,89]. However the potential for higher level multiplexing has not been explored at either the bioinformatic or operational levels. To accomplish this, informatic analysis aimed at identifying minimal sequence read requirements for robust copy number identification are required. Furthermore, while technically feasible, amplifying and creating barcoded sequencing libraries from many single cells using traditional library preparation protocols involving sonication, end repair, A-tailing, and adaptor ligation is time consuming and expensive. Hence, an optimized multiplexing process that informs the minimum number of reads that can be used to determine genome-wide copy number profiles at specific levels of resolution and a simplified preparative method that is faster and cheaper and yet maximizes the amount of information that can be extracted from each sequencing read

from a single sequencing lane of the Illumina HiSeq machine would represent a key step in developing technologies suited to address intra-tumoral genetic heterogeneity.

Here, a robust and affordable, high-throughput method is described that employs a modified version of Degenerate Oligo-nucleotide Priming-PCR (DOP-PCR) amplification, simplified library preparation, and multiplex sequencing that facilitates the retrieval of the genome-wide copy number landscape of hundreds of individual cancer cells. The method drastically lowers the cost of profiling single cell genomes (down to ~ $30 per single cell), significantly cuts sequence library preparation time, and maximizes the amount of information extracted from each single cell sequencing data set. The approach is applied to human cancer cell lines and clinical cancer biopsies to demonstrate its power to reveal population heterogeneity.

2. MATERIALS AND METHODS

2.1. Cell culture and clinical samples

315A lymphoblastoid cells were cultured as suspension cultures in RPMI 1640 (Gibco-Invitrogen) supplemented with 10%FBS (HyClone), 100 U ml^{-1} penicillin and 100ug ml^{-1} streptomycin (Gibco-Invitrogen). SK-BR-3 and MDA-MB-231 were cultured as adherent cultures in DMEM (Gibco-Invitrogen) supplemented with 10% FBS, 100 U ml^{-1} penicillin and 100ug ml^{-1} streptomycin. All lines were cultured at 37°C and 5%CO_2. Core biopsies, obtained prior to treatment, were processed by formalin fixation and paraffin embedding (FFPE) or frozen down and stored in OCT compound (2 cores each per biopsy event). Both specimen types were subjected to sectioning, hematoxylin and eosin staining, and histologic evaluation by the study pathologist. Frozen cores were processed for single nuclei isolation as described below. FPPE sections were used for tumor histology and immunohistochemistry.

2.2. NST-DAPI nuclei isolation buffer

NST buffer was prepared by mixing the following components in ddH_2O for a final volume of 800 ml; 146 nM NaCl, 10mM Tris Base pH7.8, 1mM $CaCl_2$, 21mM $MgCl_2$, 0.05% BSA, and 0.2% NP40. NST-DAPI buffer was prepared by adding 200 mL of $MgCl_2$ at a concentration of 106mM to the 800 ml of NST buffer followed by dissolving 10mg of DAPI and storing at 4 °C protected from light.

2.3. Nuclei isolation, DNA staining and single cell flow cytometry

For cell lines (both adherent and suspension) nuclei were prepared by collecting suspension cells (following trypsinization in the case of adherent cells) in PBS to a 15 ml conical centrifuge tube and gently centrifuging at 105 xg for 4 min followed by medium aspiration. Cells were subsequently re-suspended in 5 ml of PBS and counted using a hemacytometer. 0.5-1.0 x 10^6 cells were centrifuged again at 105xg for 4 mins. Following centrifugation, media was aspirated without disturbing cellular pellets. Cells were dispersed by gentling flicking the 15 ml canonical tube several times. This was followed by addition of 1 ml of NST-DAPI buffer and holding on wet ice. For frozen core biopsies, nuclei were prepared by finely mincing tissue in a 60mm TC plate with 0.5ml NST-DAPI buffer using two fine-point disposable scalpels until pieces are very fine. Prior to sorting, NST-DAPI suspended nuclei (from both cell lines and human tissue) were run thru a 5 ml Falcon round-bottom tube with cell-strainer cap to select against cellular debris and clumps that might clog the flow sorting machine. Single-cell sorting was performed using a FACS AriaIIU SORP (BD Biosciences, San Jose, CA) with the ACDU option (Automated Cell Deposition Unit). The sorter was run inside a BioProtect IV Safety Cabinet (Baker Company, Sanford, ME) to maintain BSL2 biosafety standards. The DAPI signal was detected by a 355 nM UV Laser (450/50 bandpass filter). Gains were set for the UV photomultiplier based on the DNA content equivalent to human diploid lymphoblast cells (315A cells). Single nuclei were determined by doublet discrimination using dot plots with DAPI area of the *y*-axis and DAPI pulse height on the *x*-axis as described by Wersto *et.al*[90]. From the single cell gate,

a histogram was derived that plots DNA content on a linear scale on the *x*-axis. Single nuclei were sorted according to DNA content. Single cells were deposited in 96 well plate format containing 9 µl of cell lysis buffer (800 µl H_2O, 6 µl Proteinase K, and 96 µl 10X Singe Cell Lysis and Fragmentation Buffer, SIGMA WGA4).

2.4. Single Cell Whole genome amplification

Single cells were lysed by incubating 96 well plates for 1 hour at 50 °C followed by 4 min at 99 °C using a thermocycler. Single cell whole genome amplification was then carried out using the SeqPlex Enhanced DNA Amplification Kit (SEQXE, SIGMA) as described below. Following single cell lysis and DNA fragmentation, DNA was denatured and primed for initial library synthesis by adding 2 ul of Library Preparation Buffer to each well and incubating in a thermocycler at 95 °C for 2 minutes followed by cooling at 4 °C. Library pre-amplification synthesis was performed by adding 1ul of Library Preparation Enzyme and incubating in a thermocycler in a temperature ramping scheme as follows: 16 °C for 20 mins, 24 °C for 20 mins , 37 °C for 20 mins, and 75 °C for 5 mins. Reactions were subsequently cooled on ice. Pre-amplification library DNA molecules containing universal sequences on the ends of the molecules are then amplified using single primer PCR by adding 15 µl 5X Amplification Mix, 1.5 µl DNA Polymerase for SeqPlex, and 42.5 µl of Water, incubating in a thermocycler, and using 24 cycles for amplification according to the following parameters: Initial denaturation: 94 °C for 2 minutes, 94 °C for 15 mins for subsequent denaturation steps and 70 °C for 5 mins to anneal/extend. After cycling, DNA molecules are incubated for 30 mins at 70 °C to

ensure filling of the DNA ends to facilitate subsequent reaction steps (i.e. restriction digestion of universal WGA sequences). Single cell amplification products were purified using QIAquick 96 well plates according to manufacturers instructions and DNA eluted in 50 µl EB solution.

2.5. Illumina library generation of single cell amplified DNA

All subsequent reactions were carried out in 96-well plate format using multi-channel pipetting. Restriction digestion of WGA universal sequences was performed interchangeably using SeqPlex supplied Primer Removal reagents (SIGMA) and Eco57I (Thermo Scientific). 1 µg of WGA DNA products in total volume of 20 µl containing 2.4 µl 10X Primer Removal Buffer/Buffer G, 0.4 µl Primer Removal Solution/SAM, and 0.5 Primer Removal Enzyme/Eco57I enzyme (Thermo Scientific). Reactions were incubated at 37 °C for 30 min followed by incubation at 65 °C for 15 min for enzyme deactivation. Reactions were subsequently cooled on ice. Following restriction digestion, 24 µl of EB and 26 µl of 2X Quick Ligase Reaction Buffer (NEB) were added to each reaction to bring the volume up to 70 µl. The addition of 26 µl of 2X Quick Ligase Reaction Buffer is critical since it facilitates selection of higher molecular weight DNA (between 200-600bps). Digested DNA was subsequently purified using Agencourt AMPure XP beads (Beckman Coulter) according to the following protocol. 30 µl of warmed beads were added to each digestion reaction. Beads and reaction products were mixed by vortexing for 7s. Mixed reactions were then incubated off-magnet for 10 min at RT after which they were transferred to DynaMag^{TM}-96 Side magnet (Life Technologies) and left to stand for

5 min. 90 µl of supernatant was withdrawn and discarded. Beads were washed with 180 µl of freshly made 80% EtOH. After a second round of EtOH washing, beads were allowed to dry on the magnet for 15 min. Dried beads were then re-suspended off-magnet in 48 µl of EB and allowed to incubate for 10 min followed by 5 min incubation on-magnet. 44 µl of the elutant was then mixed with 26 µl of 2X Quick Ligase Reaction Buffer and purified again using AMPure XP beads according to the steps described above. The final elution volume was 44 µl of EB of which 41 µl was transfer to another 96-well plate for ligation. 2 µl of HPLC purified custom barcoded Illumina adaptors (PE5/7) were added to each bead purified digested WGA DNA. Ligation reactions were carried in total volume of 70 µl with 1 µl of ligase and 26 µl of ligase buffer. Ligation reactions were incubated as follows: 20 °C for 30 min, 65 °C for 15 min, and 4 °C forever. After adaptor ligation, 2.3 µl of each 96 adaptor ligated library was pooled and distributed equally into 3 fresh tubes (~70 µl). Pools were purified 1X using 30 µl beads as described above and eluted in 30 µl of buffer EB. Following bead purification of the pools, PCR enrichment was performed in total volume of 62.5 µl containing 2.5 µl of 10 µM PE5/7 primers and 30 µl of Phusion® High-Fidelity PCR Master Mix (NEB) according to the following parameters: (1) 98 °C for 30 s, (2) 98 °C for 10 s, (3) 65 °C for 30 s, (4) 72 °C for 30 s, (5) return to (2) for a total 10X, (6) 72 °C for 5 min, (7) Hold at 4 °C. Samples were then quantified using the Bioanalyzer and qPCR, and subsequently run on HiSeq machines.

2.6. DNA purification of bulk samples and Illumina library generation

For bulk extraction of genomic DNA from cell lines as well as clinical tissue, leftover nuclei suspensions (from which single cells were retrieved) were mixed with equal volume of 2X lysis buffer (1 ml 1M Tris-HCl pH 8.0, 200 µl 0.5M EDTA pH 8.0, 200 µl 5M NaCl, 500 µl 10% SDS, 1ml 1M DTT, 1.1ml H_2O). Lysis nuclei mixtures were then treated with 50 µl Proteinase K (20mg/ml) and incubated for 16 hrs at 55 °C. Digestion mixture was allowed to cool to room temperature followed by RNase A treatment using 5 µl of 20mg/ml RNase A. RNase A treatment was performed at 37 °C for 1 hr. Genomic DNA was then purified from Proteinase K and RNase A treated nuclei using phenol-chloroform extraction as follows: Equal volume of Phenol was added to nuclei digestion mixtures and allowed to mix gently in a rotator for 10 min. Mixtures were then spun at 13,000g at 4 °C. Aqueous phase material was carefully retrieved (avoiding interface material) and saved in a fresh tube. Phenol extraction was repeated 2X. Phenol extracted material was further purified by adding equal volume of Phenol:Chloroform:Isoamyl alcohol. Mixing, centrifugation and aqeous phase extraction were performed as described above. Phenol:Chrloroform:Isoamyl alcohol extraction was repeated 1X. DNA was then further extracted using Chloroform:Isoamyl alcohol following the steps as described above. Chloroform extraction was also repeated 1X. Chloroform extracted DNA was subsequently precipitated by adding 1/10 volume of 3M NaOAc pH 5.2, with gentle mixing, followed by the addition of equal volume Isopropanol, mixing by inverting the tube ~40X and centifigation at 13,000g for 30 min at 4 °C. Supernatent was removed by pipetting carefully as not to disturb the pellet. DNA pellets were washed with 300 µl ice-

cold 70% EtOH (1X) followed by an additional wash using 300 µl ice-cold 100% EtOH (1X). DNA pellets were allowed to dry at room temp for ~ 15 min and re-suspended in H_2O at 4 °C overnight. 0.25 – 1 ug of high molecular weight genome DNA was then sonicated using the Covaris machine at 300+/- using the following parameters: duty cycle – 10%, Intensity – 4, cycles/burst – 200, and time 80 s. Sonicated genomic DNA was then prepared for Illumina library generation using custom built barcoded adaptors as described previously in Iossifov *et al*[91], with the exception that bead purification was performed 2X.

2.7. RNA purification and Illumina library generation

Core biopsies were removed from the OCT and homogenized in lysis buffer using Hard Tissue Omni Tip Homogenizing Probes. DNA and RNA were extracted from the lysate using the Qiagen AllPrep DNA/RNA Mini Kit (Qiagen). RNA concentrations and 260/280 ratios were determined using a NanoDrop. RNA integrity was assessed using a Bioanalyzer (Agilent). Fifty to 100 ng of total RNA were reverse-transcribed and amplified using the Ovation RNA-Seq System (NuGEN). Amplified cDNA was purified using the Qiagen MinElute Reaction Cleanup Kit (Qiagen) and quantified using a NanoDrop. RNA-Sequencing was performed on the amplified cDNA at the Yale Center for Genome Analysis (West Haven, CT) or Expression Analysis, Inc. (Durham, NC). Paired-end sequencing was performed on the Illumina GAII platform using amplified total RNA with 74bp read length, yielding data on transcript abundance for a total of 22,160 genes and 34,449 transcripts, yielding about 50M reads per sample.

2.8 RNA-Seq analysis

Raw sequencing data were analyzed using RNA-SEQ Version 2 pipeline. Reads were aligned to human reference genome hg19 with Mapsplice2[92] and gene expression was quantitated using RSEM[93] (RNA-SEQ by Expectation Maximization). Each gene expression profile (Pt31 and Pt41) was normalized in the same manner as the TCGA breast cancer cohort, by setting the upper quartile value to 1000. To perform subtyping, a nearest centroid classifier was implemented using TCGA level-3 gene expression profiles along with study samples. Out of 1100 TCGA breast cancer (BRCA) samples with available gene expression data, 542 samples were mapped out with subtype information from the published study[94], including 96 Basal-like, 58 Her2-Enriched, 231 LumA, 128 LumB and 29 Normal-like samples. With the log2 transformed gene expression data aligned on PAM50 list[95], centroids for samples from each subtype were estimated and further used to subtype study samples based on 'nearest distance' criteria. Pt31 and Pt41 gene expression profiles were pre-processed in the same manner prior to the the prediction. For visualization purpose, a principal component analysis was conducted to identify the top principal components (PCs) based on pre-processed 542 TCGA breast cancer samples. Using the top 2 principal components (PC), the samples were projected onto the PC subspace. In the PCA analysis, the top 2 principal components were capable in explaining 37.7% and 25.6% of the variance of the TCGA sample matrix respectively.

2.9. Variable bin (varbin) method

In dividing the genome into bins for copy number estimation, a method is utilized that partitions the genome into bins of variable sizes based on the unique mappability of sequences across the human genome, with each bin containing the same number of mappable positions (Varbin). This is accomplished by taking 50 base pair sequences starting at each position in the reference genome, mapping them back to the reference, eliminating reads that map to multiple places in the genome (multimappers), and then setting bin boundaries such that each bin contains roughly the same number of uniquely mappable positions. This was done to compute bin boundaries when dividing the genome to 50K, 20K, and 5K bins. Importantly, because single end reads were extracted from hg19 and mapped using bowtie to define the variable bins, only Illumina single end sequencing data that are mapped with Bowtie are useful for determining the copy number profile with the boundaries computed. If BWA is preferred to Bowtie, then the simulations will have to be repeated to define a new set of bin boundaries. The same applies to paired-end sequencing data or sequence data obtained using a different platform (ABI SOLiD for example). The output file of the Varbin algorithm contains sequence counts in the assigned genomic bins. This data is further normalized and processed to yield integer copy number values (see section 2.10). Additionally, for a number of regions in the genome, very high read depth compared to the expected norm occurs. These regions are found to consistently display the high read depth in both bulk as well as single cell sequencing data and many are found in bins surrounding centromeres. Using data from 54 normal diploid single cells, these bins (designated as

"bad bins") were determined as follows. Bincounts were divided by the mean for each cell to normalize for differences in total read count between each cell. For each chromosome, the mean of the bins over all cells is subsequently subtracted from each normalized bin count to normalize for differences between chromosomes. The mean and standard deviation of the autosomes was then used to compute an outlier threshold corresponding to a p-value of 1/N, where N is the number of bins used. This was done for the 5K, 20K and 50K bin data sets. These bins are masked from down-stream copy number analysis.

2.10. Sequence alignment and single cell copy number analysis

Multiplexed single cell sequencing libraries were split according to their unique barcode identifiers specified by the first 8 bases of the sequencing reads. Single cell sequencing data were aligned to the human reference genome hg19 using bowtie[96]. Reads were sorted, PCR duplicates removed, and then indexed using SAMtools[97]. Uniquely mapping reads were counted for each bin and normalized for GC bias using lowess smoothing. Normalized read count data were then segmented using circular binary segmentation (CBS)[98]. For copy number estimation in single cells, an approach based on least-squares fit is utilized as follows: When analyzing data from a single cell, the copy number at any point in the genome must be an integer. Thus, if the data were accurate, then after segmentation, segmented mean values should have a clear multimodal distribution with a peak representing each copy number present in the genome. The data at this point in the analysis is centered around 1, meaning that the mean value across the bins (5000, 20000,

or 50,000) of the segmented value is close to 1. In a diploid genome this would represent a copy number of 2, with regions of copy number 1 having a segmented value near 0.5 and regions of copy number 3 having a segmented value near 1.5. These could easily be converted to copy number estimates by multiplying the segmented value for each bin by 2 and rounding to the nearest integer. This is the basic idea used to estimate copy number in single cell data. In rearranged cancer cells where the copy number of genomic segments is unknown, in order to find the best multiplier, the segmented profile is multiplied by 1.5, 1.55, 1.6, 1.65, ... 5.5 (81 different values) to compute what is known as quantal error for each multiplier. This is the sum of the squared difference between the multiplied segmented profile and the multiplied segmented profile rounded to the nearest integer. The multiplier that gives the smallest quantal error is deemed the best fit and used to estimate copy number. This quantal error can also be used as a quality control parameter. Cells with a large quantal error can really be multiple cells, parts of cells, or have degraded DNA. For Heatmap plots, single cells were hierarchically clustered based on their genome-wide copy number profiles using Manhattan distance function and clustered according to Ward method.

2.11 CORE (Cores of Recurrent Events)

Core analysis was performed as described in Krasnitz *et al*[99]. Briefly, segments with integer copy number values above/below the reference were considered amplified/deleted. Copy-number events in each cell were derived by slicing, and cores, i.e., regions of significantly recurrent ($p<0.05$) gains and losses, were determined by

applying the CORE method to the entire set of single cell genomes. Finally, the incidence table was computed, with rows and columns corresponding respectively to cells and cores and with values in the [0,1] interval quantifying the best match between an event in the cell and the core. Single cells that contained statistically significant cores were judged to be part of the cancer phylogeny and used for downstream analysis while cells lacking cores (mostly cells with the vast majority of the genome at copy number 2) were judged to be normal cells.

3. CHAPTER 1: Optimizing coverage in a multiplexing strategy

Work done in collaboration with Jude Kendall and Anthony Leotta

3.1. Down-sampling analysis reveals minimal data requirements for copy number determination at a resolution of 50K bins

CNV analysis by sequencing typically counts the number of reads that uniquely map to bioinformatically computed segments or 'bins' of genomic sequence[65,66]. Recently it has been shown, from sequencing data of uniformly amplified single cell genomic DNA, that the copy number of a particular bin is directly proportional to the number of sequencing reads that map within it[79,80]. 50 thousand bins were used to divide the genome (50K bins), with an average bin length of 60 kb. The profiles produced have clean breakpoints and segments with quantal values, as one expects from single cell data. At the published coverage, this averaged 160 maps per bin, clearly an excess. But how much data (measured as the number of sequencing reads) is required to produce a clean, quantal, genome-wide copy number profile from a single cell at 50K bin resolution? Although the answer can be approached mathematically on assumptions about binomial sampling distributions, the confident detection of minimum features, and expectations of quantal values, an empirical approach using the same cancer cells previously analyzed was utilized. Single cell sequencing data[79] for a rearranged cancer cell (DNA content=2.95N) for which 8 million uniquely mapped reads were available was retrieved and correlation and copy number analysis on down-sampled data sets was performed.

Normalized read counts of data down-sampled to 4, 2, 1, 0.5, and 0.25 million reads plotted against the original 8 million reads data set demonstrate strong correlations down to 1 million reads (R2 = 0.939) (**Figure 1**). The 2 million read copy number profile (about 40 reads per bin) was highly similar to the profile generated from the original 8 million read single cell data set using 50K bins (**Figure 2**). Using fewer reads than this retained features of the breakpoint profile, but the quantal nature of the copy number segments became less clear (**Figure 3**). Two million uniquely mapped reads were also sufficient to recapitulate the copy number landscape of tumor cells with different DNA contents (**Figure 4**). Thus, irrespective of DNA content, 2 million uniquely mapped reads are sufficient to retrieve the genome-wide copy number profile of a single cell when dividing the genome in 50K bins.

