Identifying gene regulatory programs of cancer cells in response to chromosomal instability using scRNA-seq

Major Bioinformatics Research Project Thesis
(XM_0070, 36 EC)

Pedro Batista Tan
p.batistatan@student.vu.nl
orcid 0000-0001-9790-9568
Vrije Universiteit Amsterdam Student number: 2658245
Universiteit van Amsterdam Student number: 12524492
17/09/21

Supervisor: Dr. Lodewyk Wessels
l.wessels@nki.nl
Computational Cancer Biology Group Leader
The Netherlands Cancer Institute

Co-supervisor: Dr. Michael Schubert
m.schubert@nki.nl
Post-doctoral researcher
The Netherlands Cancer Institute

Host group: Computational Cancer Biology
The Netherlands Cancer Institute

University examiner: Dr. Evert Bosdriesz
e.bosdriesz@vu.nl
Assistant Professor
Vrije Universiteit Amsterdam

Contents

Identifying gene regulatory programs of cancer cells in response to chromosomal instability using scRNA-seq

Pedro Batista Tan
p.batistatan@student.vu.nl
Vrije Universiteit Amsterdam Student number: 2658245
Universiteit van Amsterdam Student number: 12524492
17/09/21

Supervisor: Dr. Lodewyk Wessels
l.wessels@nki.nl
Group leader
The Netherlands Cancer Institute

Co-supervisor: Dr. Michael Schubert
m.schubert@nki.nl
Post-doctoral researcher
The Netherlands Cancer Institute

University examiner: Dr. Evert Bosdriesz
e.bosdriesz@vu.nl
Assistant Professor
Vrije Universiteit Amsterdam

ABSTRACT

Chromosomal instability (CIN) is the propensity to mis-segregate chromosomes, compromising genome integrity during cell division. CIN has been associated with poor cancer clinical prognosis, therapeutic resistance, metastasis, and pro-inflammatory responses. However, despite its prevalence, the role of CIN in tumor initiation and progression is not entirely understood. Bakhoum and Cantley have proposed a working model to explain distinct inflammatory forces over the course of tumor development, associating CIN with downregulation of Interferon responses, chronic inflammation via NF-κB and EMT in developed tumors. In this project, we leveraged previously published scRNA-seq datasets to study gene signatures associated with CIN and verify these model predictions. Inferred copy number profiles were used to assign tumor samples to CIN-groups, then differential expression and hallmark gene set enrichment analysis were performed. We found that Interferon-α/γ and NF-κB signatures were generally downregulated in CIN-high samples, indicating mechanisms for suppression of pro-inflammatory responses in cancer cells. We also recovered signatures associated with cell cycle, proliferation and oncogenes such as E2F, G2M checkpoint and Myc targets, but we could not detect a robust enrichment of EMT signature. Overall, this work contributes to understanding which transcriptional programs are associated with CIN with single-cell resolution.

1 INTRODUCTION

For genomic stability to be maintained in proliferating cells, it is crucial that genome integrity is preserved by DNA damage repair mechanisms, accurate DNA replication, and cell cycle checkpoints. Cells must also be able to faithfully distribute chromosome copies to daughter cells during mitosis. This process can be disrupted in various ways leading to chromosomal instability (CIN), defined as the propensity to mis-segregate chromosomes during cell division [60].

The plethora of factors that induce CIN includes DNA damage by genotoxic agents, replicative stress during the S phase, and defects in cell cycle checkpoints, chromatid cohesion, centrosome replication, or microtubule attachment. Cell populations that exhibit CIN have frequent mitotic aberrations, such as lagging or acentric chromosomes, extrachromosomal DNA, multipolar spindles, anaphase bridges, and ultrafine DNA bridges. Following these various mitotic defects, daughter cells inherit compromised genome copies, promoting karyotype heterogeneity and accumulating non-euploid chromosome copy numbers known as aneuploidies [60]. These aberrations are usually associated with decreased cell growth and senescence, consistent with the low levels observed in healthy tissues and in early tumorigenesis [48, 56]. Nevertheless, chromosomal abnormalities and aneuploidies are found in most human tumors, indicating that CIN is an important phenomenon in cancer development [11, 44, 62]. At first, it is not clear how this process that compromises fitness and function in normal tissues can be beneficial to cancerous cells, resulting in the aneuploidy paradox [55].

CIN is associated with poor clinical prognosis, resistance to antineoplastic agents and metastasis, which is explained in part by the promotion of karyotype heterogeneity, accelerating tumor evolution and acquisition of resistant subpopulations [21, 41]. Given its clinical significance, it is highly relevant to understand how CIN affects these processes and shapes tumor development in patients. However, excessive CIN has also been shown to sensitize tumor cells to additional genotoxic stress [4, 5, 31, 48, 65], and tolerance to genomic instability varies depending on cancer type. Although CIN and aneuploidies are initially associated with decreased fitness, cells may acquire additional alterations that are permissive to CIN and adapt to aneuploidy states. This allows them to overcome their negative effects and contribute to tumor progression, with preferential selection of chromosome gains and losses for particular cell types [48, 52, 55, 56]. Tissue-specific adaptations and differential sensitivity to genomic instability highlights the importance of determining what CIN-associated gene signatures are generalizable across various cancer types.

To study the impact of CIN and understand its role in tumorigenesis, multiple animal and cell line models have been generated. Chromosomal instability can be induced by overexpression and

deletion of proteins involved in the mitotic spindle assembly checkpoint (SAC) such as Mad2, BubR1, Mps1, and Hec1 [9, 51, 52, 57, 60]. Mis-segregation occurs if anaphase starts before the SAC ensures chromosomes are properly attached to mitotic spindles, but also if the SAC is overactivated, delaying mitotic progression. Experimental CIN can also be generated by interfering with regulators of kinetochore-microtubule attachment such as Kif2b and Kif2c (MCAK) [6, 60]. These proteins are recruited to kinetochores and destabilize improper microtubule attachments that continuously occur during mitosis. Loss of function of these regulators stabilizes erroneous attachments to chromosomes. Altogether, these and many other models have provided key insights on the cellular and molecular basis of CIN [52, 57].

Experimental models of CIN have also revealed important links with pro-inflammatory signalling pathways. Mitotic defects related to CIN frequently cause chromosomes not to be integrated into nuclear envelopes of daughter cells, forming micronuclei structures. These small compartments contain extranuclear DNA and often rupture exposing double-stranded DNA in the cytosol [17, 30], activating cGAS (cyclic GMP-AMP synthase) and increasing concentrations of the cyclic dinucleotide cGAMP. This causes a conformational change in stimulator of interferon genes (STING), its association with TANK-binding kinase 1 (TBK1), and activation of transcription factors like interferon regulatory factor 3 (IRF3), p50/RelA and p52/RelB. In turn, IRF3 activates type I Interferon responses expressing interferon stimulated genes (ISGs), while p50/RelA and p52/RelB activate nuclear-factor kappa-light-chain-enhancer of activated B cells (NF-κB) inflammatory responses [1, 6, 20, 40]. The cGAS-STING pathway is part of innate immunity against DNA virus infections reacting to cytosolic dsDNA but is an important source of immune activation in tumor populations that exhibit CIN.

Acute CIN and its cGAS-STING associated inflammatory response can lead to clearance of cancer cells [20, 49, 53, 63], but given that CIN is present to some extent in most tumors and associated with progression and metastasis, cancer cells need to overcome or bypass this immune antitumoral surveillance. A working model has been put forward by Bakhoum and Cantley to reconcile opposing inflammatory forces, proposing that the link between CIN and cGAS-STING represents a trigger for immune editing throughout tumorigenesis [3]. The authors suggest that during early tumor development, cells exposed to acute CIN respond with inflammatory interferon and NF-κB responses, which leads to a senescence-associated secretory phenotype (SASP), senescence, cell death, and immune clearance of tumor cells. As tumors progress, the STING interferon response is suppressed, but chronic inflammation would persist involving NF-κB signaling. In turn, chronic inflammation and chronic SASP have been associated with the evasion of cell senescence, promotion of epithelial-mesenchymal transition (EMT), and metastasis [6, 10, 20].

Evidence for the model proposed by Bakhoum and Cantley can be obtained by analyzing transcriptomics data of tumor samples with varying levels of CIN. Previous studies have correlated CIN to gene expression signatures, but this was mostly based on tumor bulk microarray or bulk RNA-seq data [8, 12, 19, 24, 26, 54]. Although there is a large quantity of this type of transcriptomics data available in resources such as The Cancer Genome Atlas (TCGA) or Gene Expression Omnibus (GEO), it does not allow to directly assess sample heterogeneity, so aneuploidy alone is used as a surrogate measure for CIN. Furthermore, it is challenging to study distinct cancer inflammatory responses in bulk transcriptomics, as findings are confounded by gene expression programs of other cells in the microenvironment. Enabling the dissection of transcriptional programs specifically in cancer subpopulations requires single-cell resolution, a technology that has only recently been scaled to measure hundreds or thousands of cells within tumors.

The objective of this project is to understand gene expression changes in response to chromosomal instability by leveraging publicly available single-cell RNA sequencing (scRNA-seq) datasets. This approach allows us to obtain measures for sample heterogeneity in addition to aneuploidy, and to distinguish CIN transcriptional signatures that are regulated in cancer cells within tumor environments. More specifically, our research question is focused on looking for evidence of distinct inflammatory responses and EMT proposed by Bakhoum and Cantley [3]. According to this model, CIN-high samples from developed tumors would present signatures of chronic inflammation via NF-κB, markers of EMT, and suppression of interferon-related genes. Our work intends to verify these predictions and identify other CIN-associated gene signatures by studying multiple scRNA-seq datasets and correlating CIN levels with cancer cell transcriptional programs.

2 METHODS

Publicly available scRNA-seq datasets for multiple cancer types were searched in literature references and data repositories. Inclusion of tumor datasets required at least 4 tumor samples and 1000 cell barcodes, with the exception of Chung et al. 2017 [16], as it contained inferred karyotype and whole exome sequencing (WES) data. scRNA-seq molecular counts and barcode metadata for the different datasets were obtained from the Gene expression Omnibus (GEO) [23], the Single Cell Expression Atlas [45], or the Chinese National Gene Bank [13] (Table 1). Exceptions are Tijhuis et al. (unpublished) and Bakhoum et al. 2018, which were obtained in personal communications with the authors. Data analysis was performed using R (v4.0.3) and associated packages according to the following steps:

2.1 Data Preprocessing

All single-cell molecular count datasets went through standard preprocessing and analysis using the Seurat package (v4.0.1) [50]. When datasets had ENSEMBL gene IDs as features, these were mapped to HGNC symbols using biomaRt (v2.46.3) [22], and only genes with HGNC symbols were retained. Counts were aggregated when multiple features mapped to the same symbol, and one symbol was chosen arbitrarily when an ID mapped to multiple symbols. Datasets then underwent quality control, filtering out barcodes with less than 200 features, less than 1000 total molecular counts, or mitochondrial content above 30%, when mitochondrial gene symbols were present (starting with "MT-"). Additional filters were

applied to cells with outlying values for these features, with cutoffs determined visually and by observing the distribution in each dataset.

Seurat objects were created using the molecular count matrices, log normalized using the median library size of all cells as a scale factor, and then scaled. The 2000 most variable features were identified for each dataset and principal component analysis (PCA) was performed for dimensionality reduction, retaining the first 200 principal components. Cells were clustered using the variable features and the manifold data structure was embedded in 2D Uniform Manifold Approximation and Projection (UMAP) space, determined using 200 principal components. Cell cycle phases and scores were assigned using gene signatures from Tirosh et al. 2016 [61], or their respective mouse homologs obtained with biomaRt. All datasets were inspected visually in UMAP plots, labeling Seurat clusters, cell cycle phase, and cell types.

2.2 Inferring Copy Number Variations

Copy number variations were inferred from scRNA-seq data using the InferCNV (v1.6.0) package from the Trinity CTAT Project (https://github.com/broadinstitute/inferCNV). The chromosome name, start and end positions were obtained for all genes using biomaRt (v2.46.3) [22], and only genes located in chromosomes X and 1 up to 22 (or 19 in mice) were considered in the analysis. Tumor cells were labeled annotating each sample, and only samples containing a minimum of 20 cells were used. For larger datasets, this cutoff was increased up to 50, discarding underrepresented samples. The euploid reference groups depended on the dataset but always consisted of a combination of non-tumor cells (table 1). The InferCNV cutoff for expression was set to 0.1, and the analysis mode was set to subclusters, assuming sample heterogeneity.

2.3 Karyotype Metrics

To measure heterogeneity and aneuploidy in scRNA-seq samples, scoring functions were developed using the InferCNV expression matrix. These functions were largely based on Aneufinder karyotype scoring functions for single-cell whole genome sequencing (scWGS) data [7].

2.3.1 Aneuploidy score. The aneuploidy score A for each gene i was determined by the absolute difference between its expression value x_i minus the reference physiological state S (1 by default) across cells. The sample aneuploidy score sA was calculated as the mean aneuploidy score across all genes A_i.

$$A_i = mean(abs(x_i - S))$$

$$sA = mean(A_i)$$

2.3.2 Heterogeneity score. To determine the heterogeneity score, the expression matrix was first discretized using intervals relative to the mean expression state M, from 0.05 up to 10.05 with regular steps of 0.05. Values outside this range were assigned to $(-\infty, 0.05]$ and $(10.05, \infty)$ bins.

$$breaks = -\infty, range(0.05, 10.05, by = 0.05)M, \infty$$

After discretization, each gene-sample combination was tabulated by counting how many cells had expression values x_i within each interval, and this was rearranged in descending order. This generates a decreasing frequency vector F of length N, where N is the number of expression bins. The heterogeneity score for each gene H_i was determined by multiplying F with a weighting vector $(0 : N - 1)$ and dividing by the total number of cells. This gives the most frequent expression state the weight of zero and increases the weight of rare expression intervals. Therefore, H_i is higher for genes where cells are distributed among many different expression states and zero if all cells fall within one expression bin. The sample heterogeneity score sH was calculated as the mean heterogeneity score across all genes H_i.

$$H_i = F[0 : N - 1]/sum(F)$$

$$sH = mean(H_i)$$

2.4 Differential Expression

Within each dataset, samples containing the highest heterogeneity and aneuploidy scores were labeled as "CIN-high", and the remaining samples as "CIN-low". Pseudo bulk samples were generated by aggregating counts across all cells for each sample and rounding to integers. Differential expression between CIN-high and CIN-low groups was performed using DESeq2 (v1.30.1) [39] using the Wald significance test.

Differential expression for Bakhoum et al. 2018 [6] bulk data was obtained similarly but without the need to generate pseudo bulk samples. Initially, all 5 samples were kept and normal log fold change shrinkage was applied to reproduce findings from the original paper. We then analyzed data using only the 3 samples present in scRNA-seq data (Kif2b, MCAK and dnMCAK; 3-sample bulk), and compared results between bulk and single-cell datasets.

2.5 Gene Set Enrichment Analysis

Pre-ranked gene set enrichment analysis (GSEA) was performed with the fgsea (v1.16.0) package [34] using DESeq2 Wald statistic values as a ranking metric. For analysis of the Bakhoum et al. 2018 bulk dataset with all cell lines, DESeq2 normal shrunk log2 fold changes were used instead to reproduce the original analysis. GSEA was performed on the molecular signatures database (MSigDB) Hallmark gene set collection obtained from msigdbr (v7.4.1) [38].