3.2. Adjusting bin lengths lowers minimal read requirements and maintains genome-wide copy number profile

Are 50K bins needed? Given that the majority of copy number alterations found in bulk analysis of tumor genomes are on the order of mega bases (Mb) or greater[31], decreasing the number of bins (i.e. increasing bin lengths) should decrease sequencing read requirements while retaining the majority of copy number alterations in the cancer genome. Reanalyzing the down-sampled data using 20K and 5K bins (calculated using the variable bin method – see Methods) revealed that strong correlations were maintained with the original 8 million data set down to 1 million and 0.25 million uniquely mapped reads for 20K and 5K bins, respectively (**Figures 5)**. Importantly, the copy number

profiles at 20K and 5K bins were largely similar to the profiles at 50K bins while maintaining the quantal nature of the copy number segments (**Figure 6).** 96% and 75% of the breakpoints, detected at a resolution of 50K bins, were called at bin resolutions of 20K and 5K, respectively, with the down-sampled data. Naturally, at lower resolutions of 20K and 5K bins, some focal alterations were missed (**Figure 6, red arrows**). Furthermore, 1 and 0.25 million sequencing reads for 20K and 5K bins respectively were also sufficient to retrieve genome-wide CNV information in cancer cells with different DNA contents (**Figure 7**). The data together, when taking into account current average HiSeq output of 200 million reads per lane, indicate that up to 500 single cell genomes can be multiplexed and analyzed on a single HiSeq lane. **Table 1** lists the multiplexing capacity and the genomic bin resolution given different # of bins.

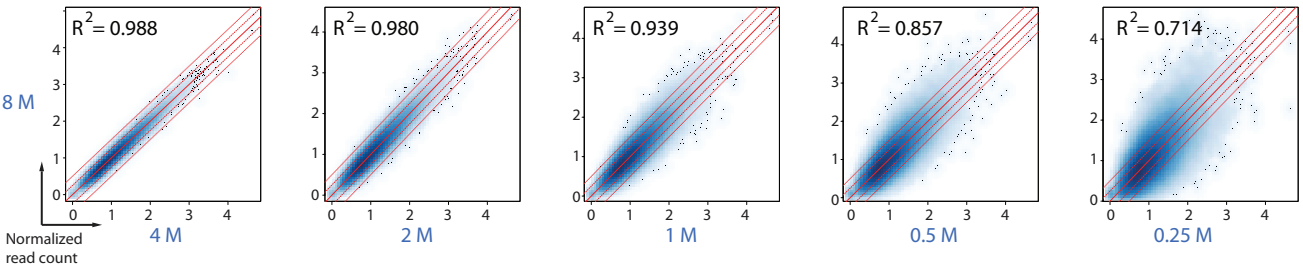


Figure 1: <u>Down-sampled data reveal strong correlations at 50K down to 1 million reads.</u>

Original 8 million read data set was down-sampled to 4, 2, 1, 0.5, and 0.25 million reads. Normalized data from down-sampled data sets using 50K bins was correlated to original data and plotted as scatter density correlation plots. Pearson R^2 correlation coefficients are shown.

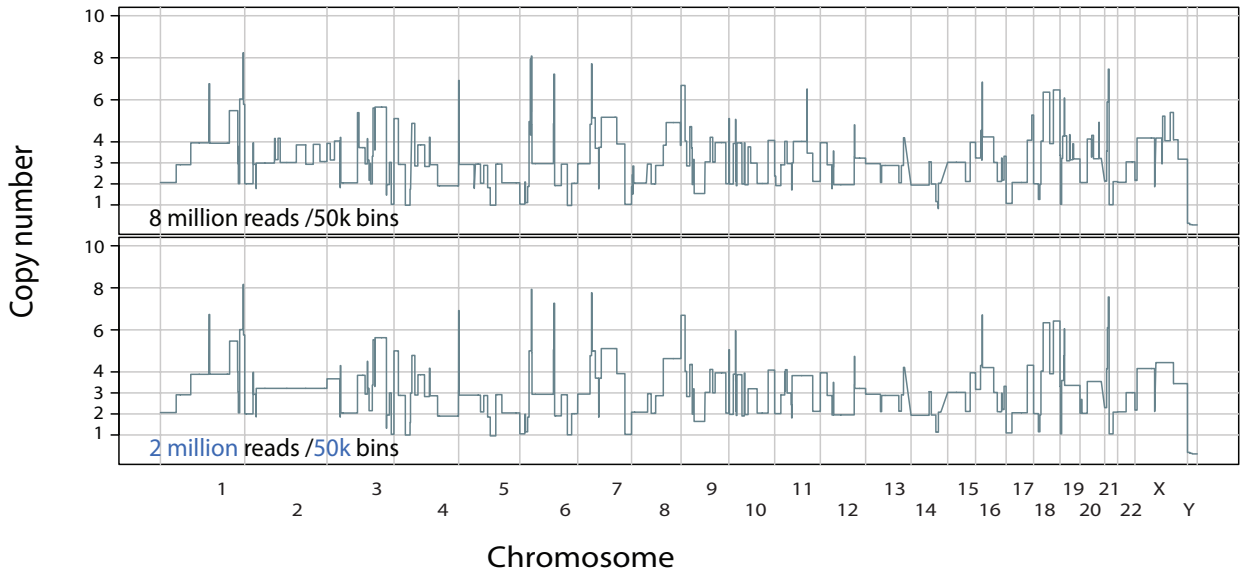


Figure 2: <u>2 million uniquely mapped reads are sufficient to reproduce a quantal genome-wide copy number profile when using 50K bins.</u>

Genome-wide CNV profile at 50K bins of 2 million read down sampled data (lower panel) was plotted with the original profile 8 million read profile (top panel).

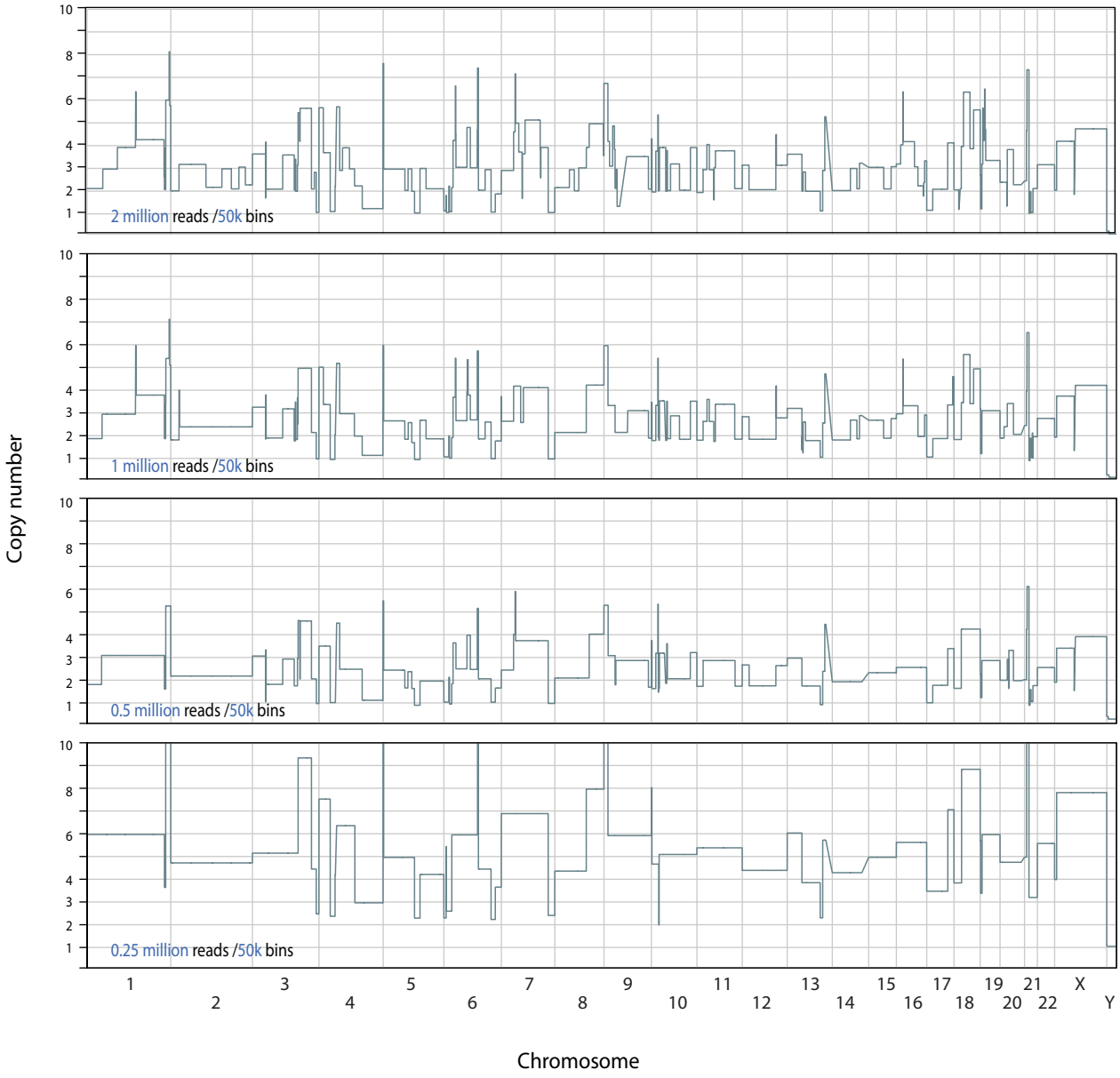


Figure 3: <u>Down-sampling below 2 million reads when using 50K bins allows observation of some features of the breakpoint profile but results in loss of the quantal nature of the CNV profile.</u>

Copy number profiles of data down-sampled to 2, 1, 0.5, and 0.25 million using 50K bins are plotted genome-wide. Some loss of quantal behavior is evident at 1 million reads with progressive deterioration with further down sampling.

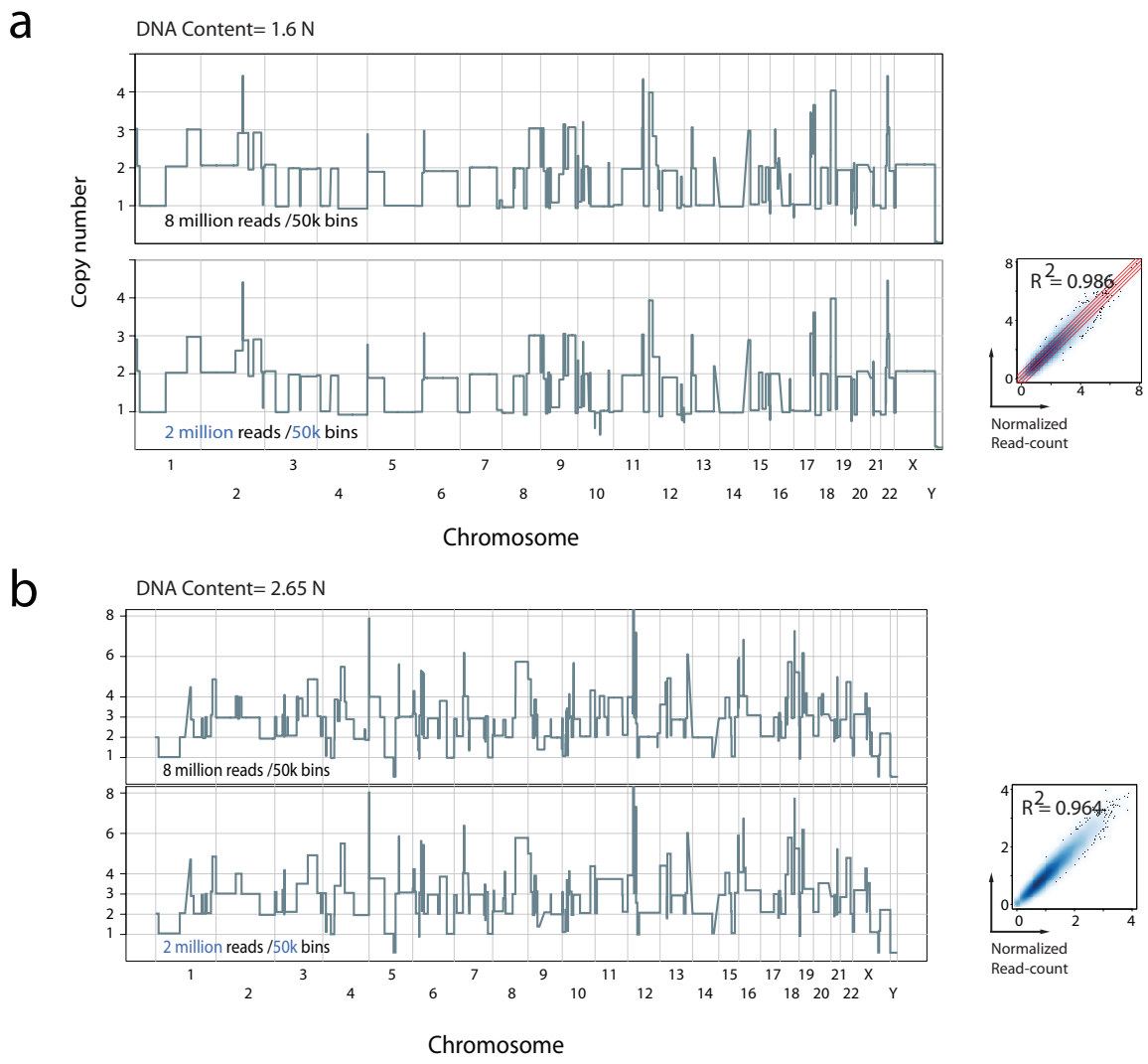


Figure 4: <u>2 million uniquely mapped reads is sufficient to reproduce genome-wide CNV profile of cancer cells with differing DNA content using 50K bins.</u>

Copy number profiles from 2 million read down-sampled data of single cancer cells of different DNA content (Panel a=1.6N, panel b=2.65N) are plotted genome-wide. Box plots next to CNV figures illustrate scatter density correlations of normalized read counts between original and down-sampled data. Pearson R^2 correlations are shown.

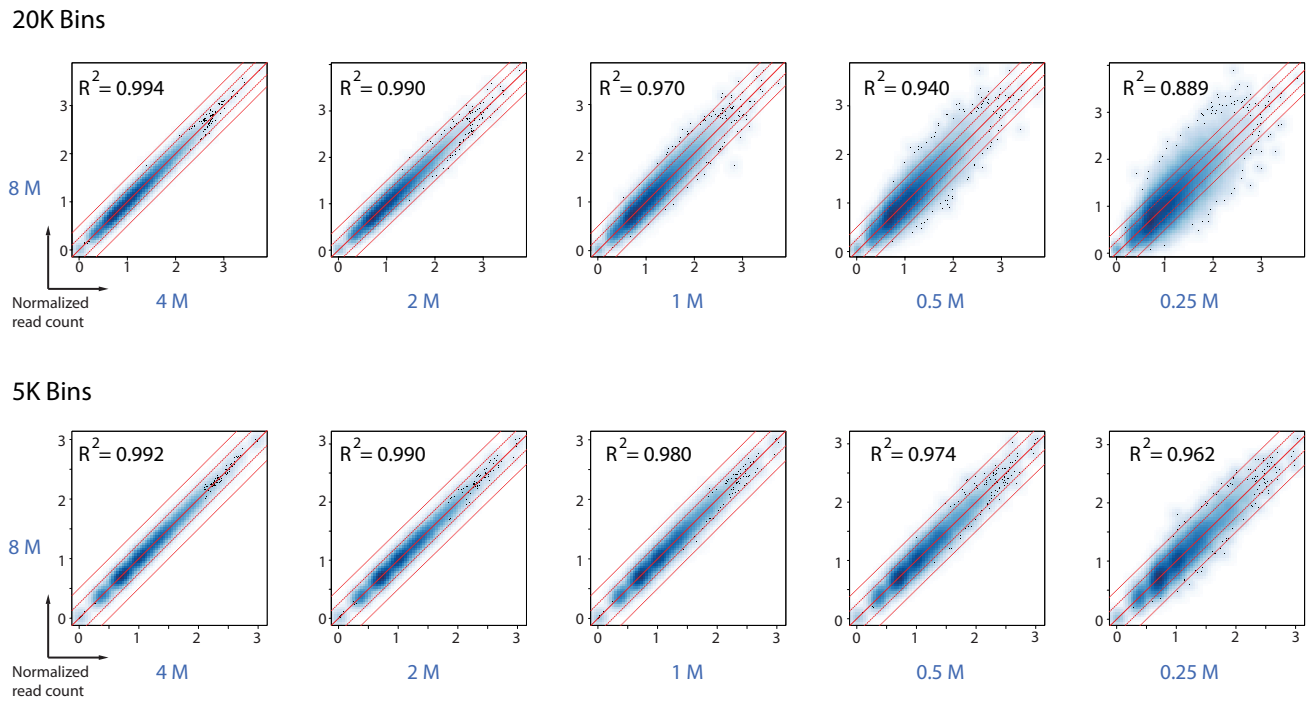


Figure 5: <u>Dividing the genome into 20K and 5K bins allows for better correlations of down sampled data down to 1 million and 0.25 million reads respectively.</u>

Normalized read counts of down sampled data at 4,2,1,0.5, and 0.25 million reads at 20K and 5K bins are plotted as density scatter correlation plots. Pearson R^2 correlation values are shown.

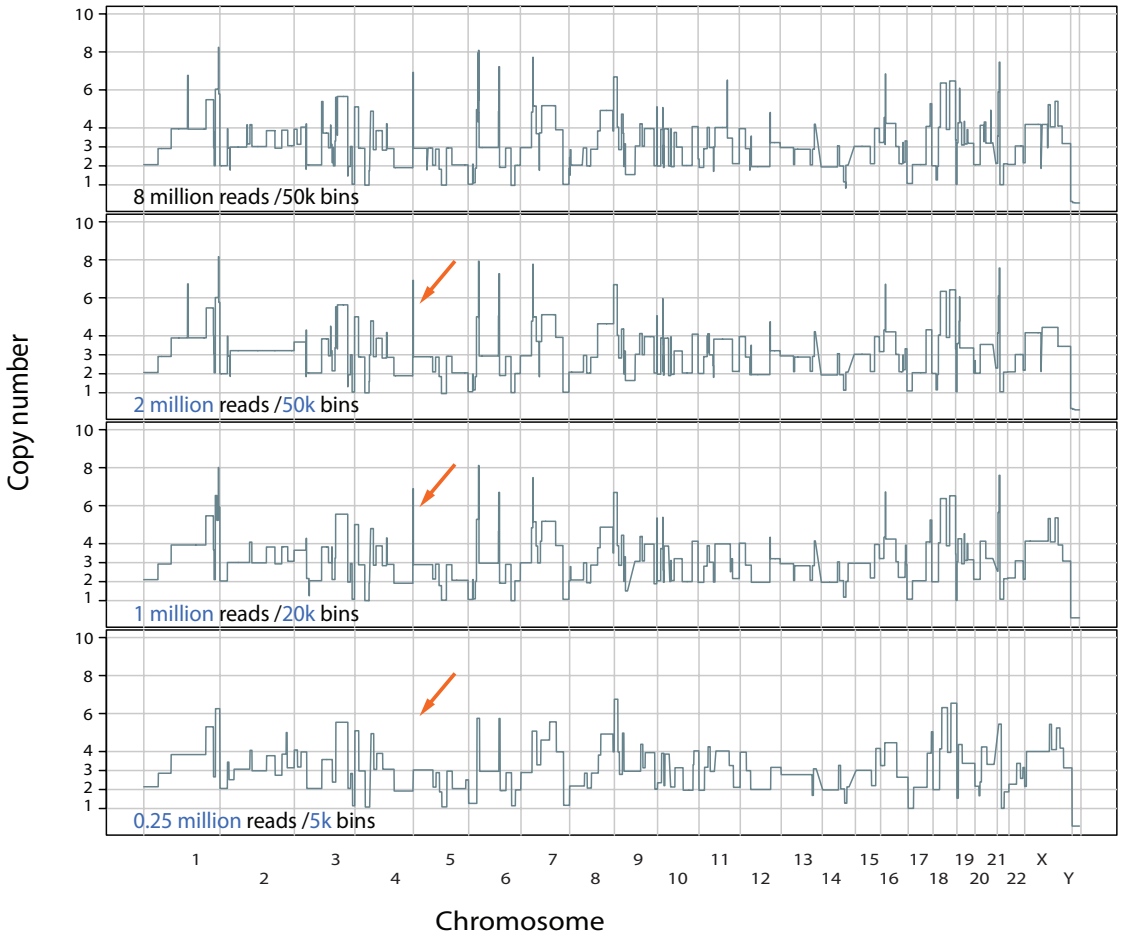


Figure 6: <u>1 million and 0.25 million reads are sufficient to recapitulate genome-wide copy number profiles at 20K and 5K respectively.</u>

CNV profiles of calculated minimal read requirements (2 million, 1 million, and 0.25 million sequencing reads) per number of bins (50K, 20K, and 5K) respectively are plotted. Red arrows point to copy number variants that are lost with lower resolutions (i.e. small number of bins/larger bin lengths).