3 RESULTS

To study differential gene expression associated with CIN, we first attempted to reproduce the bulk and single-cell transcriptomic analysis from Bakhoum et al. 2018 [6], which represented our main reference for an experimental CIN dataset. In this study, the authors generated breast cancer cell lines with varying levels of CIN by overexpressing different Kinesin-13 microtubule-kinetochore attachment regulators. Parental MDA-MB-231 control cells are already chromosomally unstable and aneuploid (~3n), and overexpression of Kif2b and MCAK led to a reduction of chromosome missegregation rates. Overexpression of a dominant negative mutant of MCAK (dnMCAK) modestly increased CIN, while no differences were observed when expressing Kif2a relative to parental

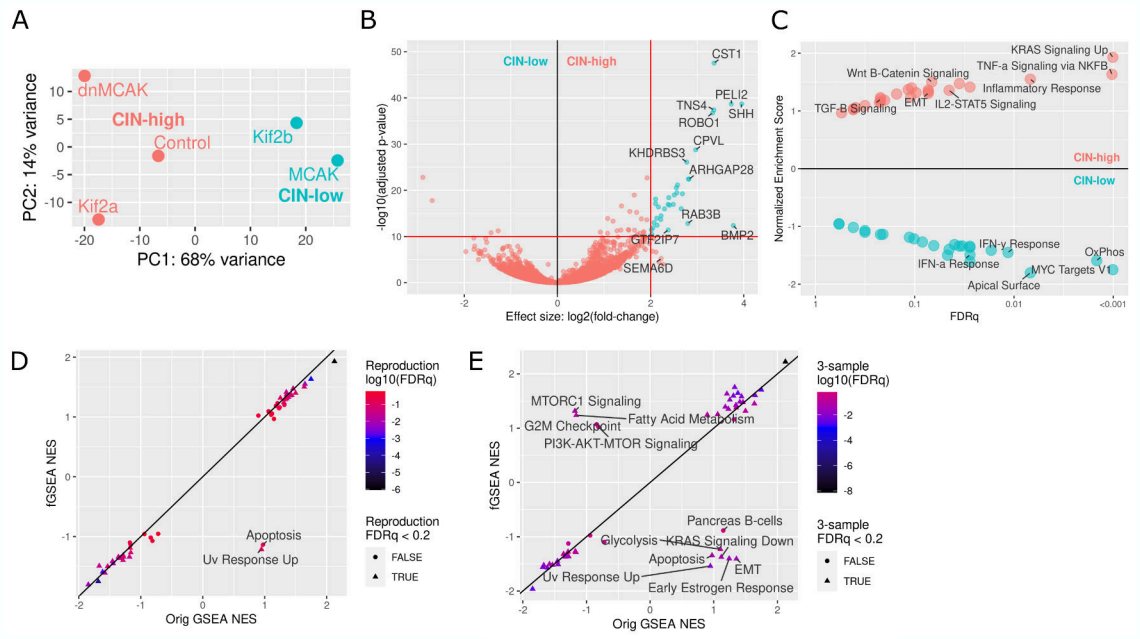

Figure 1: Bulk transcriptomics analysis of Bakhoum et al. 2018 [6]. (A) Principal Component Analysis score plot of MDA-MB-231 cell lines based on bulk RNA expression. CIN-high samples are labeled in red and CIN-low in blue. (B) Volcano plot displaying DESeq2 differential expression results between CIN-high and CIN-low samples after normal log fold change shrinkage. The horizontal and vertical lines indicate the adjusted p-value and log2 fold change thresholds used to include genes in the CIN signature. The annotated genes are part of the CIN signature of the original analysis. (C) Pre-ranked gene set enrichment analysis plot using DESeq2 shrunk log2 fold changes and the MSigDB hallmark gene sets. (D-E) Comparison between original GSEA results and results obtained with fGSEA using (D) shrunk log2 fold changes on all cell lines or (E) Wald statistic only on dnMCAK, MCAK, and Kif2b samples (3 sample bulk).

cells. These cell lines were then divided into CIN-high (control, dnMCAK and Kif2a) and CIN-low groups (Kif2b and MCAK). The authors proceeded to show that CIN-high samples are associated with a higher metastatic burden and lower survival when injected in the mammary fat pads or cardiac ventricles of mice, and that metastatic samples had higher levels of CIN relative to primary tumors.

When reproducing bulk results, cell lines generated in this study clustered according to CIN levels in PCA space using bulk transcriptomics data (Fig. 1A). Groups could be clearly distinguished by the scores on the first principal component (PC1), which accounts for 68% of the variance. Differential expression between CIN groups revealed genes that are highly expressed in CIN-high samples (Fig. 1B). In the original analysis, the top 23 genes were used to derive a CIN-signature based on DESeq2 normal shrunk log2 fold changes above 3 and adjusted p-values below 10^{-15}. Stratification based on this CIN signature had predictive value for distant metastasis-free survival in a breast cancer patient cohort [6]. Adjusting the log2 fold change and p-value cutoffs to 2 and 10^{-10}, respectively, we detected 34 most differentially expressed genes that recover 22 out of 23 genes in the original CIN signature.

Using DESeq2 shrunk log2 fold changes, we performed pre-ranked gene set enrichment analysis (GSEA) on the MSigDB hallmark gene sets. CIN-high samples were found to be enriched for inflammatory responses, EMT and NF-κB via TNF-α pathways while downregulating Interferon-α and Interferon-γ responses (Fig. 1C). These results are largely consistent with the original bulk analysis (Fig. 1D), except for the UV response up gene set, which was negatively enriched but close to the FDR cutoff.

However, our main interest was studying scRNA-seq data, which provides single-cell resolution to study karyotype heterogeneity. As single-cell transcriptomics data was only available for Kif2b, MCAK and dnMCAK samples, we also performed differential expression using only these cell lines (3 sample bulk). There were considerable disagreements between the sign of normalized enrichment scores (NES) of original results when only using 3 samples (Fig. 1E). One of the most important was the negative enrichment of the EMT gene set, which does not agree with the model proposed by Bakhoum and Cantley, and indicates that the bulk positive EMT signal was driven mostly by inclusion of control and Kif2a cell lines.

We next investigated the Bakhoum single-cell transcriptomics data. After quality control filtering, the dataset contained a total of 6778 cells, including 2223 dnMCAK (CIN-high), 2132 Kif2b and 2423 MCAK barcodes (CIN-low) (Fig. 2A). Differential expression