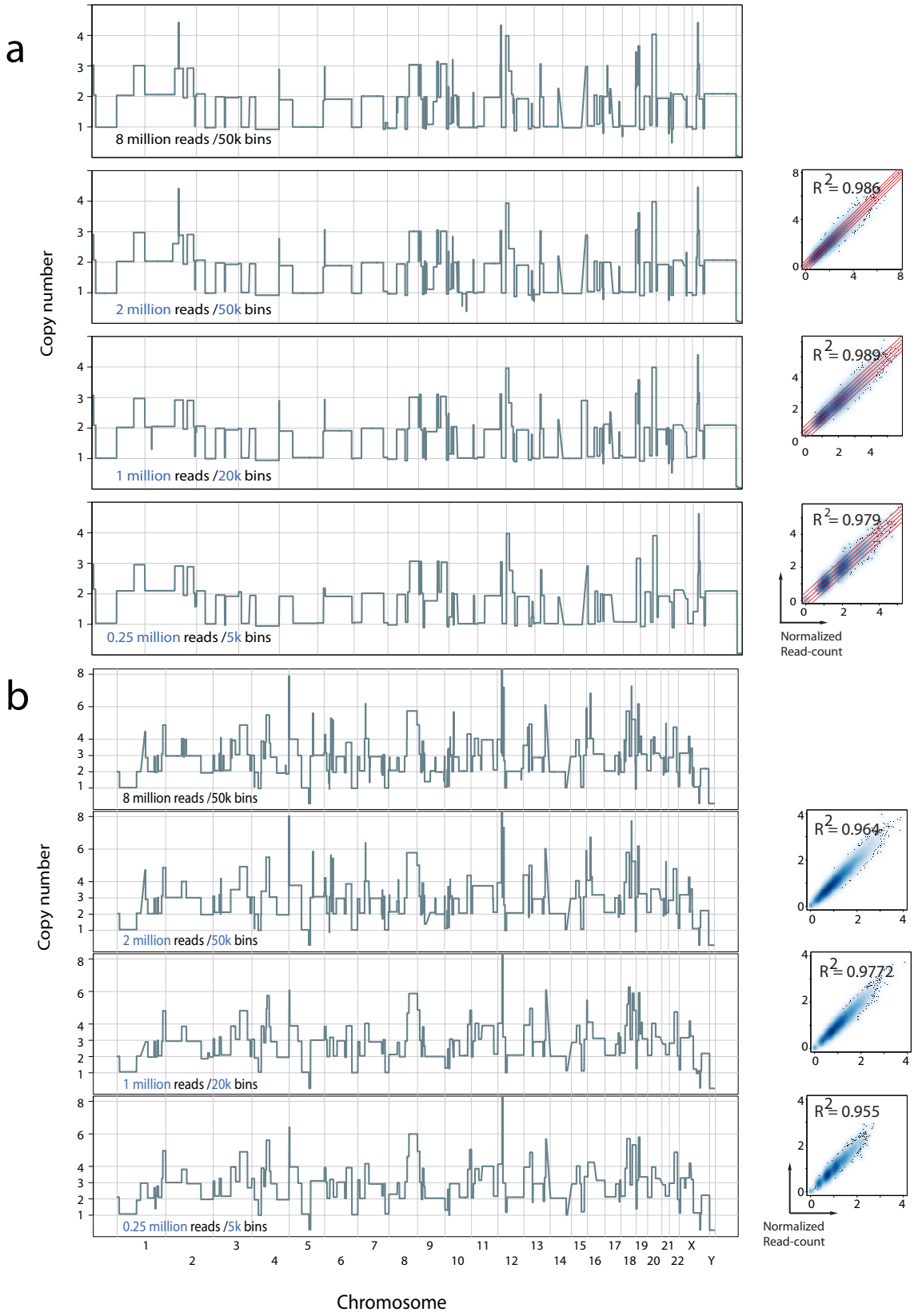

a
8 million reads /50k bins
2 million reads /50k bins
1 million reads /20k bins
0.25 million reads /5k bins
Copy number
R2 = 0.986
R2 = 0.989
R2 = 0.979
Normalized Read-count
b
8 million reads /50k bins
2 million reads /50k bins
1 million reads /20k bins
0.25 million reads /5k bins
Copy number
R2 = 0.964
R2 = 0.9772
R2 = 0.955
Normalized Read-count
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome

Figure 7: Down-sampled data at 1 million and 0.25 million reads reproduces quantal genome-wide copy number profiles at 20K and 5K bins respectively of cells with different DNA content.

CNV profiles were plotted for each caculated minimal read requirement/# of bins (2 million/50K, 1 million/20K, and 0.25 million/5K). Scatter density correlation plots of down-sampled data sets correlated to original 8 million read data are illustrated.

Number of bins	Bin Size *	# of reads required (in millions)	**MiSeq** ** Approximate multiplexing capacity	**HiSeq** *** Approximate multiplexing capacity
50K	60 Kb	2	5	70
20K	150 Kb	1	10	140
5K	600 Kb	0.25	42	**560**

Table 1: Multiplex capacity and bin parameters for copy number determination of single cells using different # of bins.

Approximate multiplexing capacity is calculated assuming equal distributions of multiplexed single cell libraries in final pool.

* Bin size calculated using the varbin algorithm.

** Calculated with the presumed output of 12 million sequencing reads per MiSeq lane for single end sequencing.

*** Calculated with the presumed output of 200 million sequencing reads per HiSeq lane for single end sequencing.

4. CHAPTER 2: An optimized DOP-PCR molecular approach for high level multiplexing; C-DOP-L

Work done in collaboration with Brian Ward

DOP-PCR methodology is employed for WGA because it amplifies more uniformly across the genome than other methods, and when the goal is CNV analysis more reproducible results with lower noise are obtained[79,80].

Maximizing the efficiency of sequencing by identifying minimal read requirements to facilitate multiplexing is not the only problem that needs to be addressed to optimize the efficiency of highly multiplex single cell CNV profiling. Performing the steps of WGA and library preparation protocols, involving sonication, end repair, A-tailing, and ligation for each cell individually takes a great deal of benchwork and can cost as much as $50 per cell in reagents alone, making the procedure itself a target for optimization. Moreover, the resulting DNA molecules following DOP-PCR carry universal 30bp sequences at the ends and even when sonicated, the universal DOP primer sequences remain on a substantial fraction of the DNA molecules, that when sequenced, cause decreased complexity (**Figure 8**), lower quality data and decreased mappability for some reads.

To circumvent the above-mentioned issues, a method was devised, termed Cleavable DOP-PCR Ligation (C-DOP-L), that incorporates restriction enzyme digestion of the universal sequences at the ends of WGA DNA via the SEQXE kit (Sigma-Aldrich), with an "NN-mediated" DNA ligation of barcoded Illumina adaptors (**Figure 9**). In the C-DOP-L method, single cell genomes are amplified using DOP-PCR similar to

what have been reported before[79,80]. However, the degenerate oligo-nucleotide differs in that it incorporates a recognition site for a type IIS restriction enzyme (isoschizomers AcuI and Eco57I, (CTGAAG 16/14). When added to the WGA DNA, the enzyme recognizes its binding site, cleaves 16/14 (top/bottom strand) bases away from its recognition sequence, effectively removing the entire universal sequence found at the ends of the DNA molecules. Furthermore and importantly, the digestion leaves 3'-NN overhangs (where N is any base). These overhangs are subsequently used in the ligation of barcoded Illumina adaptors designed to carry 3'-NN overhangs on the P5 adaptor. To test the method, 96 modified Illumina adaptors carrying custom barcoded adaptors with sufficient complexity (equal distributions of A, T, C and G base pairs) in the first 4 bases were designed, synthesized and used in subsequent experiments (**Figure 10**).

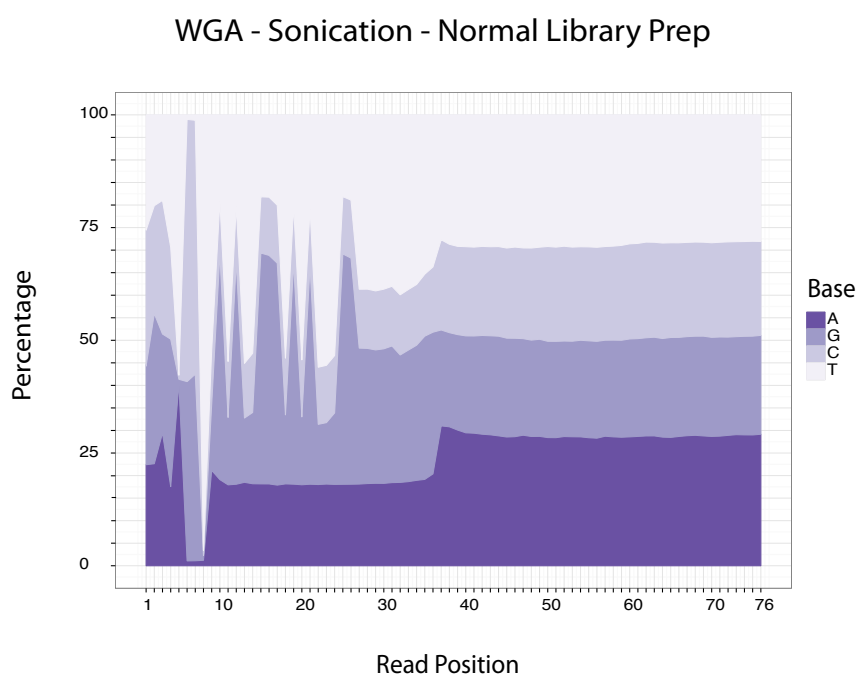


Figure 8: <u>Sonication of WGA DNA using DOP-PCR does not remove universal sequences at the ends of the DNA molecules.</u>

Nucleotide frequencies at sequenced read positions of the output sequencing data are plotted. First 8 bases denote the custom barcodes utilized in sequencing. Nucleotide percentages from position 9 thru 38 illustrate features of the universal sequences that are still present even after DNA sonication of the WGA DNA.

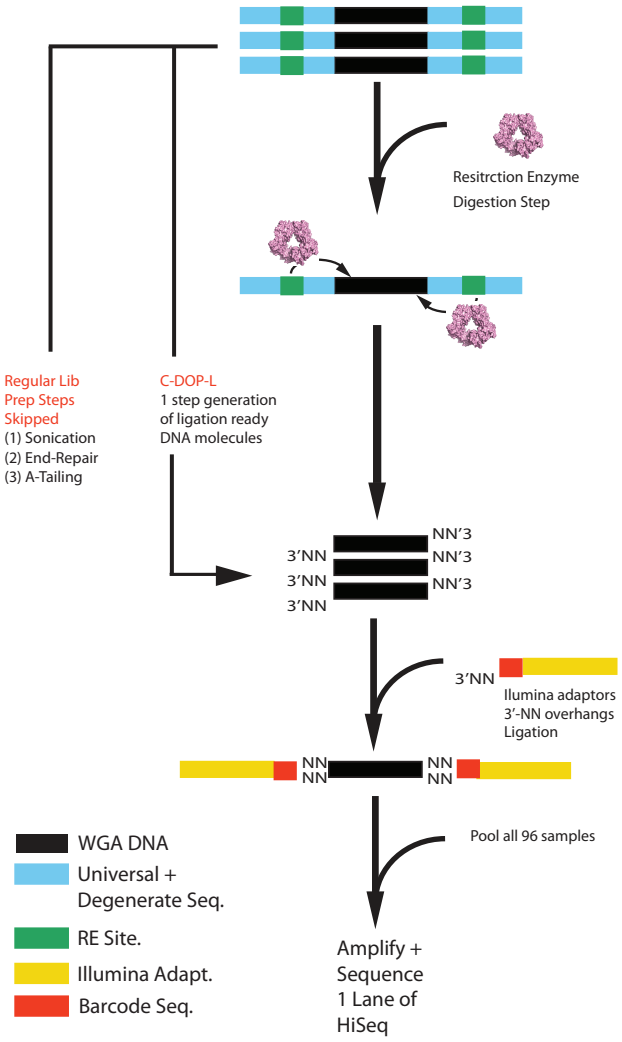


Figure 9: Schematic overview of C-DOP-L approach for high level multiplex sequencing of singe cell genomes.

In brief, WGA DNA is digestion with a type IIS restrcition enzyme to cleave the universal sequences found at the ends of WGA DNA. The digestion reaction leaves 5'-NN overhangs (where N is any base: A,T,C,G). Digested DNA is then ligated to barcoded Illumina sequencing adaptors that are designed to contain 3'-NN overhangs. After barcode addition, samples are pooled, amplified and sequenced on a HiSeq machine. WGA; Whole Genome Amplified. RE; Restriction enzyme. N; any base (A,T,C,G).

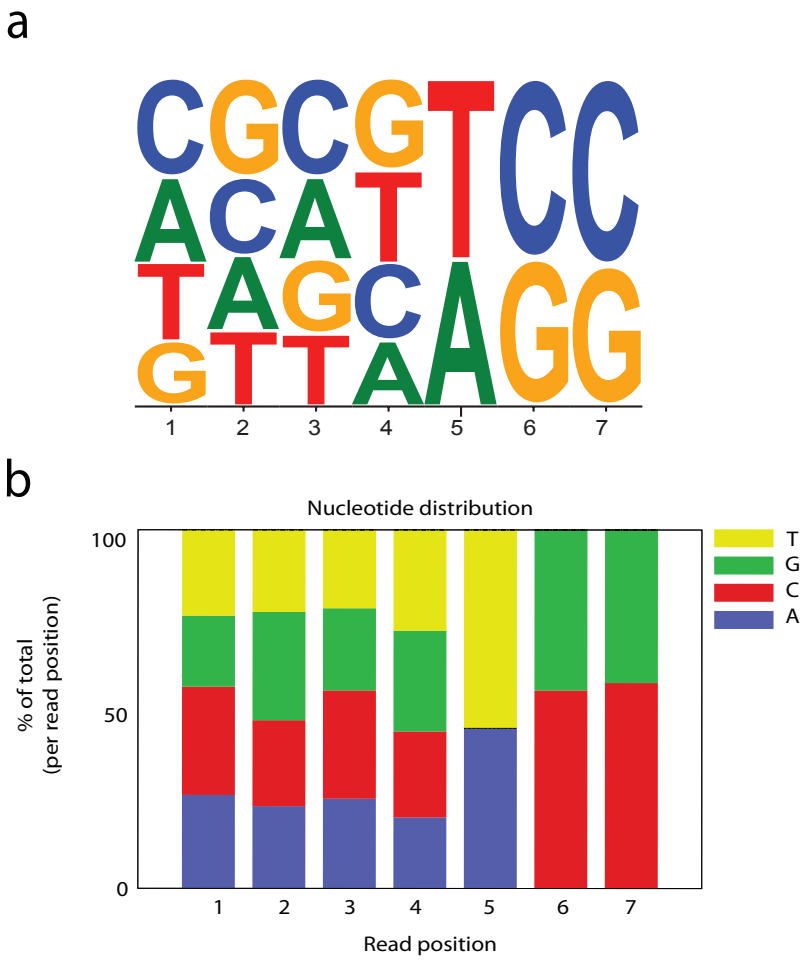


Figure 10: <u>Custom Illumina barcodes used in the C-DOP-L method.</u>

Barcode nucleotide distributions of all 96 barcodes plotted in WebLogo format (a) and chat plots (b).

5. CHAPTER 3: Validation of C-DOP-L with cell lines

Work done in collaboration with Hilary Cox, Sean D'Italia, Linda Rodgers and Anthony Leota

5.1. C-DOP-L provides uniform whole genome amplification and does not introduce biases

To ensure that the modification of the degenerate oligo-nucleotide primer does not affect the uniformity of the WGA reaction or introduce distortions to the genome normal non-genomically rearranged cells were examined. Approximately 100 genomes per HiSeq Illumina lane was chosen, a convenient number for microplate processing.

Single nuclei from a diploid EBV immortalized lymphoblastoid cell line (315A) derived from a normal male were sorted, selecting for diploid nuclei, deposited into a 96-well plate and amplified. Of the 96 sorted single nuclei, 95 were successfully amplified (i.e. yielding a minimum of 2 µg of total WGA DNA), processed using the C-DOP-L method, and sequenced on a single lane of HiSeq2000. Sequencing reads for single cells were deconvoluted, mapped to the human genome, and processed using the variable bin algorithm for copy number determination[79,80] (See Methods). Sequence reads displayed normal nucleotide complexity as expected **(Figure 11).** For all single cells processed an average of 1.5 million uniquely mapped reads with a range of 0.25 to 3.6 million, with all cells having minimum of 0.25 million (**Figure 12**). Sequenced single cell DNA displayed GC amplification bias that was comparable in magnitude to previous work using DOP-PCR and was easily corrected using lowess smoothing (**Figure 13**). Importantly, the C-

DOP-L method maintained the minimal sequence bias exhibited in previous work using the DOP-PCR approach. The uniformity of the amplification reaction was maintained as demonstrated by the tight histogram distributions of the normalized read count data as well as the genome-wide copy number profiles, revealing the vast majority of the genome at copy number 2 (**Figure 14**). Furthermore, multidimensional scaling of the 315A single cell copy number profiles showed tight clustering for the majority of single cells (88 single cells out of 96 sequenced single cells) (**Figure 15**). All of these 88 cells displayed consistent normal genome-wide copy number profiles with all of the autosomes at copy number 2 and the sex chromosomes at copy number 1 attesting to the reproducibility of the method (**Figure 14**). Two cells were distant in the multi-dimensional scaling graph from the cluster (**Figure 15, red arrows**) with one cell displaying a chromosome wide duplication of chromosome 2 and another cell displaying heterozygous focal deletions on chromosome 4 (**Figure 16 a and b**). Another 5 cells (outside of the black circle in figure 3b) displayed deviations from discrete integer copy number profiles and more spread distributions of normalized read count data (**Figure 16 c, d, and e**). These profiles could be the result of an error in the WGA amplification process, or cells caught early in S phase of the cell cycle. Occasionally non-recurrent focal deletions or duplications are observed (**Figure 15, red arrows**) in otherwise normal cells. The nature of these events is currently unknown and likely represent somatic events.

5.2. C-DOP-L provides accurate determination of rearranged copy number profiles of single cancer cells in a highly multiplex manner

Work done in collaboration with Linda Rodgers, Sean D'Italia, Michael Riggs, and Anthony Leotta

To further validate the approach, single nuclei from a rearranged human breast cancer cell line were profiled. Flow sorting 96 single nuclei from the pseudo triploid (apparent DNA content 3.65N by FACS) breast cancer cell line, SK-BR-3, followed by WGA and C-DOP-L library preparation resulted in 94 successfully amplified and ligated products (97.9%). These were loaded on a single HiSeq 2000 lane and after informatic processing produced genome-wide copy number profiles (20K bins) that very closely recapitulated that of the corresponding SK-BR-3 bulk DNA (R^2 Pearson correlation = 0.963) (**Figure 17**). Importantly, smoothing kernel density plots of the normalized sequencing data revealed the quantized nature of the single cell data with densities corresponding to discrete copy number integer values (**Figure 18**). In addition, presumed driver genomic alterations observed in the bulk copy number profile, such as high level amplification of the MYC locus on chromosome 8, the heterozygous deletion of DCC on chromosome 18 and the homozygous deletion of a cluster of zinc finger proteins on chromosome 18, were observed in 100% of the single cells sequenced (**Figure 19**). Interestingly, multidimensional scaling of all 94 integer copy number profiles resolved two distinct clusters corresponding to a major subpopulation (sub-population 1) and a minor sub-population (sub-population 2) (**Figure 20**). Hierarchal clustering of the single cell profiles plotted in the form of a copy number heatmap clearly illustrates that the two

sub-populations are derived from the same lineage with the vast majority of the genome present at the same copy number in almost all single cells, for example chromosomes 2, 7, and 11 (**Figure 21, black arrows**). Importantly, the two sub-populations differed significantly with different copy number states on chromosomes 5, 14, and 19, among others (**Figure 21, red arrows and Figure 22**). Some of these events are also evident in the bulk SK-BR-3 copy number profile as segments with non-integer copy number values (**Figure 23**). This genomic heterogeneity of a cancer cell line is not restricted to SK-BR-3 as another breast cancer cell line (MDA-MB-231) also revealed substantial heterogeneity were three distinct subpopulations were observed (**Figure 24**). Thus, the data demonstrate the robustness and accuracy of the C-DOP-L highly multiplex single cell sequencing approach in profiling cancer genomic heterogeneity.

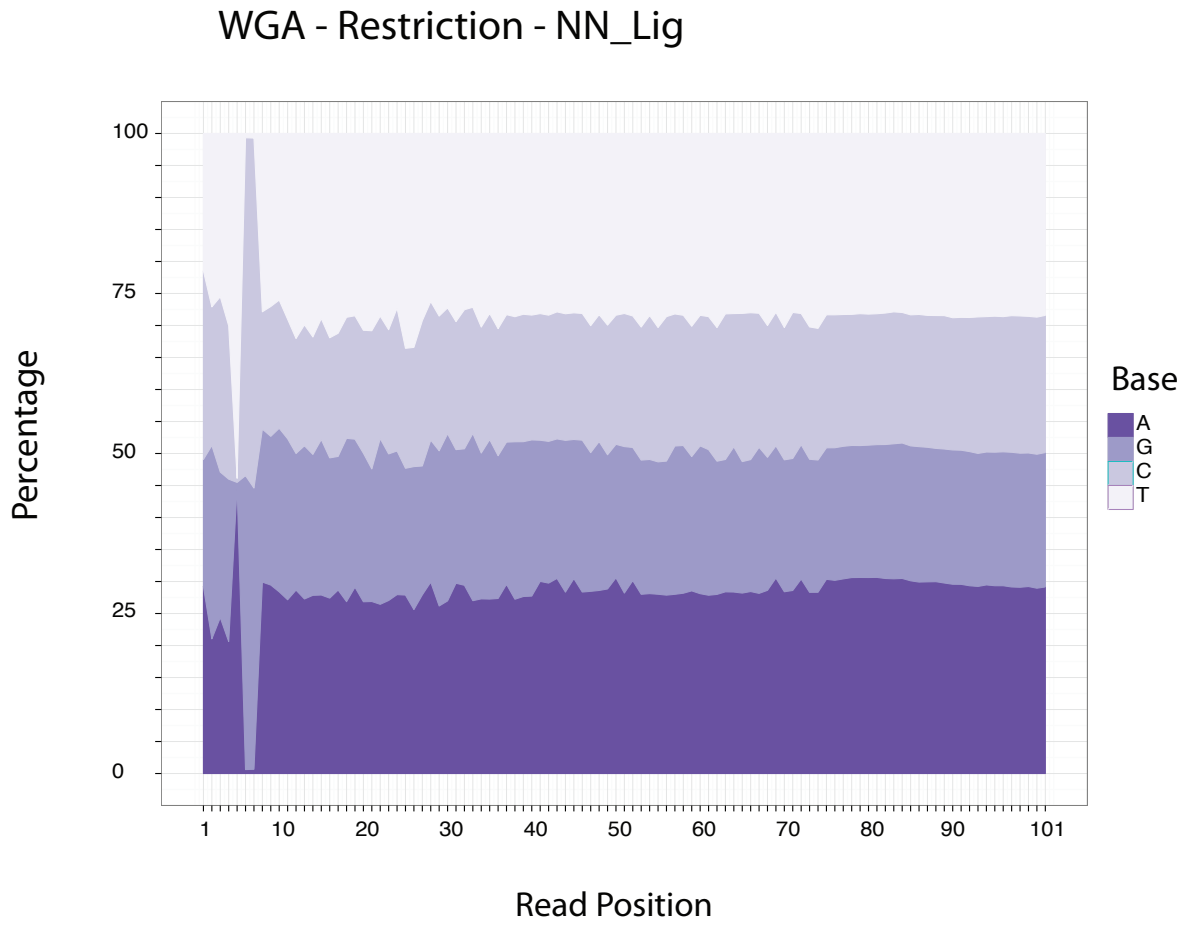


Figure 11: <u>C-DOP-L approach facilitates removal of universal sequences and re-introduces nucleotide complexity to sequenced DNA.</u>

Nucleotide frequencies at sequenced read positions of the sequencing data are plotted. First 8 bases denote the custom barcodes utilized in sequencing. Compared to Figure 8, data illustrates the effecive removal of WGA universal nucleotide sequences at the ends of DNA molecules.

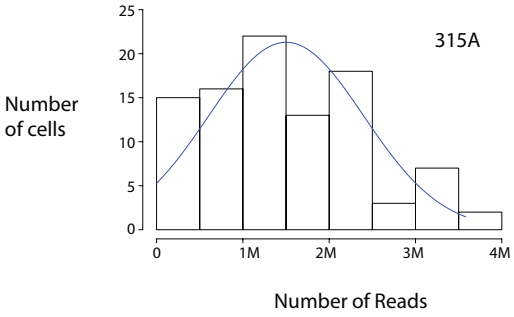


Figure 12: <u>Sequencing 96 single nuclei on a single HiSeq lane results in sufficient number of reads for analysis of all cells processed.</u>

Histogram distribution of uniquely mapped reads for all 95 single nuclei of the 315A cell line sequenced using the C-DOP-L approach were plotted.

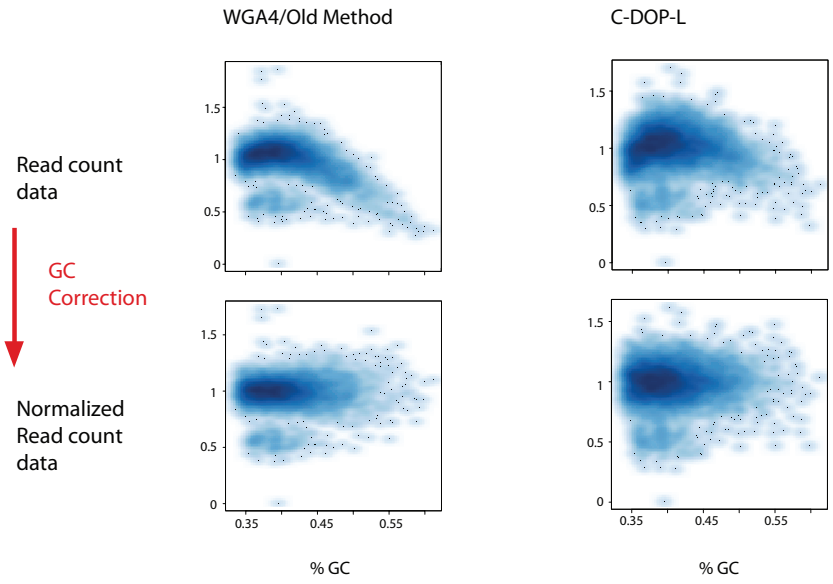


Figure 13: <u>C-DOP-L approach displays the GC bais inherent in PCR amplification and can be corrected for using lowess smoothing.</u>

Read count data were plotted as a function of bin GC content before and after lowess normalization for both WGA4 and C-DOP-L.

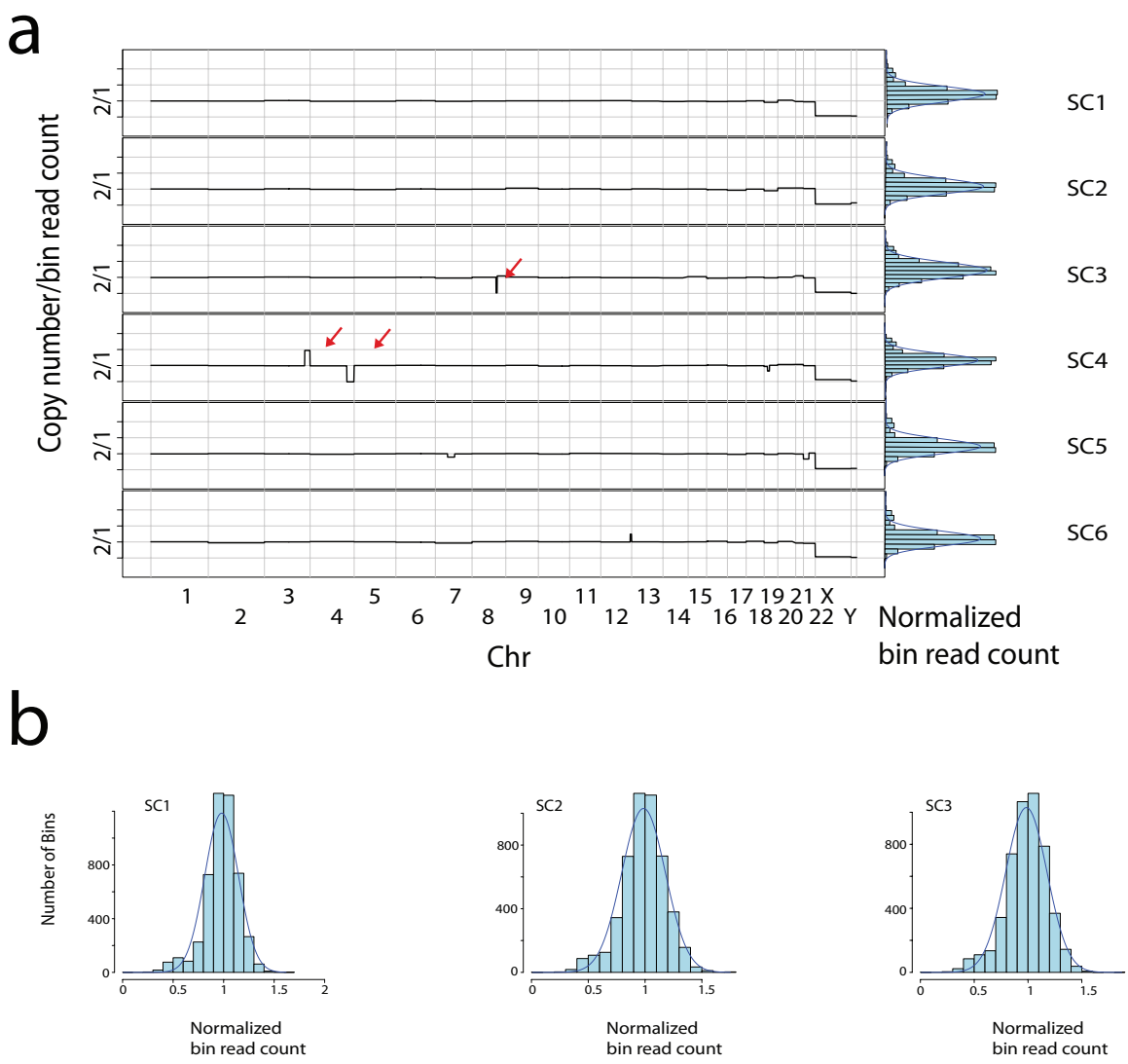


Figure 14: <u>C-DOP-L approach uniformly amplifies the genome and exhibits minimal amplication bias.</u>

(a) Left panel - Sequence data of 315A cells was processed for copy number determination and plotted genome wide. Right panel - Normalized read coun data plotted as histogram distrubtions illustrating the uniformity of the amplification reaction. (b) Expanded view of normalized read count data for single cells processed using C-DOP-L. SC, single cell. Chr; chromosome.

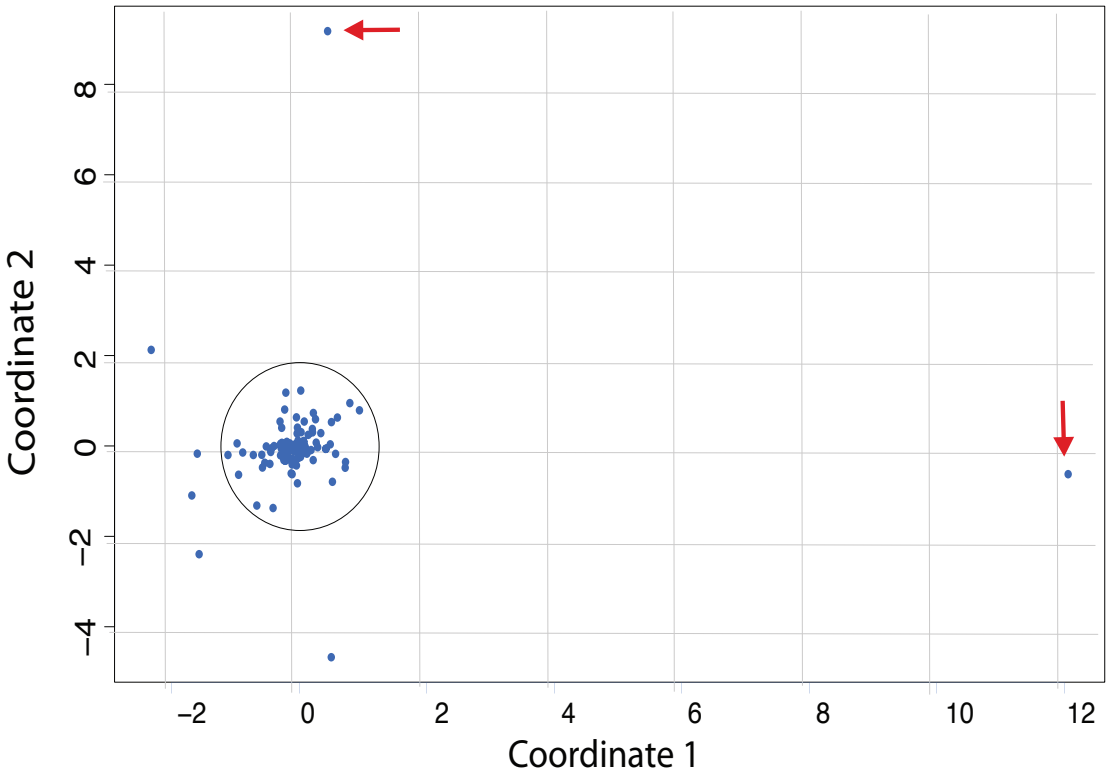


Figure 15: <u>Multi-dimensional scaling illustrates tight clustering of the 315A CNV profiles</u>

315A single cell data was scaled according to their genome-wide copy number profiles and plotted. Black circle denotes the tight clustering of the majority of the profiles displaying normal copy number profiles (i.e. autosomes at copy number 2 and sex chromosomes at copy number 1). Red arrows point to outlier cells that carry large somatic genomic rearrangements.

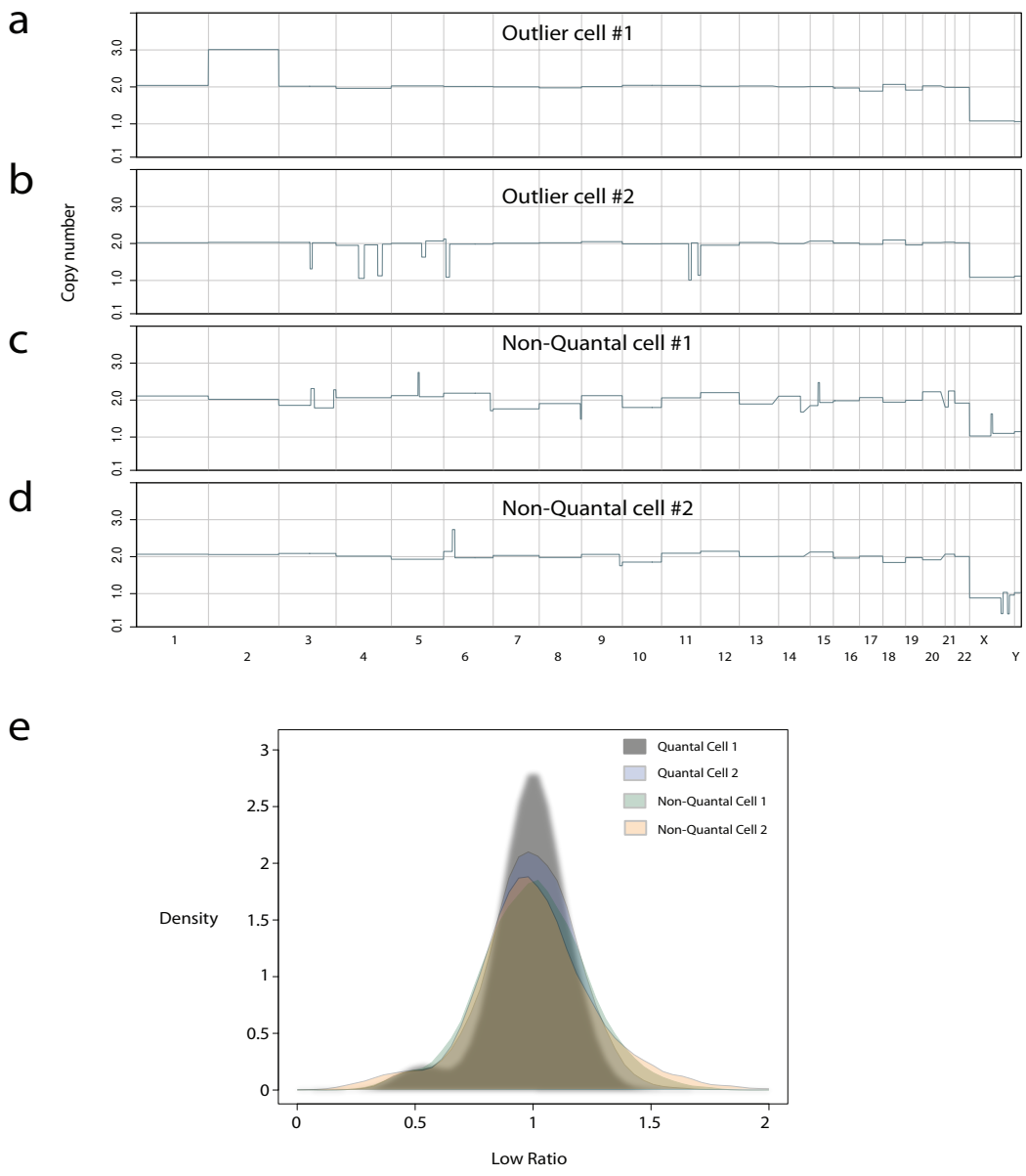


Figure 16: <u>A minority of cells display large somatic rearrangements or non-quantal copy number values.</u>

Outlier cells from the multi-dimensional graph (Figure 14) were plotted genome wide. CNV profiles illustrate a whole chromosome gain of chromosome 2 (a) and a cluster of deletions of chromsome 4 (b). Cells displaying non-quantal copy number values plotted genome wide (c and d).

(e) Quantal and non-quantal histogram distributions of normalized bin read counts illustrate more spread distributions in non-quantal cells.

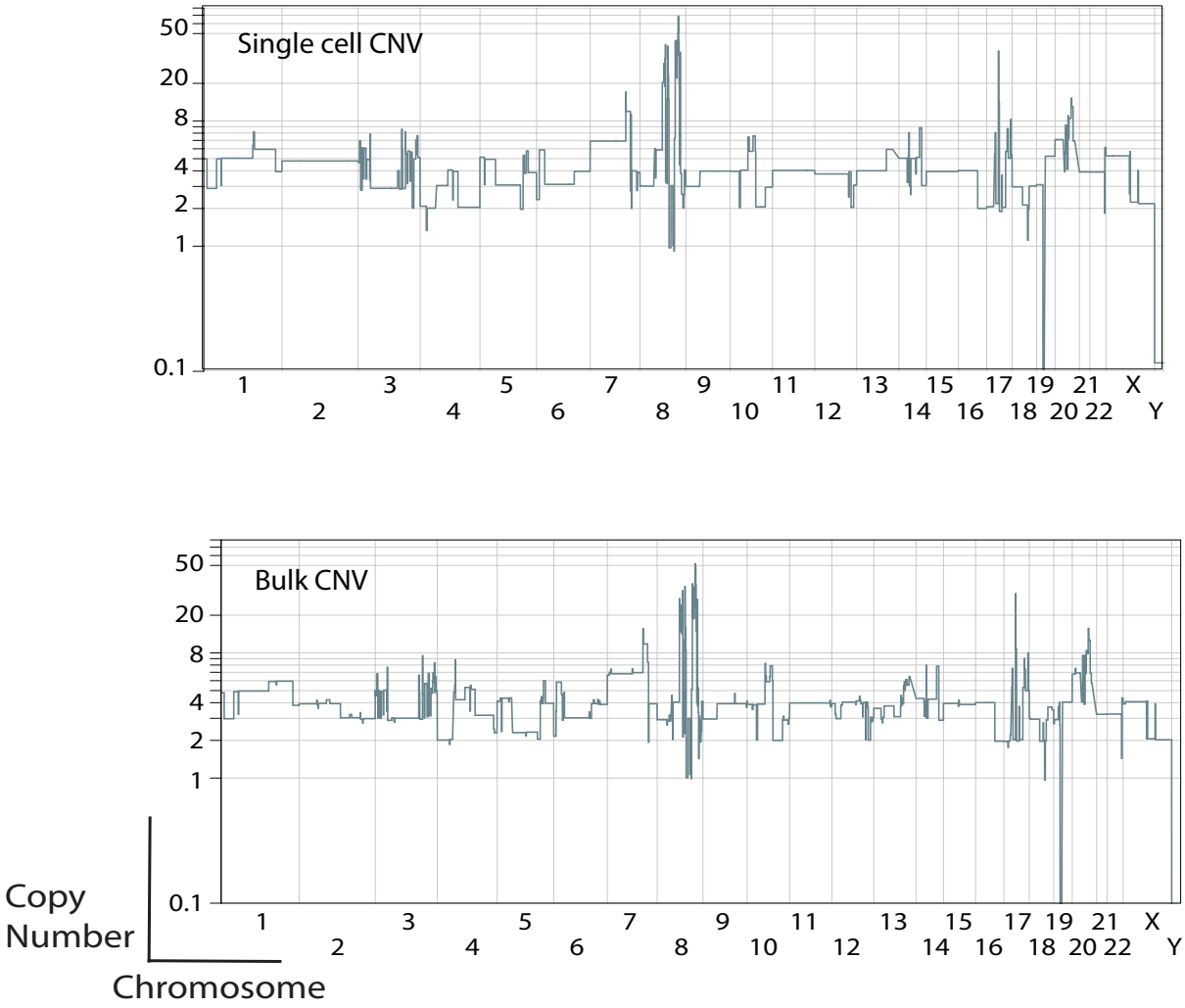


Figure 17: <u>C-DOP-L provides accurate copy number determination of rearranged cancer genomes from the SK-BR-3 breast cancer cell line.</u>

Representative genome-wide single cell copy number profile produced using the C-DOP-L approach compared to bulk profile produce via the sequencing of million cell DNA. Pearson correlation of the copy number values across the genome between single cell and bulk profile (R^2) = 0.963.

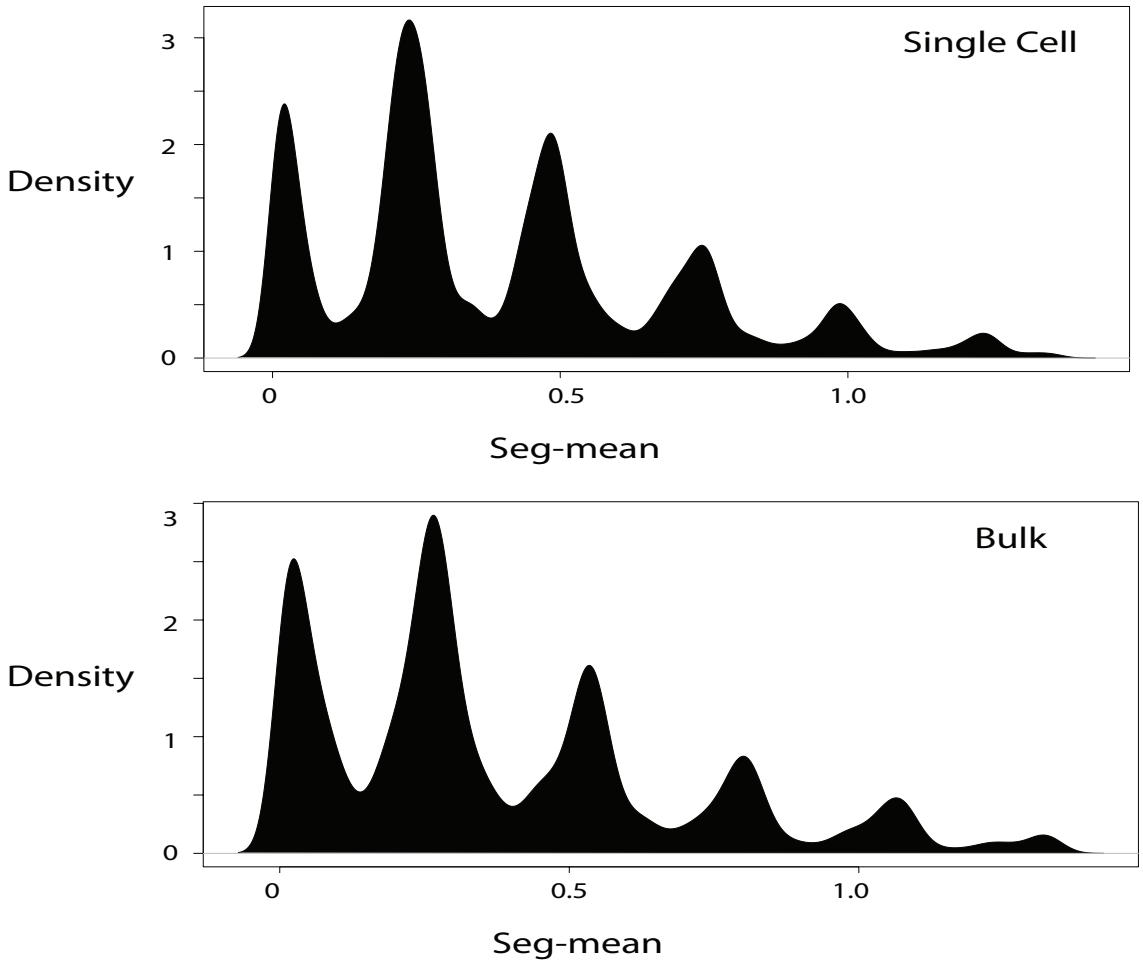


Figure 18: **<u>Single cell copy number data from C-DOP-L display quantal behavior.</u>**

Smoothened kernel density plots of the segmented read count data are plotted to display the discrete densities corresponding to integer copy number values. Density plots are shown for a representative single cell profile as well as the bulk profile.