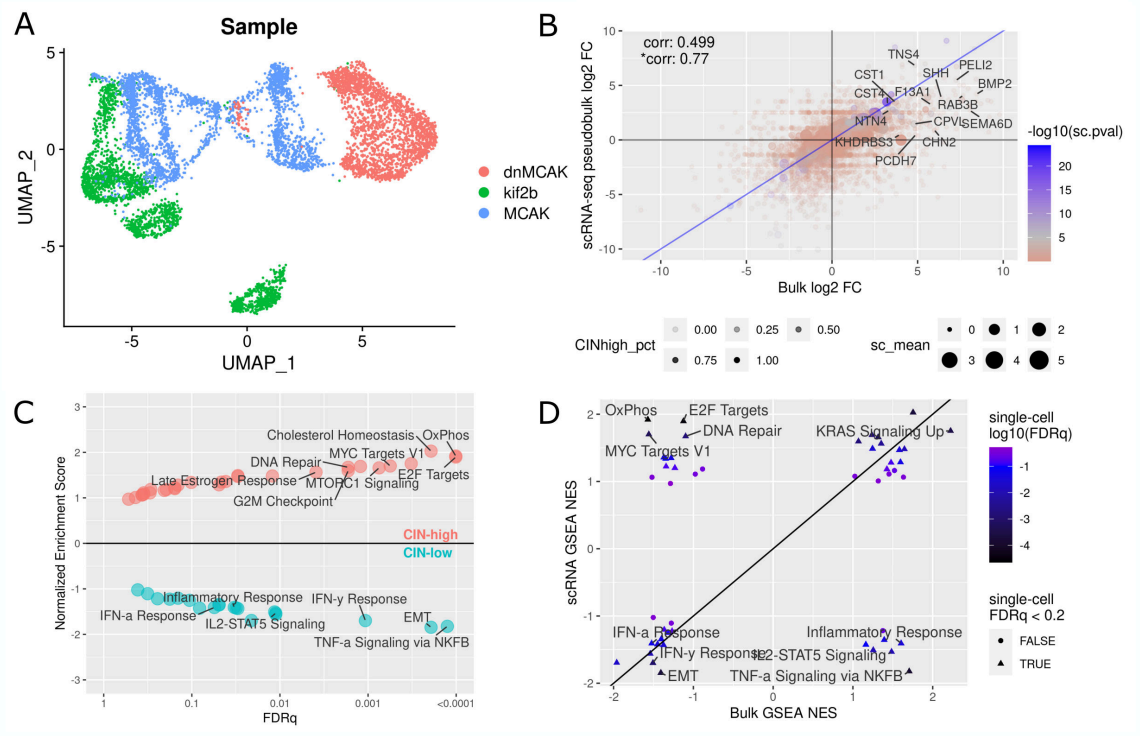

Figure 2: Analysis of Bakhoum et al. 2018 [6] scRNA-seq data. (A) UMAP plot of single-cell data based on principal components, labeling cell line samples. (B) Comparison between log2 fold changes obtained in single-cell pseudo bulk and bulk differential expression analysis using dnMCAK, MCAK and Kif2b samples. The alpha transparency channel corresponds to the percentage of CIN-high cells that expressed each gene and the size represents the mean log1p expression across all cells. The correlation between log2 fold changes is indicated on the top left, and the value after filtering genes expressed in at least 1% of CIN-high cells is indicated with an asterisk below. (C) Pre-ranked gene set enrichment analysis plot using DESeq2 Wald statistic values and the MSigDB hallmark gene sets. (D) Comparison between fGSEA results of single-cell pseudo bulk and the 3-sample bulk results, using the Wald statistic as a ranking metric.

between pseudo bulk samples was performed, and results were compared to the 3 sample bulk for the 14728 genes present in both datasets (Fig. 2B). Although there is considerable noise, especially for genes with low expression levels, there was a reasonable correlation structure between bulk and single-cell pseudo bulk log2 fold changes (corr. = 0.499). This was improved when filtering genes that are expressed in at least 1% of CIN-high cells (10859 out of 14728 genes), reaching a correlation of 0.770. These results illustrate a tradeoff between the number of genes that are included and the reliability of log2 fold change estimates. We also compared the Wald statistic for each gene, which corresponds to the ratio between log2 fold changes and DESeq2 estimates for the standard error (lfcSE). This prioritizes significant features with smaller errors, penalizing genes that achieve large fold changes due to high variability. Comparisons of DESeq2 Wald statistic values had a larger correlation (0.695), even without filtering low expression of genes.

After differential expression, pre-ranked GSEA was performed on single-cell pseudo bulk Wald statistic values on the MSigDB hallmark sets (Fig. 2C). When comparing these results to the 3-sample bulk, most gene sets had consistent NES values (Fig. 2D), importantly showing a consistent downregulation of Interferon responses and EMT. The original analysis suggested an association between CIN-high cells and EMT, but it focused only on studying cell subpopulations [6]. The authors described an “M” group that comprised ~45% of cells and was enriched for EMT genes. We were unable to detect a particular enrichment of EMT in dnMCAK cells when studying cell subpopulations with Seurat clusters. While an association between CIN and EMT may exist, it does not seem to be robustly supported by our analysis of the Bakhoum bulk and single-cell data.

Cell populations that exhibit CIN have a tendency to generate aneuploidies and karyotype heterogeneity over the course of cell divisions. Therefore, this should likely be reflected in the patterns of single-cell copy number alterations (CNAs). To study karyotype profiles using single cells, the inferCNV package was used to assign CNAs based on transcriptomics data using CIN-low cells as the reference and CIN-high as the observation group (Fig. 3A). Surprisingly, karyotype heterogeneity was much lower than expected in dnMCAK cells. We observed patterns of gains and losses that are almost clonal, such as in chromosomes 12, X, and 18. Karyotypes were scored for aneuploidy and heterogeneity using the inferCNV

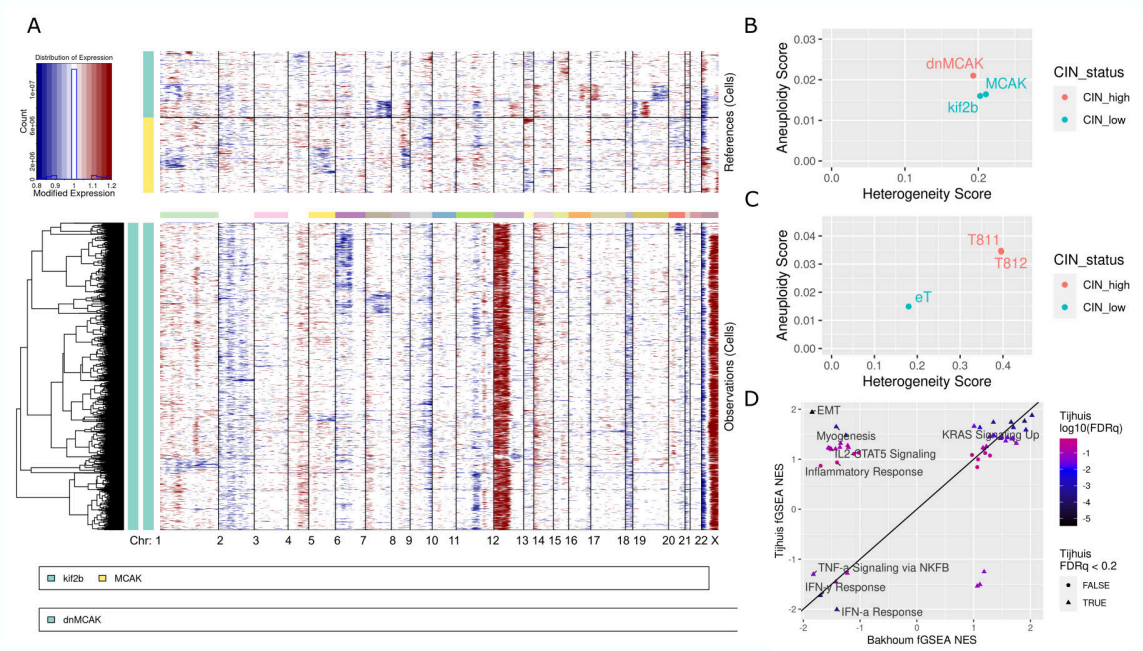

Figure 3: InferCNV copy number profiles and karyotype scores. (A) InferCNV plot of the Bakhoum scRNA-seq data. The Kif2b and MCAK cell lines (CIN-low) were used as references (top) and dnMCAK (CIN-high) as observations (bottom). Chromosome names are identified along the columns, red represents copy number gains and blue copy number losses relative to reference groups. (B) Karyotype scores of Bakhoum cell line samples. (C) Karyotype scores of the Tijhuis mouse thymic samples. (D) Comparison between fGSEA results of Bakhoum and Tijhuis single-cell pseudo bulk, using the Wald statistic as a ranking metric.

expression matrix and functions analogous to Aneufinder karyotype metrics for single-cell whole-genome sequencing (scWGS) (Fig. 3B). In broad terms, the aneuploidy score measures the mean copy number deviation from a physiological state, while the heterogeneity score measures how spread is the distribution of cells across different copy number states. CIN-high cells had a slightly higher aneuploidy score, but this was not the case for karyotype heterogeneity. It should be noted, however, that parental MDA-MB-231 cells used to derive these cell lines have high levels of CIN and aneuploidy. Overexpression of Kif2b and MCAK may reduce missegregation rates, but this would not revert accumulated aneuploid states. Ideally, inferCNV should use euploid cells as a reference group, but this was not available for the Bakhoum single-cell data.

This analysis was also performed on another experimental CIN single-cell dataset of Tijhuis et al. (unpublished) that consists of thymic cells from T-cell acute lymphocytic leukemia (T-ALL) mice models. CIN was induced specifically in mouse thymocytes by co-deletion of the SAC protein Mad2 and p53 using a conditional Lck-Cre-lox system in two mouse samples (T811 and T812) [9, 51, 52]. CIN-low control cells were obtained from a MutS homolog 2 knock-out sample (eT; Msh2 -/-). These mice have deficient DNA mismatch repair and develop spontaneous tumors that are mostly euploid, as determined previously by scWGS. Consistently, Tijhuis karyotype scores based on InferCNV profiles reflect the CIN status of samples (Fig. 3C), with less aneuploidy and heterogeneity in eT references. Differential expression and GSEA were performed between CIN-high and CIN-low cells, and these were compared with Bakhoum gene set NES values (Fig. 3D). Notably, there is a consistent down-regulation of Interferon-α, Interferon-γ and NF-κB inflammatory responses between both datasets. While inflammatory responses and EMT gene sets were negatively enriched in the Bakhoum CIN-high cell line, the Tijhuis thymic CIN-high samples had positive NES for these signatures.

To study how general these CIN transcriptional responses are, we expanded our analysis to multiple scRNA-seq datasets of human patient tumors across different cancer types (Table 1). We searched for publicly available datasets in the single-cell expression atlas, the recount3 R package and in literature references. Small datasets with less than 1000 cells were generally excluded, except for Chung et al. 2017 , as it represented another breast cancer dataset and contained inferCNV analysis with matching bulk WES data [16]. There was a high correlation between aneuploidy and heterogeneity scores (0.795) when observing the distribution of samples over all datasets (Fig. 4). For most of them, there was a reasonable spread of scores across samples, allowing straightforward assignment of CIN groups. Noteworthy exceptions are Chung et al. 2017 [16] and Nelson et al. 2020 [43]. The Chung et al. dataset was the smallest one included, and most of the tumors had high aneuploidy and heterogeneity scores compared to all other samples. Nelson et al. is an ovarian carcinoma biobank dataset, where samples also had scWGS data available. InferCNV was unable to detect most copy number gains

Table 1: Description of scRNA-seq datasets. The number of barcodes after quality control filtering and the number of samples after filtering based on a minimum number of tumor cells are indicated in parenthesis.

Reference	Cancer Type	# of Barcodes (post filtering)	# of Samples (post filtering)	InferCNV Reference Groups	Data Source	Accession Number/ID
Bakhoum et al. 2018 [6]	Luminal breast cancer cell lines (MDA-MB-231)	2211840 (6778)	3 (3)	Kif2b and MCAK CIN-low cell lines	Authors	NA
Tijhuis et al. (unpublished)	T-Cell Acute Lymphoblastic Leukemia	2090421 (5645)	3 (3)	eT euploid tumor cells	Authors	NA
Sun et al. 2020 [59]	Hepatocellular Carcinoma	16498 (16498)	18 (10)	Epithelial, Endothelial and Hepatocellular Stem Cells	China National Genebank	CNP0000650
Nelson et al. 2020 [43]	Ovarian Carcinoma	20321 (19631)	4 (4)	non-tumor cells (derived from Seurat clusters)	Single Cell Expression Atlas	E-MTAB-8559
Chung et al. 2017 [16]	Breast cancer, luminal and basal	540 (485)	12 (6)	Stromal, Myeloid, B and T cells	Single Cell Expression Atlas	E-GEOD-75688
Giustacchini et al. 2017 [28]	Chronic Myeloid Leukemia	2151 (2102)	26 (12)	Normal Bone Marrow Mononuclear cells	Single Cell Expression Atlas	E-GEOD-76312
Darmanis et al. 2017 [18]	Glioblastoma	3589 (3558)	4 (4)	Astrocytes, mature and Oligodendrocytes precursors	GEO	GSE84465
Karaayvaz et al. 2018 [32]	Triple Negative Breast Cancer	1534 (1183)	6	Macrophages, Stromal, T, B and Endothelial Cells	GEO	GSE118390
Tirosh et al. 2016 [61]	Melanoma	4645 (4645)	15 (8)	Macrophages, Fibroblasts, T, B, Endothelial and NK cells	GEO	GSE72056
Puram et al. 2017 [46]	Head and Neck Squamous Cell Carcinoma	5903 (5903)	19 (13)	Fibroblast and Endothelial Cells	GEO	GSE103322
Lee et al.2020 [37]	Colorectal Cancer	91103 (91103)	29 (26)	Epithelial cells	GEO	GSE144735 GSE132257

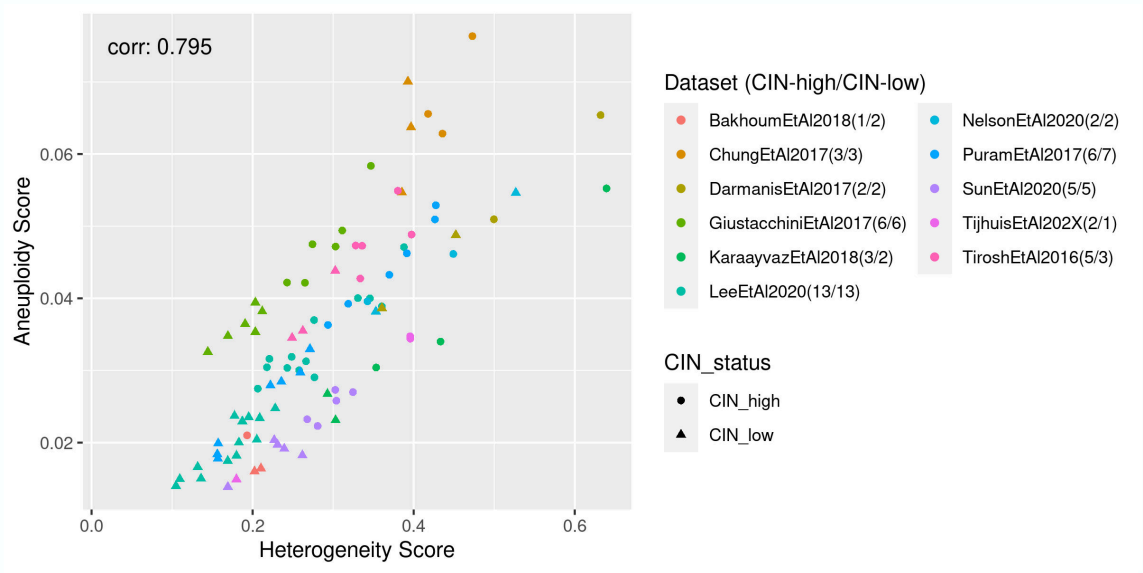

Figure 4: Distribution of sample aneuploidy and heterogeneity scores across datasets. The correlation between the heterogeneity and aneuploidy scores across all samples is indicated on the top left. The number of samples CIN-high and CIN-low samples within each dataset are indicated in parenthesis, respectively.

in a particularly aneuploid sample (patient 38b). Therefore, CIN group assignment for this dataset was based on Aneufinder scWGS karyotype scores instead, which are provided in the original supplementary material [43].