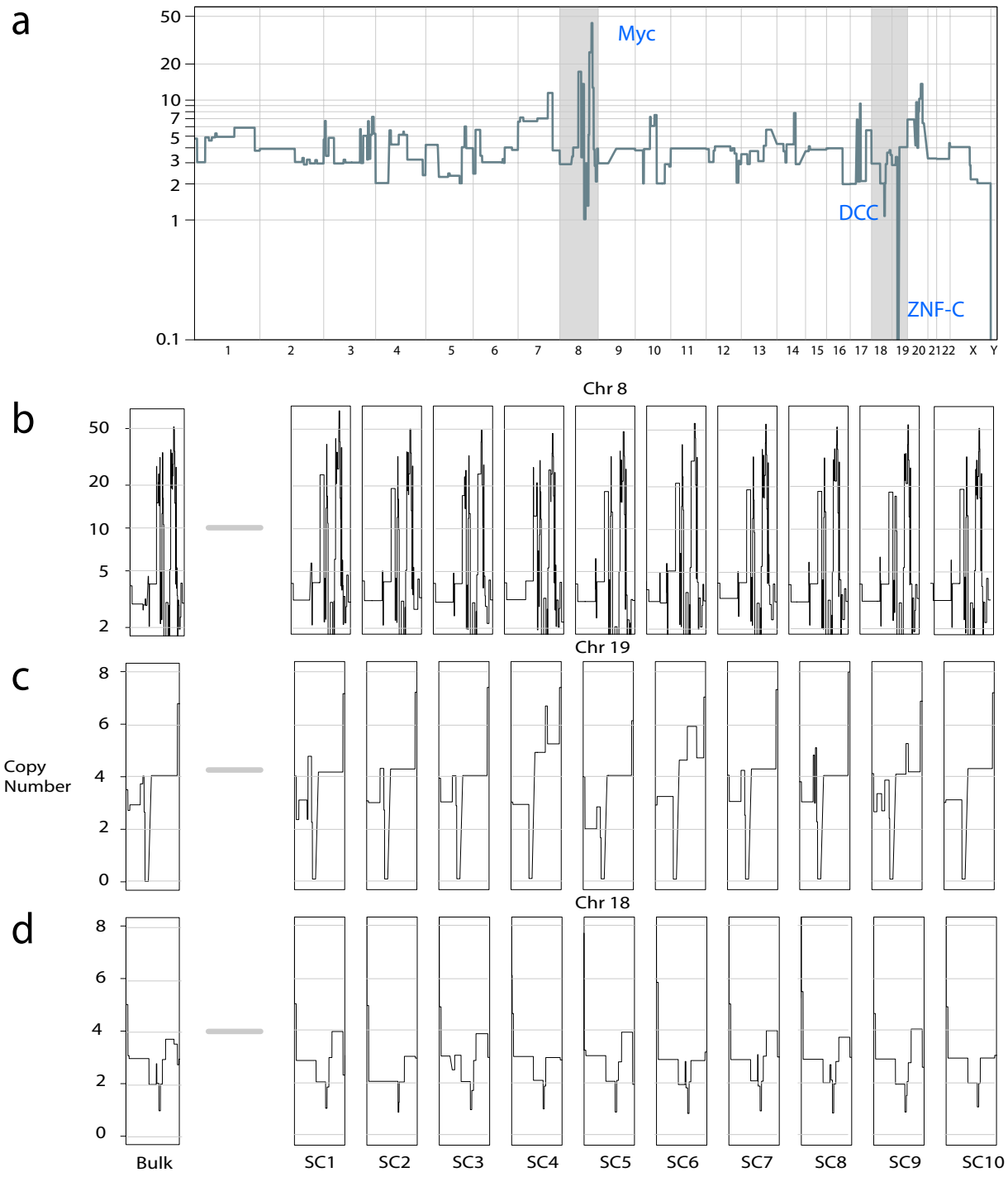


Figure 19: <u>C-DOP-L displays robust sensetivity in detection of copy number alterations.</u>

(a) Representative view of bulk SK-BR-3 copy number profile. (b) Snap-shots of copy number variants (b-amplification , c-homozygous deletions, and d-heterozygous deletions) across bulk and 10 representative single cells illustrating the detection of the variants in 100% of single cells sequenced. SC; single cell. Chr; Chromosome.

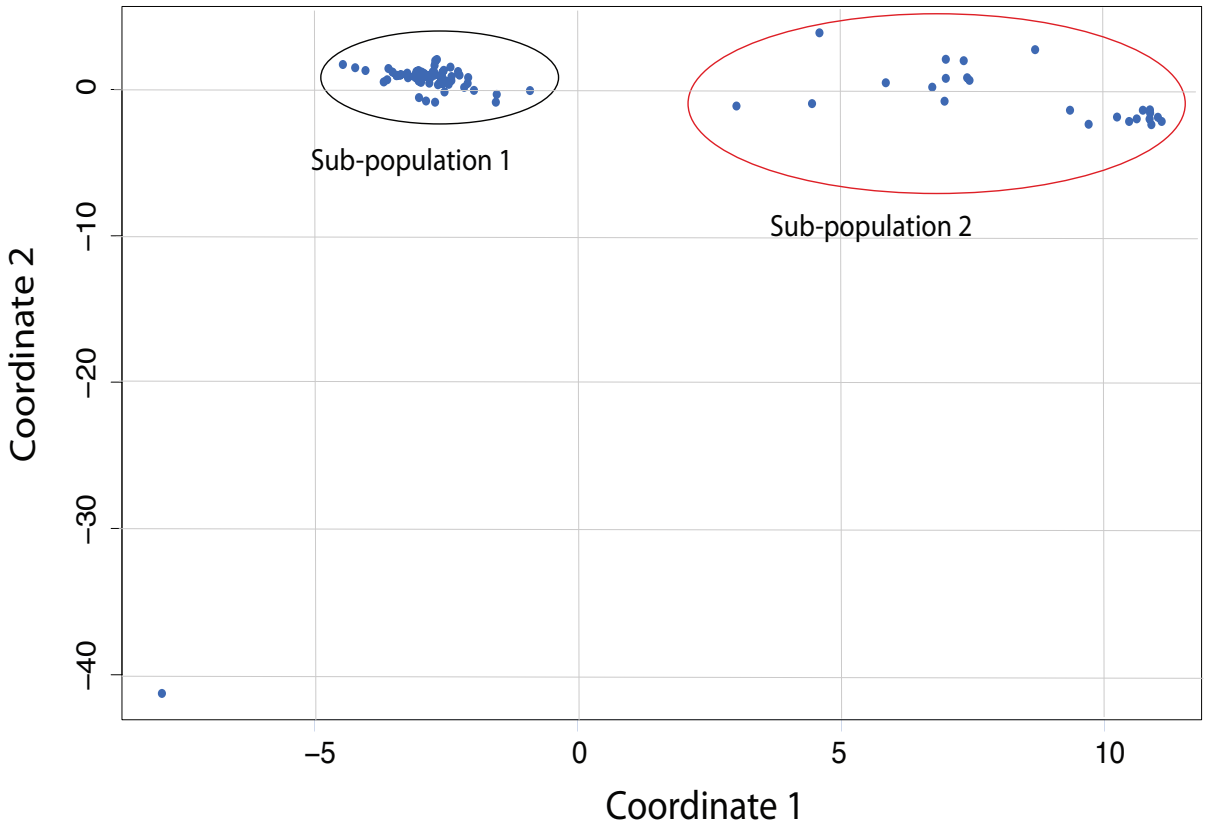


Figure 20: <u>High level multiplexing of single cells from the SK-BR-3 breast cancer cell line identifies distinct sub-populations.</u>

Sequenced SK-BR-3 genomes were plotted based on multi-dimensional scalling of their genome-wide copy number profiles. Ellipses denotes the two sub-clonal populations identified (Sub-population 1 and Sub-population 2).

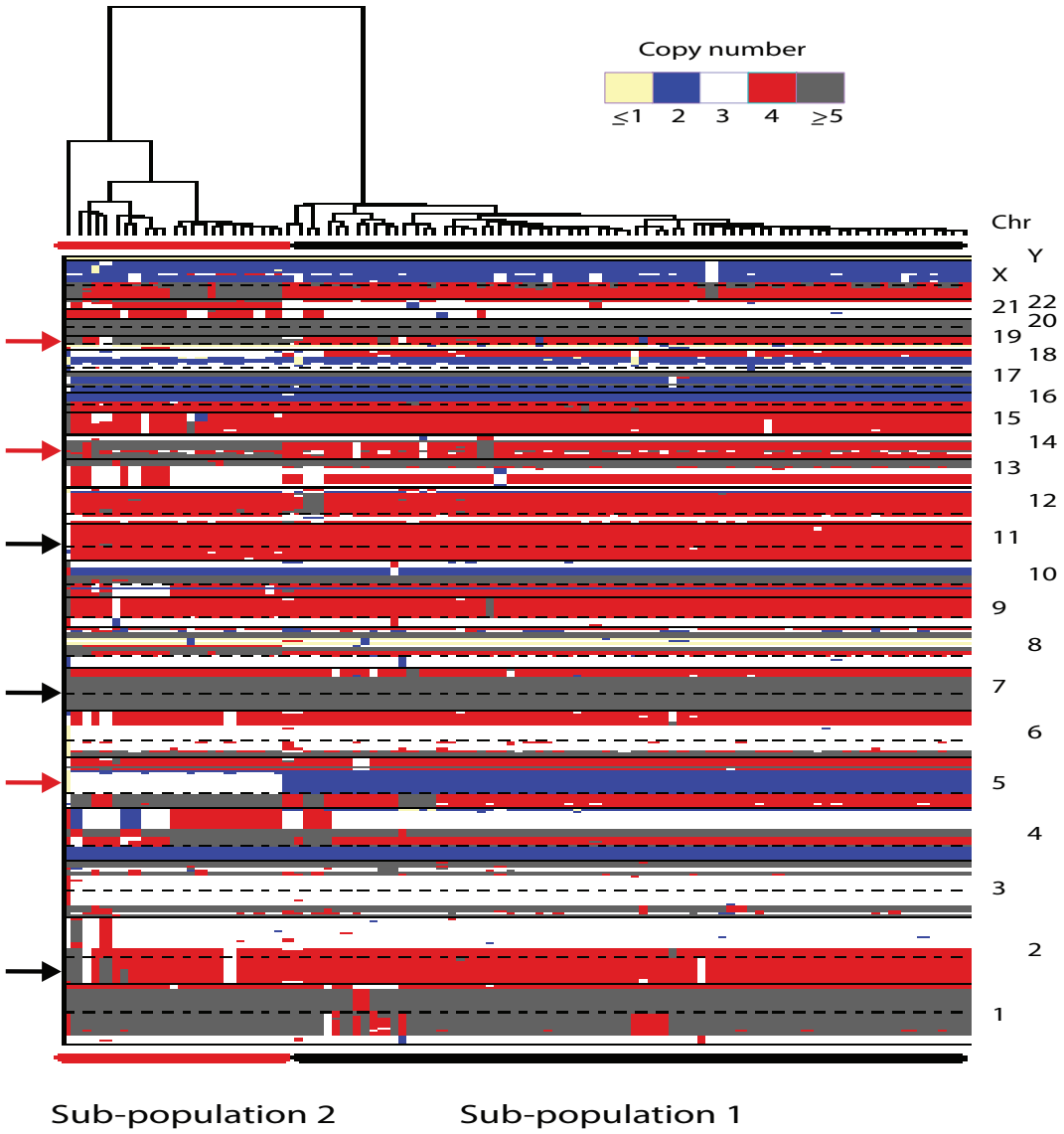


Figure 21: <u>The 2 SK-BR-3 sub-populations are derived from the same lineage, share the vast majority of copy number alterations, and differ significantly.</u>

SK-BR-3 single cell genomes were clustered using manhattan distance and Ward method in heatmap format. Black arrows point to represntative examples of copy number alterations shared between the vast majority of cells. Red arrows point to representative examples of copy number variants that distinguish each sub-population.

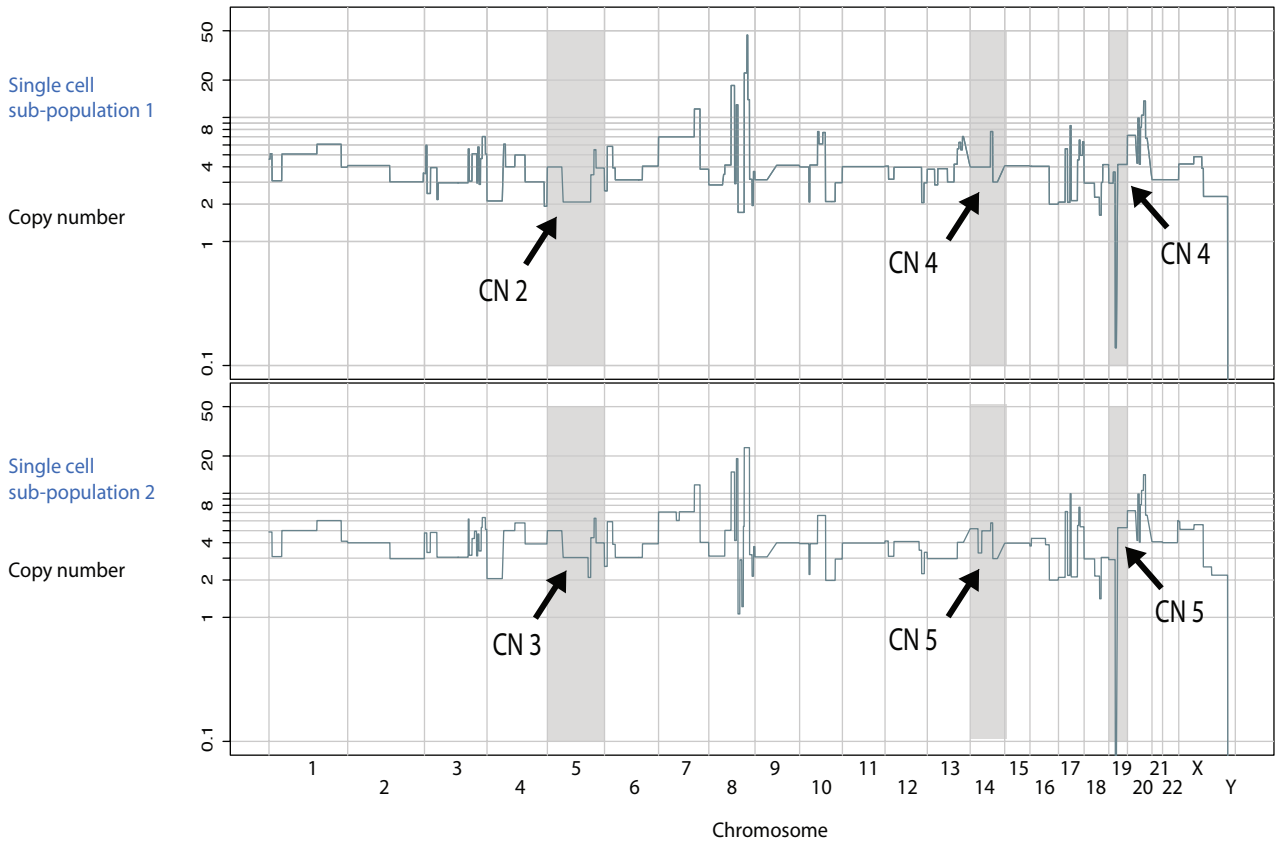


Figure 22: <u>Single cells from the 2 SK-BR-3 sub-populations differ significantly in copy number alterations.</u>

Representative single cell genomes from each SK-BR-3 sub-population are plotted genome-wide. Chromosomes where heterogenous copy number variants exit are highlighted in gray in background. Arrows point to specific regions on the chromosomes that are altered.

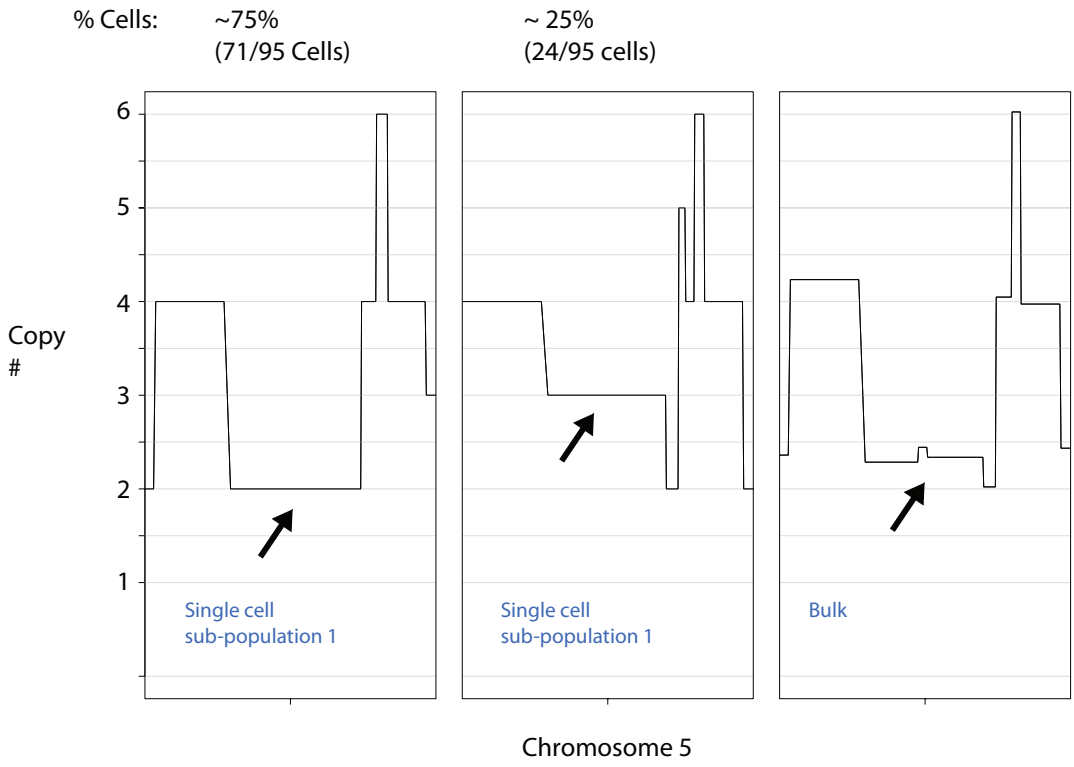
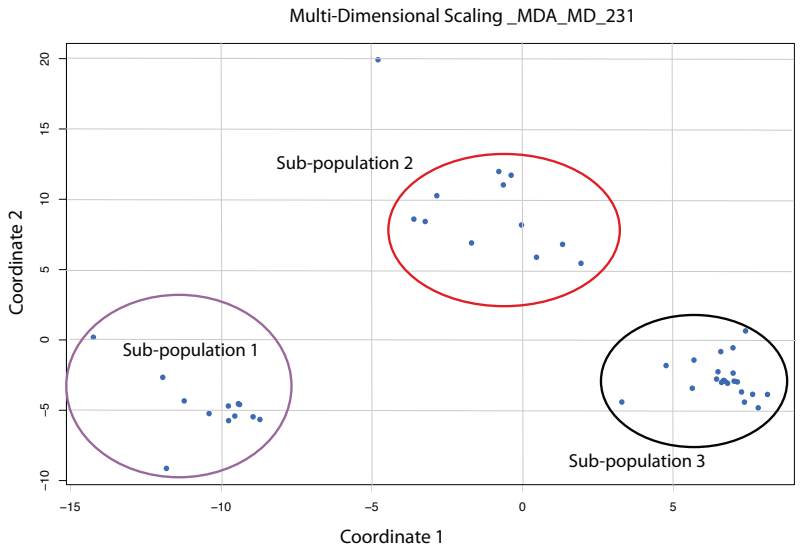


Figure 23: <u>Signatures of the sub-clonality of copy number variants identified via single cell sequencing are observed in some , but not all cases. in the bulk profile.</u>

Chromosome 5 view of copy number variation from representative single cells of the two SK-BR-3 sub-populations and the bulk profile is plotted. Signature of the sub-clonal variation is evident in the bulk population and reflects the frequency of each sub-population.