After CIN-groups were assigned based on Karyotype scores, pseudo bulk differential expression and fGSEA was performed on Wald statistic values using the MSigDB hallmarks (Fig. 5). Hallmark gene sets are presented in a heatmap and ordered according to the mean NES across datasets, shown in the first row. Consistent with the model proposed by Bakhoum and Cantley, the most downregulated gene sets in CIN-high samples across datasets were Interferon-α and Interferon-γ responses. The Chung and Darmanis datasets were exceptions with significant positive enrichment for

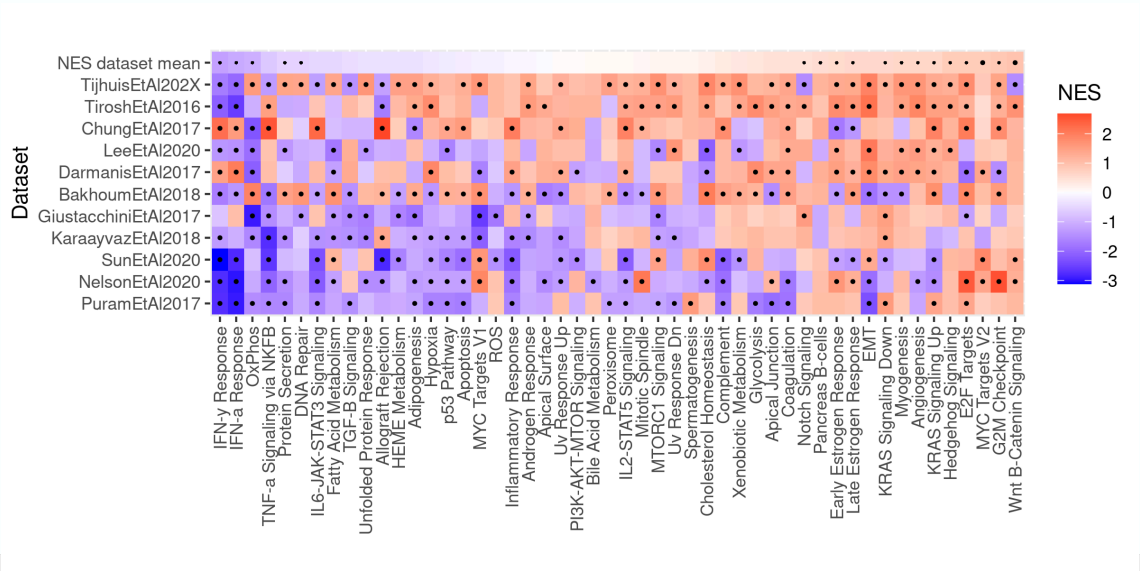

Figure 5: CIN-high samples downregulate interferon and NF-κB responses. Comparison of GSEA normalized enrichment scores (NES) across datasets. Gene sets with a FDRq value below 0.2 are annotated with large dots. Gene sets are ordered by values in the first row, which contains the mean NES across datasets. Small dots indicate gene sets with p-value < 0.2 in a single sample t-test, testing the mean NES against a null model of 0.

these gene sets. We did not find evidence for chronic inflammation involving NF-κB signaling in CIN-high samples, as it was among the most negatively enriched across datasets. The inflammatory response and EMT gene sets were also not robustly enriched in CIN-high samples in either direction, so more data should be included to support conclusions. Notably, among the most upregulated gene sets we detected oncogenic, cell cycle and proliferation gene sets like Wnt/β-catenin, E2F, MYC targets and G2M checkpoint.

4 DISCUSSION

4.1 Transcriptional responses to CIN

In this work, we provide evidence for transcriptional programs that are associated with CIN across multiple cancer scRNA-seq datasets. Previous studies have also investigated transcriptional changes correlated with CIN or aneuploidy, but these were largely based on microarray or bulk transcriptomics data [8, 12, 19, 24, 26, 54]. This type of data does not allow to distinguish within sample heterogeneity, so aneuploidy alone is commonly used as a surrogate measure for CIN. In addition to patient tumor samples, our analysis included two datasets with experimentally induced CIN: Bakhoum et al. with breast cancer MDA-MB-231 cell lines and Tijhuis et al. with T-ALL mouse thymic samples.

In both experimental CIN datasets, we detected downregulation of interferon and NF-κB responses and upregulation of E2F and G2M checkpoint in CIN-high samples, but Tijhuijs and Bakhoum disagreed on many hallmark gene sets, including EMT. This result could be explained by differences in cell types, animal species, or between in vivo and in vitro consequences of CIN. When expanding the analysis to patient tumor datasets, we expected larger variability due to heterogeneity between patient samples, specific mutations, tumor stage and cancer types. Despite this, some gene set observations hold for most patient tumor datasets, and seem to indicate robust transcriptional changes associated with CIN.

The downregulation of interferon responses is in line with the working model by Bakhoum and Cantley and previous observations where cancer cells lose the ability to respond to cytosolic DNA or oncolytic DNA viruses [35, 64]. Additionally, the Interferon-γ gene signature has been negatively correlated with CNAs and mutational burden in TCGA samples [8, 19]. Signalling by cGAS-STING is key to sensing of dsDNA in the cytosol linking CIN with inflammatory responses. Different mechanisms have been suggested to explain the loss of cGAS-STING interferon signaling in cancer cells. Suppression of cGAS-STING associated responses does not seem to be associated with mutations in either cGAS or STING, as these are rarely detected in cancer samples and known mutations usually do not impact protein functions [33, 64]. Also, activation of cGAS-STING in MDA-MB-231 cells did not lead to increased interferon and ISG responses [6], which argues for other forms of suppressing cGAS-STING related pathways. These include antagonism by oncogenes, reduction of STING expression by epigenetic silencing, loss of interferon gene clusters on chromosome 9p and activation of the p38 MAPK stress pathway, that is able to inhibit STING and its interferon signaling [3, 14, 33, 35, 58]. In our results, only the Chung et al. 2017 [16] and Darmanis et al. 2017 [18] data had a positive enrichment for Interferon-α /γ gene sets. Chung samples had overall high aneuploidy and heterogeneity scores, so it is possible that CIN-low assignment did not accurately reflect their CIN status. Darmanis consists of glioblastoma multiforme samples, a

cancer type which was previously reported to not be associated with downregulation of Interferon [19].

Bakhoum and colleagues proposed that even if interferon responses are lost, cGAS-STING activity could be maintained, activating NF-κB targets in tumor cells and contributing to chronic inflammation and SASP [3, 6]. This was not observed for most datasets included in this study, as the NF-κB inflammatory pathway was generally downregulated. For this to occur, cGAS-STING activity would have to be preserved, while selectively inhibiting downstream targets related to interferon, or NF-κB would need to be activated through other signaling pathways. An alternative explanation is that NF-κB signaling is still active in cancer cells, but transcriptional changes associated with its gene signature might be suppressed through other pathways. This could happen due to confounding overlap in downstream targets with interferon or other inflammatory responses, masking NF-κB activity.

Interestingly, we also observed a positive association between CIN and cell cycle and proliferation gene sets. Upregulation of proliferation and cell cycle signatures such as E2F and G2M checkpoint are consistent with previous reports in tumor samples [8, 19, 24], although negative associations have also been detected in cancer cell lines [54]. Furthermore, oncogenic pathways like Myc targets and Wnt/β-catenin also obtained high mean NES across datasets. Oncogene expression and deregulation of the cell cycle is expected to generate CIN, especially if this induces replicative stress and compromises cell cycle checkpoints. Expression of Myc targets has been previously associated with aneuploidy in pan-cancer TCGA data [8, 26]. Myc activation has also been mechanistically linked with CIN by perturbation of cell cycle checkpoints, progression to mitosis with DNA damage and deregulation of mitotic spindle genes leading to defects in spindle formation [42, 47]. Activation of Wnt/β-catenin is caused by CIN in colon cancer due to loss of APC function, and Wnt/β-catenin was also shown to induce CIN in cell line models [2, 29]. Overall, these results contribute to the understanding of transcriptional programs associated with CIN, but analysis should be expanded to include more datasets and cancer types.

4.2 Limitations in CIN assignment

Although some robust gene signatures were detected, important limitations to this study need to be addressed. Except for the experimental CIN datasets in Bakhoum and Tijhuis, the CIN status of samples was not actually known. CIN-high and CIN-low assignments were based on surrogate measures taken from inferCNV copy number profiles.