a


Multi-Dimensional Scaling _MDA_MD_231
Sub-population 2
Sub-population 1
Sub-population 3
Coordinate 2
Coordinate 1
20
15
10
5
0
−5
−10
−15
−10
−5
0
5

b

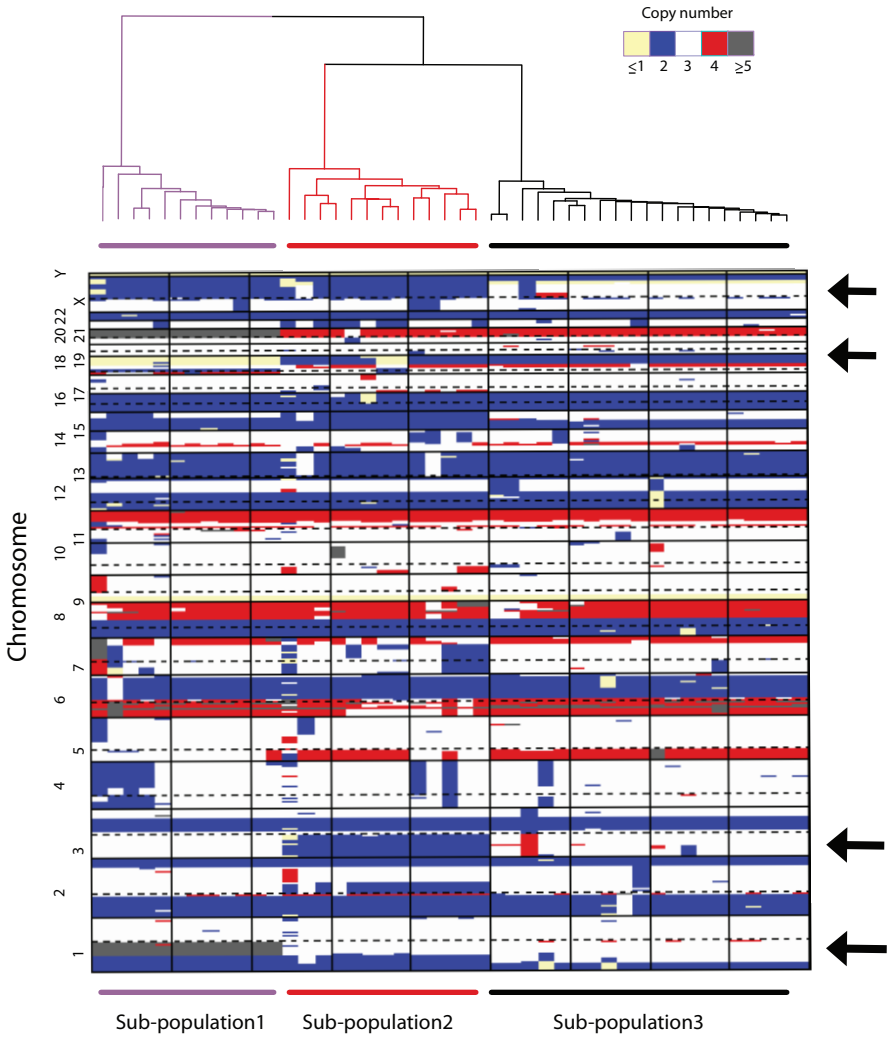

Copy number
≤1 2 3 4 ≥5
Chromosome
Y X 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
Sub-population1
Sub-population2
Sub-population3

Figure 24: <u>Sub-clonal heterogeneity is also observed in the MDA-MB-231 breast cancer cell line.</u> (a) Multi-dimensional scaling of 45 single MDA-MB-231 genomes based on genome-wide copy is plotted. Ellipses denote the 3 distinct sub-populatios. (b) Hierarchal clustering heatmap tree illustrates genomic regions that display heterogeneity between the different sub-populations (black arrows).

6. CHAPTER 4: Application of C-DOP-L to clinical breast cancer biopsies.

Work done in collaboration with Hilary Cox, Linda Rodger, Sean D'Italia, Mao Yong, Hannah Gilmore, Guoli Sun, Kristy Miskimen, Anthony Leota, and Jude Kendall.

6.1 Highly multiplex single-cell sequencing of clinical breast cancer tissue reveals genetic heterogeneity and sub-clonal populations.

To determine the feasibility of high level multiplexing for actual clinical samples, two estrogen receptor (ER) positive breast cancer cases (Pt31 and Pt41) were analyzed. Both were determined to be diploid in DNA content, with similar histopathology, and from the same gene expression sub-type (luminal B) as determined by RNA sequencing and PAM50 analysis (**Figure 25 and 26**). Bulk copy number analysis revealed characteristic ER positive copy number alterations, such as gains of chromosome 1q and 8q and deletion of chromosome 11q [32,48] (**Figure 27**) in both cases. To allow comparison with a previous single cell CNV approach (WGA4 amplification (Sigma-Aldrich) followed by Standard Illumina library prep), core needle biopsies from both cases (8mm in length) were cut evenly into two sections for processing using WGA4 and C-DOP-L (**Figure 28**). For each section, 96 nuclei were sorted and the plates were processed with either WGA4 or C-DOP-L. Each 96 multiplexed pool was sequenced on a single lane of Illumina HiSeq instrument. Cells yielding at least 0.25 million uniquely mapped reads were considered successful for the complete process. Compared with the cell lines the clinical samples were somewhat more variable. The number of successfully profiled cells

for Pt41 was 86/96 using WGA4 and 89/96 using C-DOP-L, while Pt31 yielded 88/96 and 69/96 respectively.

Single tumor cells from both cases where then plotted and clustered in a heatmap format based on their genome-wide copy number profile (**Figure 29**). Cells with normal profiles were omitted from the figure and CORE[99] (Cores of Recurrent Events) (see methods) was used to select the cancer cells that are part of a clonal lineage. This way tumor cellularity for each biopsy was approximated (~60% tumor for Pt31 and ~90% tumor for Pt41). Chromosome 1q and 8q duplications as well as loss of 11q were found in virtually all single cells from both tumors using both approaches, consistent with these events occurring very early in the evolution of the tumor genome and further attesting to the sensitivity and specificity of the approach (**Figure 29, black arrows**). Interestingly, whereas Pt41 tumor profile contained more copy number alterations than Pt31 (measured as % of genome altered), single cell copy number profiles from Pt41 displayed homogeneity, with almost all cells sharing all chromosomal alterations. By contrast, Pt31 had 3 sub-populations that differed in their copy number status at multiple chromosomes, for example chromosomes 5, 7, 11, and 13 (**Figure 29, red arrows**). These populations were also found to differ in proportion between the two adjacent sections. Phylogenetic analysis of the sub-populations based on their genomic alterations revealed that the two divergent populations, 2 and 3, arose from the earlier ancestral population 1 via the acquisition of additional genomic alterations (**Figure 30**), yet, interestingly, population 1 also persisted.

6.2 Highly multiplex single-cell sequencing of clinical breast cancer tissue reveals mosaicism in chromosomal amplifications.

Upon further examination of the single cell copy number profiles of the tumor (Pt31), additional heterogeneity in the form of mosaic copy number amplifications was noted (**Figure 31**). Some occurred at genes with established clinical significance in breast cancer such as the amplification of Cyclin D1[100] (CCND1) on chromosome 11 and TOP2A[101] on chromosome 17, while others occurred at genes for which experimental evidence exists for involvement in cancer, such as the homeobox protein SIX6[102] on chromosome 14 and PREX1[103] on chromosome 20. Together, these data provide strong evidence of the power of highly multiplex single cell sequencing in resolving sub-clonal structure and illustrating genomic heterogeneity present within the genomes of human tumors.

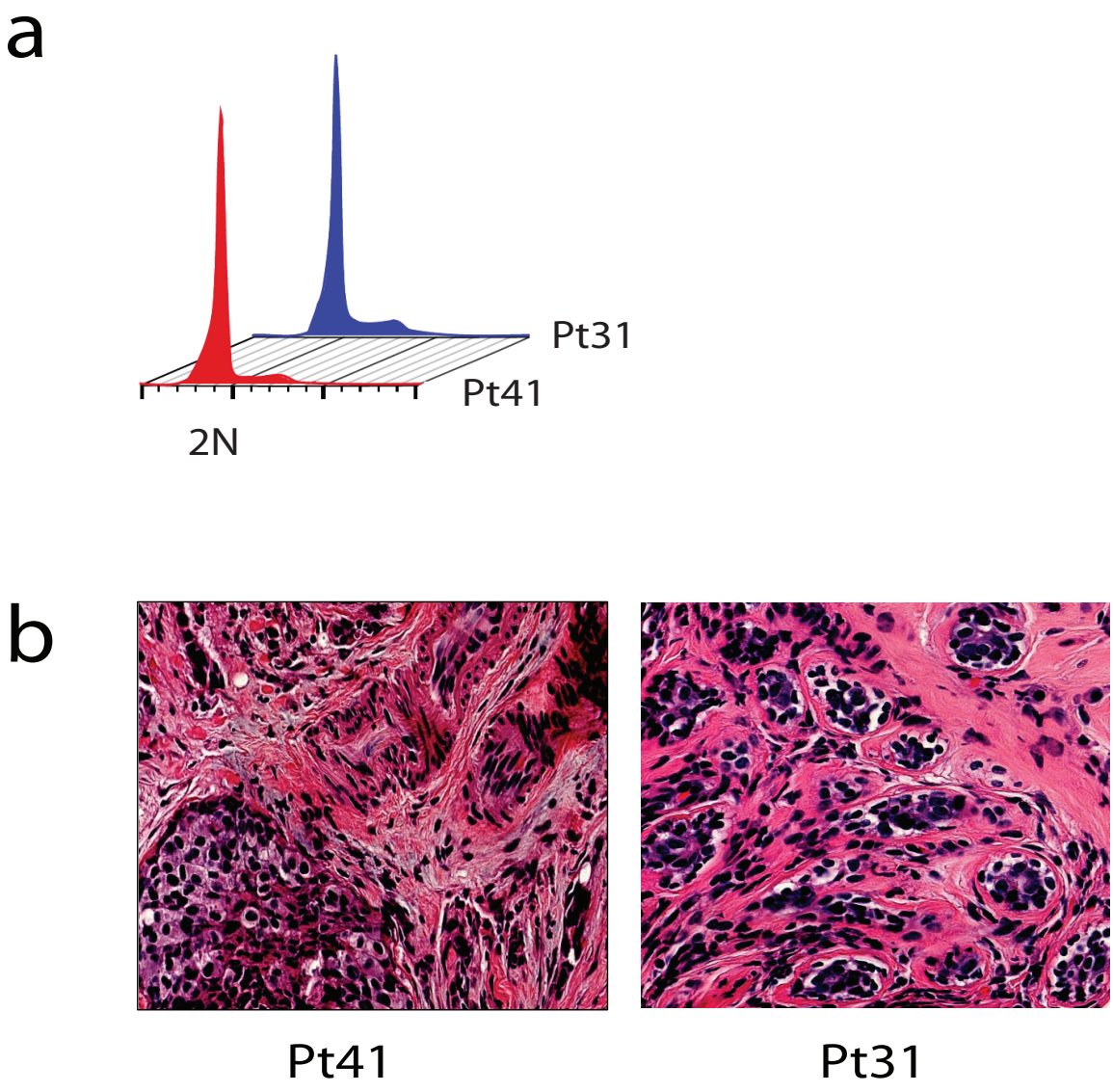


Figure 25: <u>Pt31 and Pt41 are similar in terms of DNA content and histopathology.</u>

(a) DNA content was determined to be diploid for both cases based on flow cytometry by DAPI staining. (b) Both tumors were judged to be similar histopathological based on hematoxylin and eosin (H&E) staining with both showing invasive ductal carcinoma with moderate differentiation and complex glandular growth pattern.

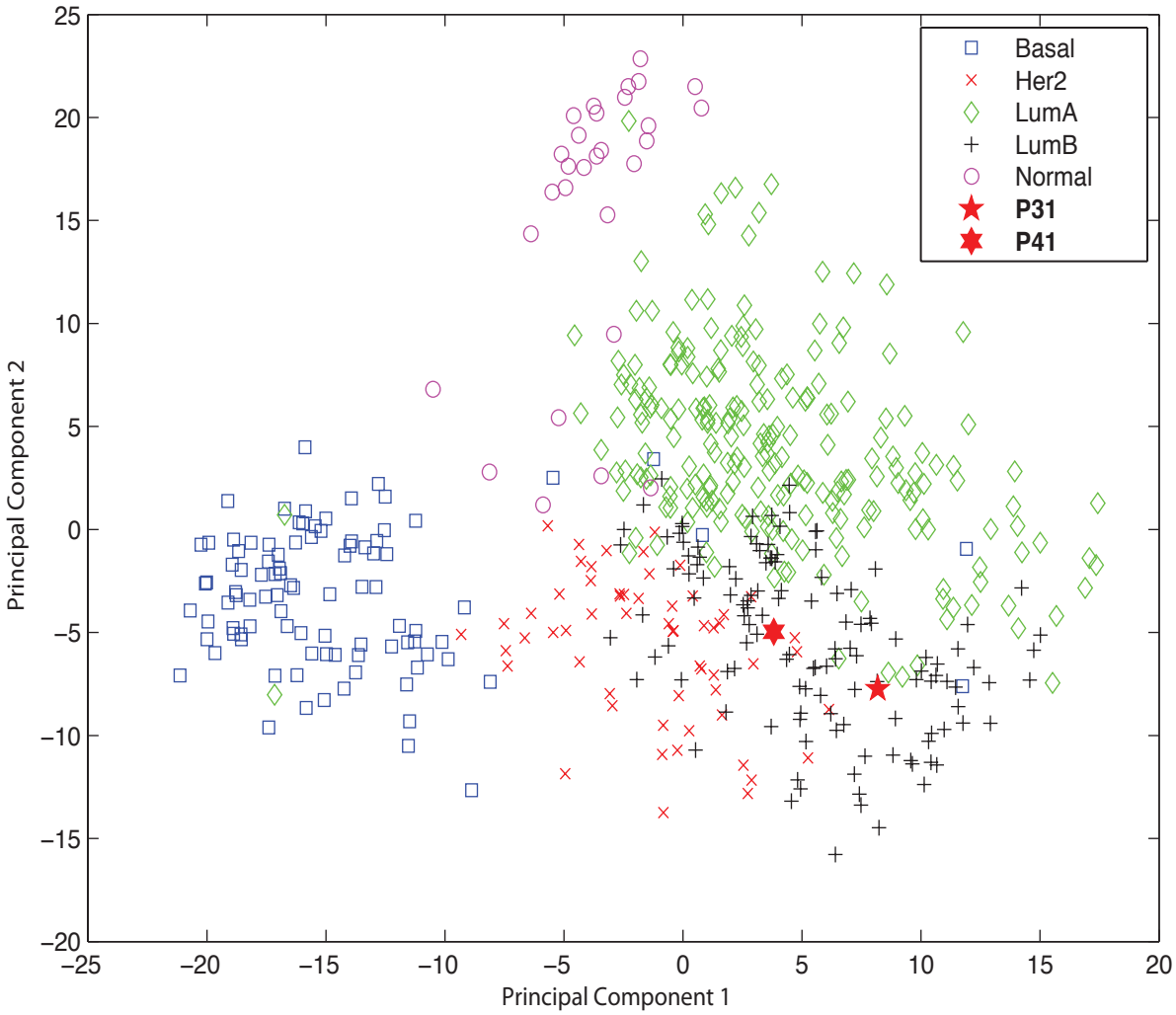


Figure 26: <u>Pt31 and Pt41 belong to the Luminal B breast cancer gene expression subtype.</u>

542 TCGA samples were projected on the PCA plot with the top 2 principal components and given subtype information obtained from TCGA data sets (see methods). It included 96, 58, 231, 128 and 29 samples respectively for Basal, Her2, LumA, LumB and Normal-like subtypes. P31 and P41 were projected on the background within the LumB cluster marked as a pentagram and a hexagon.

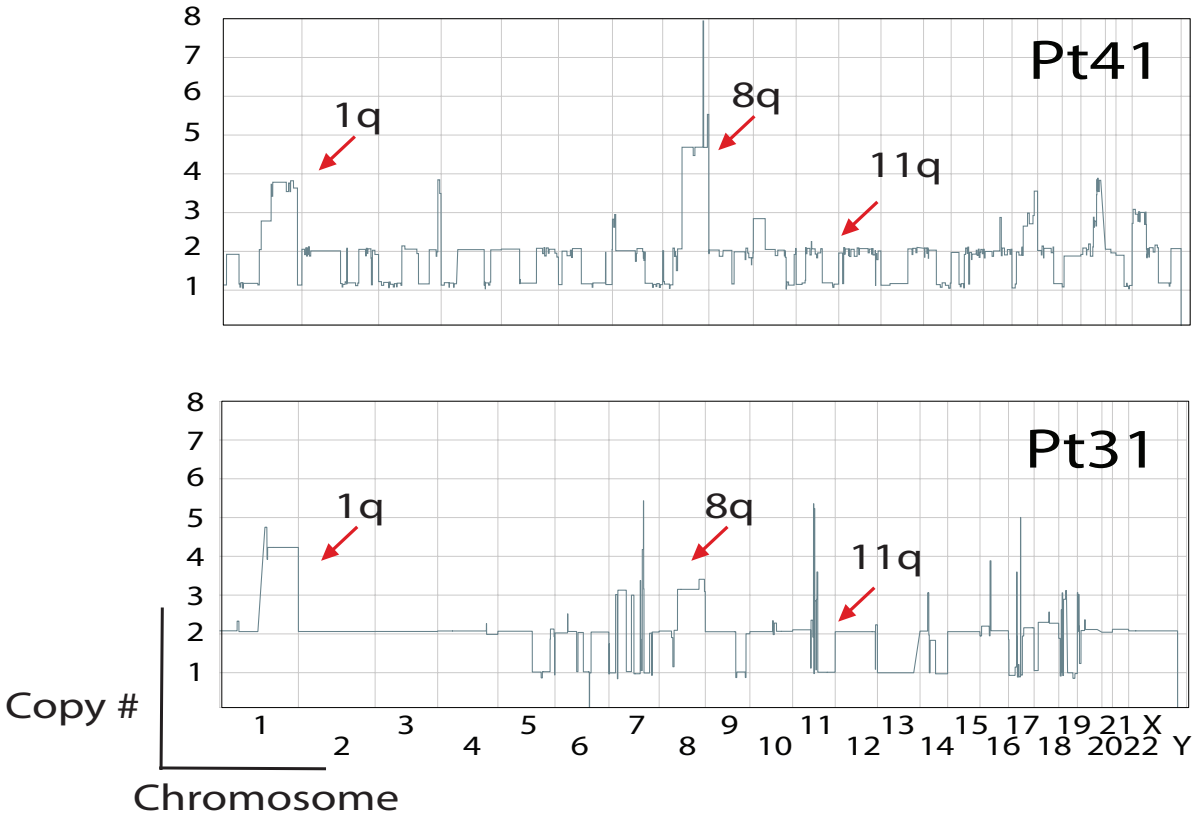


Figure 27: <u>Bulk copy number profiles of Pt31 and Pt41 reveal genetic alterations characteristic of ER positive breast cancer disease.</u>

Copy number alterations of bulk Pt31 and Pt41 were plotted genome-wide. Red arrows point to genetic alterations that have been found to be characteristic for ER positive breast cancer disease.

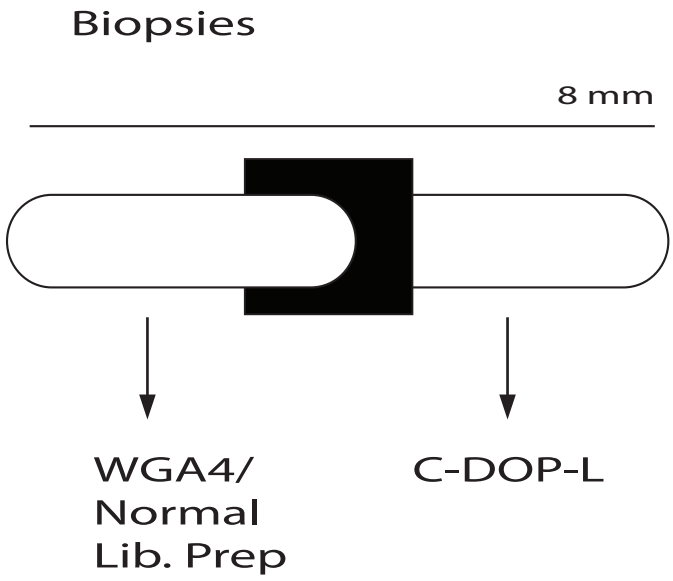


Figure 28: <u>Schema of biopsy dissection and single cell processing.</u>

In brief, each tumor biopsy, measuring 8 mm in diameter, was cut evenly into two sections for processing using the C-DOP-L approach as well as WGA4 (previous method).