For the Bakhoum dataset, karyotype heterogeneity in CIN-high dnMCAK cell line was lower than expected, as reflected in the inferCNV CNA profile and heterogeneity score. This could be explained by strong selection pressures on karyotypes, purifying the cell population from cells that deviate from optimal CNA profiles in culture. Another possibility is that inferCNV was not sensitive enough to assign copy number gains and losses in Bakhoum CIN-high cells, underestimating aneuploidy and heterogeneity. By contrast, karyotype aneuploidy and heterogeneity were readily observable in the Tijhuis T811 and T812 samples (Mad2; p53 double knockout). However, this dataset contained a proper euploid reference in eT (Msh2 -/-) cells, indicating this may help improve results for inferCNV CNA detection.

Some patient tumor samples also underwent WGS or WES in the original publications, which allowed a comparison between these karyotypes and the ones obtained with inferCNV. Specifically, this data was available for Chung et al. 2017 [16], Nelson et al. 2020 [43] and Karaayvaz et al. 2018 [32]. For the majority of samples, there was a high agreement between the assigned copy number alterations and WGS/WES data. However, inferCNV failed to retrieve most copy number gains for a highly aneuploid ovarian carcinoma tumor in Nelson et al. 2020. For this reason, Nelson et al. aneufinder scWGS karyotype metrics were used instead of the inferCNV derived aneuploidy and heterogeneity scores. InferCNV only quantifies expression changes from the mean, and whole-genome ploidy gains might fail to be detected when normalizing cells across inferCNV groups. This is problematic when studying extremely aneuploid samples, which are more common depending on cancer type, such as ovarian carcinoma. In future work, it would be interesting to compare the performance of inferCNV and other tools to obtain copy number profiles from scRNA-seq data such as Honeybadger [25] or Copykat [27]. Copykat is a recently published method that is particularly promising, as this tool enables copy number assignment with a high spatial resolution at 5 Mb bins. It is supposedly more accurate in resolving chromosomal breakpoints and could be more powerful to detect heterogeneity in tumor subpopulations.

Additionally, the karyotype scoring functions developed in this study could be improved. As currently implemented, scores are calculated and averaged at the gene level. It would be possible to use bins for the genome instead, and derive weighted measures that assign each chromosome an equal weight, similar to the Weighted Genome Integrity Index (wGII) [15, 36]. This would take into account differences in gene density across the genome, and decrease the relative importance of regions which are particularly enriched in genes.

Another limitation of CIN assignment is that this required a spread of karyotype scores within datasets. This is not as problematic when many tumor samples are present, but especially for smaller datasets, incorrect assignment or lack of distinct CIN-groups can drastically impact differential expression results. A possible approach to circumvent this would be to integrate datasets, correct for batch effects and create large Seurat and inferCNV objects. As this is expected to yield a continuous distribution of karyotype scores across many samples, it could also be preferable to run differential expression using a continuous karyotype score metric instead of labeling samples and using discrete CIN-high and CIN-low groups.

5 CONCLUSION

In this study, hallmark transcriptional programs related to chromosomal instability were identified in multiple scRNA-seq datasets. To our knowledge, this is the first cross-tissue report of CIN gene expression responses specifically in cancer cells and within tumor environments, an approach that is only possible with single-cell transcriptomics resolution. We conclude that CIN is associated with downregulation of interferon and NF-κB esponses and upregulation of MYC and proliferation signatures across different cancer types. Our results do not support the enrichment of EMT and general inflammatory pathways related to CIN. Altogether, this work contributes to the understanding of gene expression regulation in response to chromosomal instability in cancer cells.

6 CODE AND DATA AVAILABILITY

All code used for analysis, to generate the figures and detailed instructions on how data was downloaded are available on Github (https://github.com/1PTan/NKI-CIN-internship). With the exception of Tijhuis et al. and Bakhoum et al. 2018, all scRNA-seq data is previously published and publicly available in repositories (Table 1). All secondary data generated in this project is available upon request.

7 ABBREVIATIONS

BubR1 - Bub1-related kinase
cGAMP - Cyclic guanosine monophosphate–adenosine monophosphate
cGAS - Cyclic GMP-AMP synthase
CIN - Chromosomal Instability CNAs - Copy number alterations
DNA - Deoxyribonucleic acid
dnMCAK - double negative Mitotic centromere-associated kinesin
EMT - Epithelial-mesenchymal transition
FC - Fold change
GEO - Gene expression omnibus
GSEA - Gene set enrichment analysis
IRF3 - Interferon regulatory factor 3
ISGs - Interferon stimulated genes
Kif2b - Kinesin family member 2B
lfcSE - Log fold change standard error
Mad2 - Mitotic arrest deficient 2
MCAK - Mitotic centromere-associated kinesin
Mps1 - Monopolar spindle 1
MSigDB - Molecular Signatures Database
NES - Normalized enrichment score
NF-κB - Nuclear-factor κ-light-chain-enhancer of activated B cells
PCA - Principal component analysis
SAC - Spindle assembly checkpoint
SASP - Senescence-associated secretory phenotype
scRNA-seq - Single-cell RNA sequencing
scWGS - Single cell whole genome sequencing
STING - Stimulator of interferon genes
T-ALL - T-cell acute lymphocytic leukemia
TBK1 - TANK-binding kinase 1
UMAP - Uniform manifold approximation and projection
WES - whole exome sequencing
wGII - Weighted Genome Integrity Index

ACKNOWLEDGMENTS

I am grateful to Michael Schubert and Lodewyk Wessels for the supervision, all the interesting discussions, for accepting me in the NKI Computational Cancer Biology group and for giving me the opportunity to work on this project. I would also like to thank Andréa Tijhuis and Floris Foijer for kindly providing their unpublished dataset of experimental mouse T-ALL models; and Bakhoum et al. for providing the single-cell molecular counts.

REFERENCES

[1] Takayuki Abe and Glen N Barber. 2014. Cytosolic-DNA-mediated, STING-dependent proinflammatory gene induction necessitates canonical NF-κB activation through TBK1. *J. Virol.* 88, 10 (May 2014), 5328–5341.

[2] K Aoki, M Aoki, M Sugai, N Harada, H Miyoshi, T Tsukamoto, T Mizoshita, M Tatematsu, H Seno, T Chiba, M Oshima, C-L Hsieh, and M M Taketo. 2007. Chromosomal instability by β-catenin/TCF transcription in APC or β-catenin mutant cells. , 3511–3520 pages.

[3] Samuel F Bakhoum and Lewis C Cantley. 2018. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. *Cell* 174, 6 (Sept. 2018), 1347–1360.

[4] Samuel F Bakhoum and Duane A Compton. 2012. Chromosomal instability and cancer: a complex relationship with therapeutic potential. , 1138–1143 pages.

[5] Samuel F Bakhoum, Lilian Kabeche, Matthew D Wood, Christopher D Laucius, Dian Qu, Ashley M Laughney, Gloria E Reynolds, Raymond J Louie, Joanna Phillips, Denise A Chan, Bassem I Zaki, John P Murnane, Claudia Petritsch, and Duane A Compton. 2015. Numerical chromosomal instability mediates susceptibility to radiation treatment. *Nat. Commun.* 6 (Jan. 2015), 5990.

[6] Samuel F Bakhoum, Bryan Ngo, Ashley M Laughney, Julie-Ann Cavallo, Charles J Murphy, Peter Ly, Pragya Shah, Roshan K Sriram, Thomas B K Watkins, Neil K Taunk, Mercedes Duran, Chantal Pauli, Christine Shaw, Kalyani Chadalavada, Vinagolu K Rajasekhar, Giulio Genovese, Subramanian Venkatesan, Nicolai J Birkbak, Nicholas McGranahan, Mark Lundquist, Quincey LaPlant, John H Healey, Olivier Elemento, Christine H Chung, Nancy Y Lee, Marcin Imielenski, Gouri Nanjangud, Dana Pe'er, Don W Cleveland, Simon N Powell, Jan Lammerding, Charles Swanton, and Lewis C Cantley. 2018. Chromosomal instability drives metastasis through a cytosolic DNA response. *Nature* 553, 7689 (Jan. 2018), 467–472.