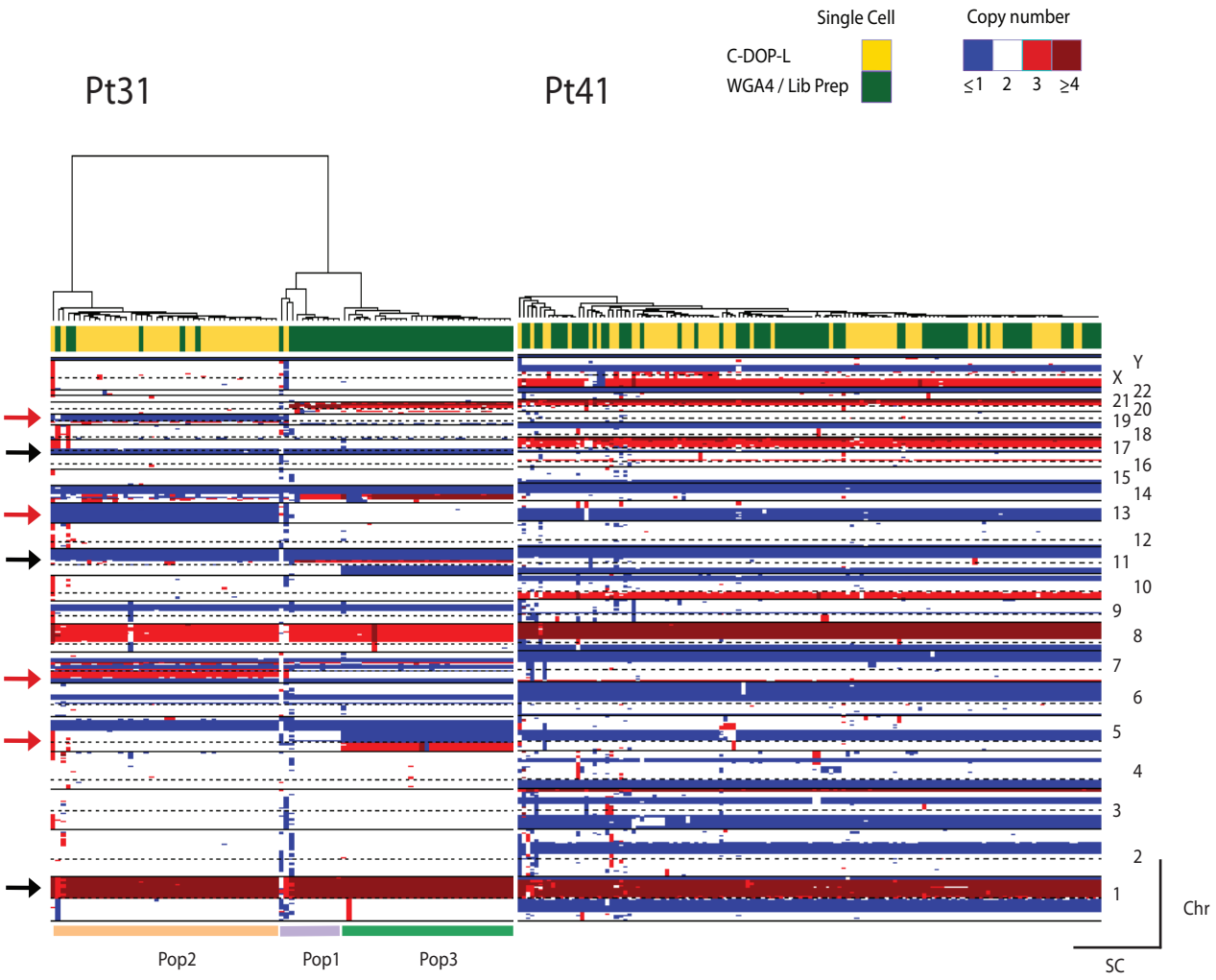


Figure 29: <u>Hierarchal clustering heatmap of the clinical cases profiled using single cell sequencing methods reveals genetic heterogeneity and sub-clonal populations.</u>

Single cell genomes from Pt31 and Pt41 were clustered using manhattan distance function and according to the Ward method. Black arrows point to copy number alterations characteristic of ER positive breast cancers and found in nearly all single cells. Red arrows point to copy number alterations that are found to be sub-clonal in Pt31 single cells.

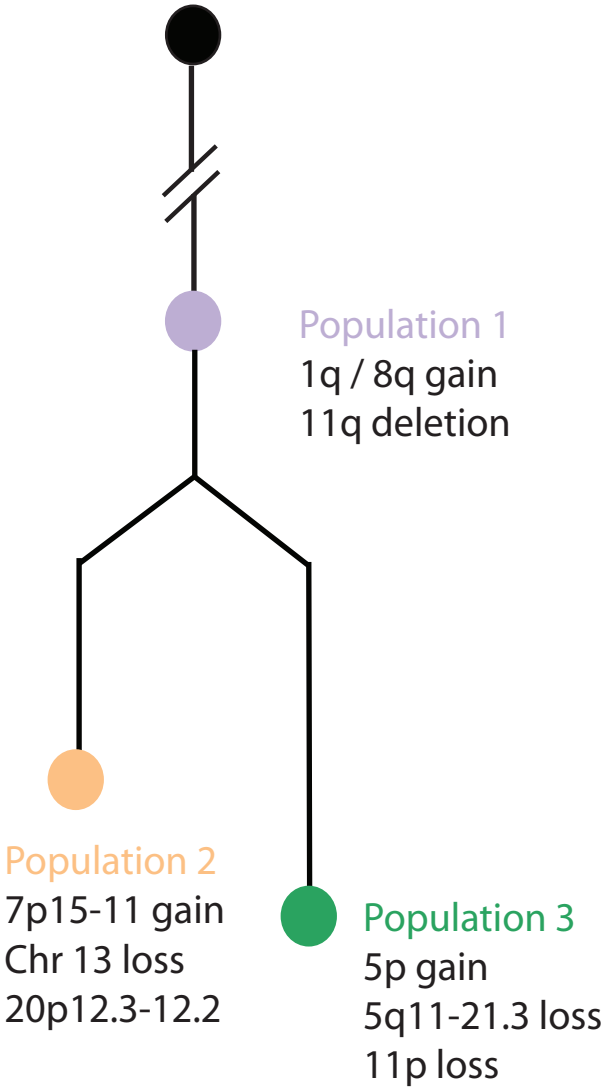


Figure 30: <u>Schematic representation of the phylogenetic tree of Pt31 sub-populations.</u>

Pt31 sub-populations were analyzed according to their sub-clonal alterations to reconstruct the phylogenetic evolutionary tree. Colors denoting each sub-population are as in Figue 28. Sub-clonal genetic alterations for Populations 2 and 3 and indicated.

7. DISCUSSION

7.1. C-DOP-L offers an affordable, robust, and high-throughput platform for highly multiplex genome-wide single cell copy number profiling

The potential of single cell genome profiling in understanding cancer heterogeneity lies in the ability to profile hundreds and even thousands of single cell genomes. The C-DOP-L approach extends on the burgeoning field of single cell genomics by offering a robust high-throughput method to examine the genome-wide copy number profile of hundreds of single cancer cells. The down-sampling simulation analysis facilitated the benchmarking of the minimal data requirements necessary to reproduce genome-wide copy number variation of cancer cells and guided the multiplexing strategy. By coupling the restriction enzyme digestion of the WGA DNA universal sequences with NN-mediated adaptor ligation the approach allows for (1) maximizing the amount of information extracted from each sequencing read via the elimination of the WGA universal sequences, (2) enhancing the quality of the sequencing output, and (3) significant reduction in the cost and effort required to generate highly multiplexed single cell sequencing libraries. In previous reports[80], single cell was sequenced on a single lane of the Illumina platform at a cost of approximately $1,000. Using the methods described here, with the multiplexing of 96 single cells on a single HiSeq lane, the cost of sequencing a single cell is reduced to approximately $30 per cell in reagents and sequencing costs. Undoubtedly, with the decreasing cost per base from NGS, this figure is likely to drop even further and facilitate the profiling of thousands of

single cells in a single lane. At that stage, microfluidics will be needed to reduce preparation costs and reduce manual labor[104,105]. In addition, C-DOP-L can easily accommodate different multiplexing platforms such as the Illumina third read TruSeq indexing system.

7.2. Limitation of the C-DOP-L method and possible solution

While the approach focuses on robustly identifying an important class of somatic mutations in copy number variants it does not focus on the identification of other sources of somatic mutations such as single nucleotides variants (SNVs) and structural variants. However, it is important to point out that with current sequencing output of NGS platforms it is still prohibitively expensive to sequence the exomes of hundreds of single cells. Furthermore, even though DOP-PCR does not cover the entire genome when sequenced at high depths, there is evidence to suggest that up to a third of the whole genome can be covered in a single cell WGA product[84] and the data indicate that genome coverage increases with more single cells sequenced (**Figure 32**). Thus an approach based on initially resolving clonal population structure via genome-wide copy number variation followed by pooling of single cell libraries and targeted capture of particular sub-populations (for example the 3 subpopulations in Pt31) may provide exome-wide views of these sub-populations. A similar strategy has recently proven effective in illustrating the clonal architecture of secondary acute myeloid leukemia[106].

7.3. Emerging Next Generation Sequencing technologies and their potential impact on single cell sequencing

Given the rapidly changing landscape of Next Generation sequencing technologies, the approaches for single cell profiling presented here are very likely to change radically as technologies evolve. Two of the most obvious changes could involve (1) a significant increase in the numbers of cells that can be multiplexed using the 'short-read' (Illumina) technology; and (2) the continued advancement of 'long read' sequencing using the Pacific Biosciences RSII Instrument or the Oxford Nanopore technology[107].

To date, most single cell studies have utilized short read sequencing obtained using the highly parallel 'short read' method dominated by the Illumina technology platforms. Illumina sequencing relies on mapping millions of short reads (approximately 100 nucleotides in length) to yield an extremely high total output of sequencing data. This method is ideal for copy number determination since CNV states are determined via molecular counting of sequenced DNA molecules in pre-defined genome segments or 'bins'[80]. Increasing the numbers of cells to be sequenced is, to a large extent, a matter of how many independent DNA fragments or 'reads' that can be extracted from one lane of a sequencing chip. The capacity to perform multiplexing of thousands of single cell genomes is already theoretically possible given that the newest version of the Illumina HiSeq instruments (HiSeq X) provides a total yield of 3 billion sequencing reads per lane. Thus, using 5K binning for copy number determination, up to 12,000 single cell genomes can be multiplexed on a single

HiSeq X lane. At that stage the limiting factor is no longer the actual sequencing capacity, but the ability to process individual cells. It is clear that significant breakthroughs in microfluidic technology will be needed to reduce preparation costs and manual labor. Additionally, to realize that level of multiplexing, the barcoding strategies implemented here are also likely to require modification. In the work presented here, each single cell was uniquely barcoded using a specifically modified Illumina adaptor oligonucleotide. Thus, 96 different oligonucleotides were synthesized and purchased separately. It is clear that one cannot utilize this approach for the multiplexing of thousands of single cells (i.e. design and synthesis of 1000 barcoded Illumina adaptors would become a bottleneck both in cost and time). As an alternative, it would be feasible to implement a combinatorial multiplexing strategy. Such a system already exists in the form of TruSeq indexing, available from Illumina.. Illumina sequencing is generally carried out through two sets of synthetic sequencing reactions (the 'first' and 'second' reads) that are initiated, one at each end of the target DNA molecule, from adaptor sequences added during library preparation. The TruSeq technology embeds separate short sequences of 8 base pairs coded in the adaptors (the "third" and "fourth" sequencing reads). The system uses 8 (or more) adaptors carrying different 'third' read sequences and 12 (or more) carrying different 'fourth' read sequences. When combined in a matrix, that yields a minimum of 96 distinguishable barcodes combining the 'third' and 'fourth' reads. Although the TruSeq matrix method as currently implemented uses only enough barcodes to create 96 separate barcodes, we have shown, using the work presented here, that it is possible to create at least 96 barcodes from each embedded 8 basepair coding read. Thus, barcoded Illumina adaptors can be designed to provide a unique set of 96 indexes via the "third"

read as well as a separate unique set of further 96 indexes via the "fourth" read. In this fashion, matching the unique barcodes of both reads (third and fourth) would result in 9216 barcodes (i.e. $96^2 = 9216$). This approach can easily be accommodated using the C-DOP-L method via further modification of the NN Illumina adaptors described in the materials and methods section.

Another direction that may impact future single cell sequencing approaches is the evolution of sequencing methods that achieve much longer read lengths than possible with the Illumina methodology. Longer read lengths, up to tens of thousands of base pairs are highly desirable for genome mapping and assembly of complete genomes from contiguous sequence, so-called 'contigs'. Such long read lengths and potentially lower cost per base pair of actual sequence, are the characteristic property of 'single molecule' sequencing via Pacific Biosciences RSII (PacBio) and Oxford Nanopore technologies. By providing sequencing reads up to fifty kilobases in length, it is conceivable to imagine that information regarding single nucleotide polymorphisms as well as structural variants can be retrieved in a high-throughput manner for single cell genomes. For that to be realized however, single molecule sequencing platforms would have to provide higher output of sequencing data. For example, currently, a single run on the PacBio instrument yields orders of magnitude less sequencing data than a HiSeq Illumina sequencing run.

Nonetheless, it is possible to contrive innovative molecular approaches that maximize the utility of longer read lengths even with the current low sequence read output afforded by the current single molecule platforms. One conceivable strategy would be to adopt an approach similar to the SAGE (Serial Analysis of Gene Expression) method for mRNA profiling[108]. The key element of SAGE is that thousands of very short

segments of the actual mRNA or cDNA (called 'tags') are ligated together and resulting concatamer cloned into a recombinant vector for amplification as DNA in bacteria or bacteriophage. This cloned DNA can then be sequenced and the individual tags identified from the bulk sequence by bioinformatic algorithms, yielding a snapshot of the mRNA profile. For copy number profiling, it would similarly be possible to create very short individual DNA molecules by restriction enzyme digestion (the equivalent of tags in the SAGE approach), ligate them to one another to yield much larger DNA concatamers (i.e. ligating molecules of 30 base pair nucleotide length to yield a 10 kilobase DNA molecule) that may be sequenced directly through one of the single molecule sequencing methods (most probably the Oxford Nanopore technology, since fewer actual molecules are required for this method). The large DNA molecules are then informatically de-convoluted to give a list of the shorter segments that are subsequently counted in genomic bins to yield copy number estimates. For example, given that any 30 base pair nucleotide sequence is sufficiently unique across the entire human genome, then sequencing a 10 kilobase DNA molecule constructed from the ligation of much smaller 30 nucleotide DNA molecules would effectively yield approximately 333 DNA molecules instead of one (10,000/30).

Although such methodology is feasible to carry out even with the current long-read instruments, it is limited by the significantly higher error rates of the 'long read' technologies. The 'long read' methods depend on the extreme length of the continuous reads for accurate mapping and the accumulation of multiple redundant reads covering the same sequence for accurate calling of individual bases. Thus, for such a SAGE-like

method to actually yield useful copy number data, the base calling accuracy of the single molecule sequencing platforms will have to be improved significantly.

7.4. Biological insights gleaned via highly multiplex single cell sequencing

The robustness of the C-DOP-L approach has allowed for the profiling of hundreds of single cell genomes from cancer cell lines and human tissue. Even within the results of the example studies presented here biological insights with important implications for tumor biology were deduced. First, the observation of sub-clonal variation in human cancer cell lines, generally presumed to be monoclonal, implies that the evolutionary process that underlies cancer development is still operative in cell culture. Second, the sub-clonal heterogeneity for genome-wide copy number in culture raises the question of how divergent a cancer cell line might be across different laboratories and how to compare different studies utilizing what are supposed to be the same cells, but which might differ significantly[109]. Third, the striking differences in sub-clonal heterogeneity between the two clinical samples, Pt41 and Pt31 is intriguing, given that Pt41 genome is more highly rearranged than Pt31. This might suggest that factors other than genomic instability might modulate intra-tumoral heterogeneity and/or diversification is dynamic throughout the history of a tumor. And fourth, the mosaicism of genomic amplifications observed in Pt31 highlights the remarkable heterogeneity that can be obtained by cancer genomes and presents the question of how these varied alterations might modulate responses in the face of selective pressures such as therapeutic intervention.

The insights gleaned via the single cell analysis also raise important biological and genetic questions, particularly with regards to sub-population identification. In the cases presented, the resolution of the analysis (i.e. sequencing approximately one hundred cells per case) generally yielded two to three sub-populations found at a minimal frequency of 10% (i.e. a particular genomic pattern is observed in 10 out of the 100 single cells sequenced). While it is almost certain that single cell analysis of a cohort of patients will provide a more dynamic view of cancer cell sub-populations in different tumors, a question relating to the relative depth, measured in the number of single cells sequenced, can be posed: How many single cells are needed to accurately ascertain the number of sub-populations present within a tumor? For example, assuming a thousand single cells were sequenced, what cutoff point marks a transition from sub-populations to individual single cells? Could clonal sub-populations exist at relative frequencies of 1%, 0.1% or lower? The answer to this question would yield valuable insights into the genetics of tumor evolution and help address clinically relevant phenomena such as late recurrences, therapeutic resistance and existence/nature of cancer stem cells, thought to contribute in some cases to therapeutic failure/resistance[110].

Additionally, the observation of the underlying genetic heterogeneity implies that there also exists phenotypic heterogeneity in the cancer sub-populations. The resolution afforded by single cell sequencing thus could allow direct associations between phenotype and genotype. Indeed, this has recently been explored via the analysis of circulating tumor cells (CTCs) during therapeutic intervention[88]. In the work, single cells from prostate cancer patients undergoing therapy were selected using an imaging modality based on the expression of protein markers (i.e. phenotype). Circulating tumor

cells were stained and selected for the expression of the androgen receptor (AR) at multiple time points following treatment with abiraterone, an inhibitor of androgen synthesis. Single CTCs where subsequently analyzed using methods similar to the ones described here. Interestingly a subset of single cells was found to have gain a unique set of somatic alterations, including MYC amplification, an event experimentally linked with hormone therapy resistance. Thus, in this case, the monitoring of circulating tumor cells during treatment facilitated the identification of genomic markers that could be of potential utility. It is likely that further studies like the one described above will provide a compendium of genetic alterations that link resistance phenotypes to somatic copy number alterations.

7.5. Foreseeable clinical applications of single cell copy number profiling

The utilization of single cell copy number profiling in studying cancer experimentally, in settings such as cancer cell lines and tumor mouse models is likely to yield valuable insights into the dynamics of the evolutionary process. But, what of the utility of single cell analysis in the clinical setting? While realizing any applicability of single cells in the clinical management of patients will require more rigorous validation of the developed methodologies as well as the automation of these methods, it is still tempting to speculate about the translational opportunities that might exit.

As described in the previous section on biological implications, one area that holds a great deal of promise in the clinic is the genomic profiling of circulating tumor

cells. The ability to non-invasively capture and genetically profile single circulating tumor cells via a "fluid biopsy" to guide targeted therapeutic modalities is distinctly appealing. This is particularly true in the setting of metastatic disease where acquiring biopsies from different metastatic sites (for example bone metastasis in breast and prostate cancer) is often not feasible. In addition, the minimal invasiveness of this approach makes it possible to perform genetic analysis of single circulating tumor cells on serially collected samples in real time during therapeutic intervention. This has the potential to facilitate rapid response from the side of clinic to any insights that might be gleaned from the cancer genome copy number profile. For example, investigations of targeted therapy in lung adenocarcinomas have revealed pronounced sensitivity of EFGR mutant Non Small Cell Lung Cancer (NSCLC) to EGFR inhibitors such as gefitinib[111]. Unfortunately, in most cases, responses are not durable and treatment resistant clones emerge. Genomic investigations at the bulk analysis level of relapse cases have shown that focal amplification of the MET proto-oncogene confers resistance to EGFR mutant cancers[112]. These amplifications are robustly detected at the single cell level using the methods described here. In addition, there is evidence that circulating tumor cells are abundant in lung cancer patients. Thus, it is conceivable to imagine a setting where the detection of increasing numbers of circulating tumor cells containing a MET amplicon during targeted gefitinib treatment signals the emergence of a clonal population that is resistant to gefitinib treatment and informs the clinician of the possible value of adding a MET specific inhibitor to the therapeutic modality.

Furthermore, in the setting of primary disease, it is also possible to envision utilizing single cell analysis on biopsies serially collected in real time during the course

of disease treatment with the purpose of monitoring the shifts in the dynamics of the sub-populations identified via their genome-wide copy number patterns. In this case, the emergence/disappearance of identified sub-populations could be construed as evidence for resistance/sensitivity towards the therapeutic agent being used. This may inform the use of an additional therapeutic agent if a particular genetic biomarker for which a therapeutic agent is known was observed in the single cell sequencing data. While the analysis of single cell data during the course of treatment from repeated biopsies could be confounded by geographic heterogeneity, certain approaches could be utilized to address this confounding variable. For example, the observation of the emergence of a sub-clonal population following treatment in two distinct sets of biopsies acquired before and after treatment could be interpreted as sufficient proof of the resistance of this population to the therapeutic agent. Furthermore, given the resolution of data, single cell genome analysis can facilitate the utilization of minimally invasive approaches, such as Fine Needle Aspirates (FNA), that yield lower material in terms of numbers of cells than traditional biopsies. The advantage of the utility of FNA for single cell genomic interrogation lies in relative safety of sample retrieval coupled with the capacity to perform comprehensive exploration of the single cell genomes.

In addition, an often overlooked area where single cell analysis could potentially contribute is in predisposition risk assessment in individuals with germline mutations of unknown pathogenicity. Given the single cell resolution of the data and the fact that genetic instability is the driving force for cancer genome evolution, measurement of the level of genetic instability via genome-wide copy number profiling in seemingly normal cells could potentially provide a metric for pre-disposition to cancer. For example, the

observation of random non-recurrent, genomic rearrangements in normal single cells from an individual with a germline BRCA1 mutation of unknown pathogenicity might indicate that the actual mutation is pathogenic. This is in a fashion similar to the DEB (diepoxybutane) test used to infer the pathogenicity of Fanconi Anemia (FA) alleles via the measurement of chromosomal breakage and radial formation of FA cells in response to treatment with DEB. Consistent with the feasibility of such an approach is the observation of single cells carrying random non-recurrent genomic alterations in otherwise normal genomes using our methods of single cell copy number profiling [79].

7.6. Applications of C-DOP-L outside of cancer biology and Future directions

Finally, while the approach was devised for the purpose of studying cancer heterogeneity and evolution, it is clear that its applications are not limited to cancer biology. The robustness of the method coupled with its high-throughput nature makes it an attractive approach to examine the CNV patterns underlying aneuploidy in human gametes[113] as well as human neurons[114]. In addition, biological phenomena such as the ploidy conveyor in hepatocytes[115] could very well be carefully dissected using the methods described here. With regards to cancer biology, the application of our high-throughput single cell genome sequencing approach to many tumors types and ultimately hundreds of cancers samples is bound to illuminate the underlying biology behind tumor heterogeneity and help in our struggle to better understand and tackle this disease.

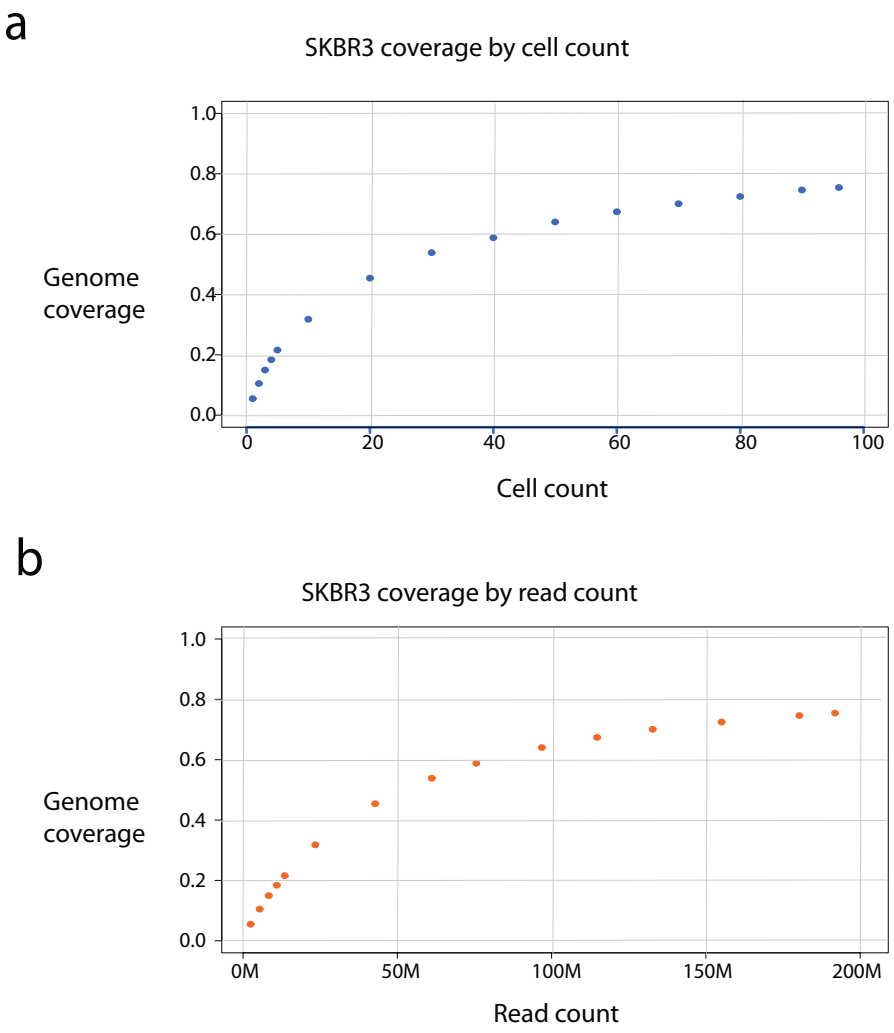


Figure 31: <u>Genome coverage increases with increasing number of single cells sequenced.</u>

Genome coverage at 1X was calculated as a function of increasing number of cells (a) and increasing number of sequencing reads (b) and plotted.

8. SUMMARY AND CONCLUSION

To summarize, high-throughput highly multiplex single cell sequencing provides a powerful approach to dissect the underlying genomic heterogeneity present in cancer genomes at the level of genomic copy number variation. The down-sampling analysis presented represents an important benchmark for single cell whole-genome copy number variation analysis upon which further investigations could build. The introduction of concepts regarding varying levels of resolution in the analysis via different number of bins also provides a good foundation for future study. The optimization of the molecular approach in generating single-cell sequencing libraries offers a resource that facilitates low-cost and high-throughput analysis of single cell genomes. In addition, the malleability of the approach gives it the capacity to adapt to the evolving Next Generation Sequencing landscape. Importantly, the robustness and accuracy of the method facilitated in-depth investigations of intra-tumoral genetic heterogeneity in the context of cancer cell lines and clinical biopsies. These investigations unraveled important biological insights in illustrating substantial genomic heterogeneity in tumor genomes, both in terms of sub-populations structure as well as somatic mosaicism of chromosomal amplifications. With the ever decreasing cost, increasing output, and continual technical evolution of Next Generation Sequencing platforms it is likely that the concepts presented here (minimal read requirements/high level multiplexing) will be further expanded to facilitate the profiling of thousands of single cancer cells and further facilitate the retrieval of additional pieces of genetic information in the form of single nucleotide variants as well as structural variants. This will enable expansive studies

across many tumor samples and will allow for a better comprehension of cancer genome heterogeneity and facilitate in-depth understanding of its contribution to disease progression, therapeutic resistance and cancer metastasis. Clinically, it is tempting to envision a role for single cell sequencing in the monitoring of therapeutic response of patients to therapy, especially within the context of circulating tumor cells. The ability to genetically profile, at a genome-wide level, the cancer genomes of single circulating tumor cells via a minimally invasive fluid biopsy over the course of treatment could offer clinicians valuable information in tackling the evolving nature of cancer. The methods described here represent a significant step towards the realization of these applications and offer a solid foundation upon which further studies will expand.

REFERENCES

1. Beckmann, J. S., Estivill, X. & Antonarakis, S. E. Copy number variants and genetic traits: closer to the resolution of phenotypic to genotypic variability. *Nat Rev Genet* **8**, 639-646 (2007).

2. Watanabe, J., Uehara, K. & Mogi, Y. Adaptation of the osmotolerant yeast *zygosaccharomyces rouxii* to an osmotic environment through copy number amplification of FLO11D. *Genetics* **195,** 393-405 (2013).

3. Jiao, Y. *et al*. Genome-wide genetic changes during modern breeding of maize. *Nat Genet* **44,** 812-815 (2012).

4. Maron, L.G. *et al*. Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 5241-5246 (2013).

5. Bickhart, D.M. *et al*. Copy number variation of individual cattle genomes using next-generation sequencing. *Genome Re* **22,** 778-790 (2012).

6. Sebat, J. *et al.* Large-scale copy number polymorphisms in the human genome. *Science* **305,** 525-528 (2004).

7. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K. Qi, Y., Scherer, S.W. & Lee, C. Detection of large-scale variation in the human genome. *Nat Genet* **36,** 949-951 (2004).

8. Hasin, Y. *et al.* High-resolution copy-number variation map reflects human olfactory receptor diversity and evolution. *PLoS Genet* **4**, e1000249 (2008).

9. Perry, G. H. *et al.* Diet and the evolution of human amylase gene copy number variation. *Nat Genet* **39**, 1256-1260 (2007).

10. Soemedi, R. *et al*. Contribution of global rare copy-number variants to the risk of sporadic congenital heart disease. *Am. J. Hum. Genet* **91,** 489-501 (2012).

11. Warburton, D. *et al*. The contribution of de novo and rare inherited copy number changes to congenital heart diease in an unseleted sample of children with conotruncal defects or hypoplastic left heart disease. *Hum. Genet* **133,** 11-27 (2014).

12. Hitz, M.P. *et al.* Rare copy number variants contribute to congential left-sided heart diease. *PLoS Genet* **8,** e1002903 (2012).

13. Sanna-Cherchi, S. *et al*. Copy-number disorders are a common cause of congential kidney malformations. *Am. J. Hum. Genet* **91,** 987-997 (2012).

14. Materna-Kiryluk, A. *et al.* De novo microduplications at 1q41, 2q37.3, and 8q24.3 in patients with VATER/VATERL association. *Eur. J. Hum. Genet* **21,** 1377-1382 (2013).

15. Sebat, J. *et al*. Strong association of de novo copy number mutations with autism. *Science* **316,** 445-449 (2007).

16. Bochukova, E.G. *et al.* Large, rare chromosomal deletions associated with severe-onset obesity. *Nature* **463,** 666-670 (2010).

17. Levy, D. *et al.* Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron* **70,** 886-897 (2011).

18. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism specturm disorders. *Nature* **466,** 368-372 (2010).

19. Wheeler, E. *et al.* Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. *Nat Genet* **5,** 513-517 (2013).

20. Wang, K., Li, W.D., Glessner, J.T., Grant, S.F., Hakonarson, H. & Price, R.A. Large copy-number variations are enriched in cases with moderate to extreme obesity. *Diabetes* **59,** 2690-2694 (2010).

21. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. *Nature* **458,** 719-724 (2009).

22. Teixeira, M.R., Pandis, N. & Heim, S. Cytogenetic clues to breast carcinogenesis. *Genes Chromosomes Cancer* **2002,** 1 (2002).

23. Fountain, J.W. *et al*. Homozygous deletions within the human chromosome band 9p21 in melanoma. *Proc. Natl. Sci. U.S.A* **89,** 10557-10561 (1995).

24. James, C.D., He, J., Carlbom, E., Nordenskjold, M., Cavenee, W.K. & Collins, V.P. Chromosome 9 deletion mapping reveals interferon alpha and inteferon beta-1 gene deletions in human glial tumors. *Cancer Res* **51,** 1684-1688 (1991).

25. Cairns, P. *et al.* Frequency of homozygous deletion at p16/CDKN2 in primary human tumors. *Nat Genet* **11,** 210-212 (1995).

26. Kallioniemi, A., Kallioniemi O.P., Sudar, D.A., Rutovitz, D. Gray, J.W., Waldman, F. & Pinkel, D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258,** 818-821 (1992).

27. Shinawi, M. & Cheung, S. W. The array CGH and its clinical applications. *Drug Discov Today* **13**, 760-770 (2008).

28. Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D. & Brown, P.O. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet* **1,** 41-46 (1999).

29. Wang, D.G. *et al*. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280,** 1077-1082 (1998).

30. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. *Nat. Genet **21*** **(1 Suppl.),** 20-24 (1999).

31. Beroukhim, R. *et al*. The landscape of somatic copy-number alterations across human cancers. *Nature* **463**, 899-905 (2010).

32. Curtis, C. *et al*. The genomic and transcriptomic architecture of 2,000 breast tumors reveals novel subgroups. *Nature* **486,** 346-352 (2012).

33. Northcott, P.A. *et al*. Subgroup-specific structural variation across 1,000 medulloblastoma genomes. *Nature* **488**, 49-56 (2012).

34. Hoadley, K.A. *et al*. Multi-platform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158,** 929-944 (2014).

35. Zack, T.I. *et al.* Pan-cancer patterns of somatic copy number alterations. *Nat. Genet* **45,** 1134-1140 (2013).

36. Demichelis, F. *et al*. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proc. Natl. Acad. Sci. U.S.A* **109,** 6686-6691 (2012).

37. Krepischi, A.C. *et al*. Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res* **14,** R24 (2012).

38. Stadler, Z.K. *et al.* Rare de novo germline copy-number variation in testicular cancer. *Am. J. Hum. Genet* **91,** 379-383 (2012).

39. Guichard, C. *et al*. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet* **44,** 694-698 (2012).

40. Rudin, C.M. *et al.* Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet* **44,** 1111-1116 (2012).

41. Peifer, M. *et al*. Integrative genome analysis identify key somatic driver mutations in small-cell lung cancer. *Nat. Genet* **44,** 1104-1110 (2012).

42. Nijhawan, D. *et al*. Cancer vulnerabilities unveiled by genomic loss. *Cell* **150,** 842-854 (2012).

43. Shiu, K. K., Natrajan, R., Geyer, F. C., Ashworth, A. & Reis-Filho, J. S. DNA amplifications in breast cancer: genotypic-phenotypic correlations. *Future Oncol* **6**, 967-984, (2010).

44. Spector, N.L. & Blackwell, K.L. Understanding the mechanism behind trastuzumab therapy for human epidermal growth factor receptor 2-positive breast cancer. *J. Clin. Oncol* **27**, 5838-5847.

45. Takano, T. *et al*. Epidermal growth factor receptor gene mutations and increased copy numbers predict gefitnib sensitivity in patients with recurrent non-small-cell lung cancer. *J Clin. Oncol* **23,** 6829-6837 (2005).

46. Cappuzzo, F. *et al*. Increased HER2 gene copy number is associated with response to gefitinib therapy in epidermal growth factor receptor-positive non-small-cell lung cancer. *J Clin.Oncol* **23,** 5007-5018 (2005).

47. Hicks, J. *et al.* Novel patterns of genome rearrangement and their association with survival in breast cancer. *Genome Res* **16,** 1465-1479 (2006).

48. Russnes, H.G. *et al*. Genomic architecture characterizes tumor progression paths and fate in breast cancer patients. *Sci Transl Med* **2,** 38ra47 (2010).

49. Janoueix-Lerosey, I. *et al.* Overall genomic pattern is a predictor of outcome in neuroblastoma. *J Clin. Oncol* **27,** 1026-1033 (2009).

50. Avet-Loiseau, H. *et al*. Prognostic significance of copy-number alterations in mutiple myeloma. *J Clin. Oncol* **27,** 4585-4590 (2009).

51. Huang, YT. *et al.* Genome-wide analysis of survival in early-stage non-small-cell lung cancer. *J Clin. Oncol* **27**, 2660-2667 (2009).

52. Vogelstein, B. *et al.* Cancer genome landscapes. *Science.* **339**, 1546-1558 (2013).

53. Burrell, R.A., McGranahan N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature.* **501**, 338-345 (2013).

54. Yates, L.R. & Campbell, P.J. Evolution of the cancer genome. *Nat. Rev. Genetics.* **13**, 795-806 (2012).
55. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumor heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer.* **12**. 323-334 (2012).

56. Nowell, PC. The clonal evolution of tumor cell populations. *Science*. **194**, 23-28 (1976).

57. Pandis, N., Heim, S., Bardi, G., Idvall, I., Mandahl, N. & Mitelman, F. Chromosome analysis of 20 breast carcinomas: cytogenetic multiclonality and karyotypic-pathologic correlations. *Genes. Chromosomes. Cancer* **6,** 51-57 (1993).

58. Teixeira, M.R., Pandis, N., Bardi, G., Andersen, J.A., Mandah, N., Mitelman, F. & Heim, S. Cytogenetic analysis of multifocal breast carcinomas: detection of karyotypically unrelated clones as well as clonal similarities between tumor foci. *Br J Cancer* **70,** 922-927 (1994).

59. Teixeira, M.R., Pandis, N., Bardi, G., Andresen, J.A., Mitelman, F. & Heim, S. Clonal heterogeneity in breast cancer: karyotypic comparisons of multiple intra- and extra-tumorous samples from 3 patients. *Int J Cancer* **63,** 63-68 (1995).

60. Adeyinka, A., Mertens, F., Bondeson, L., Garne, J.P., Borg, A., Baldetorp, B. & Pandis, N. Cytogenetic heterogeneity and clonal evolution in synchronous bilateral breast carcinomas and their lymph node metastasis from a male patient without any detectable BRCA2 germline mutation. *Cancer Genet Cytogenet* **118,** 42-47 (2000).

61. Metzker, M. L. Sequencing technologies - the next generation. *Nat Rev Genet* **11**, 31-46, (2010).

62. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet* **24**, 133-141 (2008).

63. Pleasance, E.D. *et al.* A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463,** 191-196 (2010).

64. Mardis, E.R. *et al*. Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361,** 1058-1066 (2009).

65. Chiang, D. Y. *et al.* High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods* **6**, 99-103, (2009).

66. Alkan, C. *et al.* Personalized copy number and segmental duplication maps using next-generation sequencing. *Nat Genet* **41**, 1061-1067 (2009).

67. Shah, S.P. *et al*. Mutational evolution in a lobular breast tumor profiled at single nucleotide resolution. *Nature* **461,** 809-813 (2009).

68. Ding, L. *et al*. Genome remodeling in a basal-like breast cancer metastasis and xenograft. *Nature* **464,** 999-1005.

69. Carter, S.L. *et al*. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol* **30,** 413-421 (2012).

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

71. Govindan, R. *et al*. Genomic landscape of non-small cell lung cancer in smokers and never-smokers. *Cell* **150,** 1121-1134 (2012).

72. Walter, M.J. *et al*. Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* **366,** 1090-1098 (2012).

73. Welch, J.S. *et al*. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264-278 (2012).

74. Campbell, P.J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467,** 1109-1113 (2010).

75. Wu, X. *et al*. Clonal selection drives genetic divergance of metastatic medulloblastoma. *Nature* **482,** 529-533 (2012).

76. Diaz, LA. Jr. *et al*. The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537-540 (2012).

77. Misale, S. *et al*. Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer. *Nature* **486,** 532-536 (2012).

78. Van Allen, E.M. *et al.* The genetic landscape of clinical resistance to RAF inihibtion in metastatic melanoma. *Cancer Discov* **4,** 94-109 (2014).

79. Navin, N. *et al*. Tumor evolution inferred from single-cell sequencing. Nature. 472, 90-94 (2011).

80. Baslan, T. *et al*. Genome-wide copy number analysis of single cells. *Nat. Protoc.* 7, 1024-1041 (2012).

81. Xu, X. *et al*. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148,** 886-895 (2012).

82. Hou, Y. *et al*. Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148,** 873-885 (2012).

83. Voet, T. *et al*. Single-cell paired-end genome sequencing reveals structural variation per cell cycle. *Nucleic. Acid. Res*. **41**, 6119-6138 (2013).

84. Smallwood, S.A., Lee, H.J., Angermueller, C., Krueger, F., Saadeh, H., Peat, J., Andrews, S.R., Stegle, O., Reik, W. & Kelsey, G. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. *Nat Methods* **11,** 817-820 (2014).

85. Lohr, J.G. *et al*. Whole-exome sequencing of circulating tumor cells provides a window into metastatic prostate cancer. *Nat. Biotechnol*. **32**, 479-484 (2014).

86. Francis, J.M. *et al*. EFGR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov*. **4**, 956-971 (2014).

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

88. Dago, A.E. *et al*. Rapid phenotypic and genomic change in response to therapeutic pressure in prostate cancer inferred by high content analysis of single circulating tumor cells. *PLOS ONE* **9**, e101777 (2014).

89. Ni, X. *et al*. Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl. Acad. Sci. U.S.A*. **110**, 21083-21088 (2013).

90. Wersto, R.P. *et al*. Doublet discrimination in DNA cell-cycle analysis. *Cytometry* **46**, 296-306, (2001).

91. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285-299, (2012).

92. Wang, K. *et al*. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. **38**, e178 (2010).

93. Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A., & Dewey, C.N. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. **26,** 493-500 (2010).

94. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70, (2012).

95. Parker, JS. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* **27,** 1160-1167 (2009).

96. Langmean, B., Trapnell, C., Pop, M. and Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** 2078-2079 (2009).

97. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* **15**, 2078-2079 (2009).

98. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*. **23,** 657-663 (2007).

99. Krasnitz, A., Sun, G., Andrews, P. & Wigler, M. Target inference from collections of genomic intervals. *Proc. Natl. Acad. Sci. U.S.A*. 110, E2271-E2278 (2013).

100. Arnold, A. & Papanikolaou, A. Cyclin D1 in breast cancer pathogenesis. *J Clin. Oncol*. **23**, 4215-4224 (2005).

101. Engstrom, M.J., Ytterhus, B. Vatten, L.J., Opdahl, S. & Bofin, A.M. *J Clin. Pathol*. **67**, 420-425 (2014).

102. Soulier, J. et al. HOXA genes are included in genetic and biologic networks defining human acute T-cell leukemia (T-ALL). *Blood*. **106**, 274-286 (2005).

103. Sosa, M.S. et al. Identification of the Rac-GEF P-Rex1 as an essential mediator of ErbB signaling in breast cancer. *Mol. Cell*. **40**, 877-892 (2010).

104. Yu, Z., Lu, S. & Huang, Y. Microfluidic whole genome amplification device for single cell sequencing. *Anal Chem* **86,** 9386-9390 (2014).

105. Yang, Y., Swennenhuis, J.F., Rho, H.S., Le Gac, S. & Terstappen, L.W. Parallel single cancer cell whole genome amplification using button-valve assisted mixing in nanoliter chambers. *PLOS One* **9,** e107958 (2014).

106. Hughes, A.E. et al. Clonal architecture of secondary cute myeloid leukemia defined by single-cell sequencing. *PLOS Genet*. **10**, e1004462 (2014).

107. Wanunu, M. Nanopores: A journey towards DNA sequencing. *Phys Life Rev*. **2,** 125-158 (2012).

108. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science.* **270,** 484-487 (1995)

109. Hatzis, C. i. Enhancing reproducibility in cancer drug screening: how do we move forward? *Cancer. Res.* **74**, 4016-4023 (2014).

110. Cho, R.W. & Clarke, M.F. Recent advances in cancer stem cells. *Curr Opin Genet Dev*. **1,** 48-53 (2008).

111. Giaccone, G. Epidermal growth factor receptor inhibitors in the treatment of non-small-cell lung cancer. *J Clin Oncol.* **23,** 3235-3242 (2005).

112. Engelman, J.A. *et al.* MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. *Science*. **316**, 1039-1043 (2007).

113. Hou, Y. *et al*. Genome analysis of single human oocytes. *Cell.* **155**, 1492-1506 (2013).

114. McConnell, M.J. *et al*. Mosaic copy number variation in human neurons. *Science.* **342**, 632-637 (2013).

115. Duncan, A.W. *et al*. The ploidy conveyor of mature hepatocytes as a source of genetic variation. *Nature.* **467**, 707-710 (2010).