[7] Bjorn Bakker, Aaron Taudt, Mirjam E Belderbos, David Porubsky, Diana C J Spierings, Tristan V de Jong, Nancy Halsema, Hinke G Kazemier, Karina Hoekstra-Wakker, Allan Bradley, Eveline S J M de Bont, Anke van den Berg, Victor Guryev, Peter M Lansdorp, Maria Colomé-Tatché, and Floris Foijer. 2016. Single-cell sequencing reveals karyotype heterogeneity in murine and human malignancies. *Genome Biol.* 17, 1 (May 2016), 115.

[8] Christopher Buccitelli, Lorena Salgueiro, Konstantina Rowald, Rocio Sotillo, Balca R Mardin, and Jan O Korbel. 2017. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res.* 27, 4 (April 2017), 501–511.

[9] Aurora A Burds, Annegret Schulze Lutum, and Peter K Sorger. 2005. Generating chromosome instability through the simultaneous deletion of Mad2 and p53. *Proc. Natl. Acad. Sci. U. S. A.* 102, 32 (Aug. 2005), 11296–11301.

[10] Israel Cañadas, Rohit Thummalapalli, Jong Wook Kim, Shunsuke Kitajima, Russell William Jenkins, Camilla Laulund Christensen, Marco Campisi, Yanan Kuang, Yanxi Zhang, Evisa Gjini, Gao Zhang, Tian Tian, Debattama Rai Sen, Diana Miao, Yu Imamura, Tran Thai, Brandon Piel, Hideki Terai, Amir Reza Aref, Timothy Hagan, Shohei Koyama, Masayuki Watanabe, Hideo Baba, Anika Elise Adeni, Christine Anne Lydon, Pablo Tamayo, Zhi Wei, Meenhard Herlyn, Thanh Uyen Barbie, Ravindra Uppaluri, Lynnette Marie Sholl, Ewa Sicinska, Jacob Sands, Scott Rodig, Kwok Kin Wong, Cloud Peter Paweletz, Hideo Watanabe, and David Allen Barbie. 2018. Tumor innate immunity primed by specific interferon-stimulated endogenous retroviruses. *Nat. Med.* 24, 8 (Aug. 2018), 1143–1150.

[11] Scott L Carter, Kristian Cibulskis, Elena Helman, Aaron McKenna, Hui Shen, Travis Zack, Peter W Laird, Robert C Onofrio, Wendy Winckler, Barbara A Weir, Rameen Beroukhim, David Pellman, Douglas A Levine, Eric S Lander, Matthew Meyerson, and Gad Getz. 2012. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30, 5 (May 2012), 413–421.

[12] Scott L Carter, Aron C Eklund, Isaac S Kohane, Lyndsay N Harris, and Zoltan Szallasi. 2006. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* 38, 9 (Sept. 2006), 1043–1048.

[13] Feng Zhen Chen, Li Jin You, Fan Yang, Li Na Wang, Xue Qin Guo, Fei Gao, Cong Hua, Cong Tan, Lin Fang, Ri Qiang Shan, Wen Jun Zeng, Bo Wang, Ren Wang, Xun Xu, and Xiao Feng Wei. 2020. CNGBdb: China National GeneBank DataBase.

Yi Chuan 42, 8 (Aug. 2020), 799–809.

[14] Yunfei Chen, Lufan Wang, Jiali Jin, Yi Luan, Cong Chen, Yu Li, Hongshang Chu, Xinbo Wang, Guanghong Liao, Yue Yu, Hongqi Teng, Yanming Wang, Weijuan Pan, Lan Fang, Lujian Liao, Zhengfan Jiang, Xin Ge, Bin Li, and Ping Wang. 2017. p38 inhibition provides anti-DNA virus immunity by regulation of USP21 phosphorylation and STING activation. *J. Exp. Med.* 214, 4 (April 2017), 991–1010.

[15] Suet F Chin, Andrew E Teschendorff, John C Marioni, Yanzhong Wang, Nuno L Barbosa-Morais, Natalie P Thorne, Jose L Costa, Sarah E Pinder, Mark A van de Wiel, Andrew R Green, Ian O Ellis, Peggy L Porter, Simon Tavaré, James D Brenton, Bauke Ylstra, and Carlos Caldas. 2007. High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.* 8, 10 (2007), R215.

[16] Woosung Chung, Hye Hyeon Eum, Hae-Ock Lee, Kyung-Min Lee, Han-Byoel Lee, Kyu-Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, Zhengyan Kan, Wonshik Han, and Woong-Yang Park. 2017. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* 8 (May 2017), 15081.

[17] Karen Crasta, Neil J Ganem, Regina Dagher, Alexandra B Lantermann, Elena V Ivanova, Yunfeng Pan, Luigi Nezi, Alexei Protopopov, Dipanjan Chowdhury, and David Pellman. 2012. DNA breaks and chromosome pulverization from errors in mitosis. *Nature* 482, 7383 (Jan. 2012), 53–58.

[18] Spyros Darmanis, Steven A Sloan, Derek Croote, Marco Mignardi, Sophia Chernikova, Peyman Samghababi, Ye Zhang, Norma Neff, Mark Kowarsky, Christine Caneda, Gordon Li, Steven D Chang, Ian David Connolly, Yingmei Li, Ben A Barres, Melanie Hayden Gephart, and Stephen R Quake. 2017. Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. , 1399–1410 pages.

[19] Teresa Davoli, Hajime Uno, Eric C Wooten, and Stephen J Elledge. 2017. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* 355, 6322 (Jan. 2017).

[20] Zhixun Dou, Kanad Ghosh, Maria Grazia Vizioli, Jiajun Zhu, Payel Sen, Kirk J Wangensteen, Johayra Simithy, Yemin Lan, Yanping Lin, Zhuo Zhou, Brian C Capell, Caiyue Xu, Mingang Xu, Julia E Kieckhaefer, Tianying Jiang, Michal Shoshkes-Carmel, K M Ahasan Al Tanim, Glen N Barber, John T Seykora, Sarah E Millar, Klaus H Kaestner, Benjamin A Garcia, Peter D Adams, and Shelley L Berger. 2017. Cytoplasmic chromatin triggers inflammation in senescence and cancer. *Nature* 550, 7676 (Oct. 2017), 402–406.

[21] P Duesberg, R Stindl, and R Hehlmann. 2000. Explaining the high mutation rates of cancer cells to drug and multidrug resistance by chromosome reassortments that are catalyzed by aneuploidy. , 14295–14300 pages.

[22] Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* 21, 16 (Aug. 2005), 3439–3440.

[23] Ron Edgar, Michael Domrachev, and Alex E Lash. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 1 (Jan. 2002), 207–210.

[24] David Endesfelder, Rebecca Burrell, Nnennaya Kanu, Nicholas McGranahan, Mike Howell, Peter J Parker, Julian Downward, Charles Swanton, and Maik Kschischo. 2014. Chromosomal instability selects gene copy-number variants encoding core regulators of proliferation in ER+ breast cancer. *Cancer Res.* 74, 17 (Sept. 2014), 4853–4863.

[25] Jean Fan, Hae-Ock Lee, Soohyun Lee, Da-Eun Ryu, Semin Lee, Catherine Xue, Seok Jin Kim, Kihyun Kim, Nikolaos Barkas, Peter J Park, Woong-Yang Park, and Peter V Kharchenko. 2018. Linking transcriptional and genetic tumor heterogeneity through allele analysis of single-cell RNA-seq data. *Genome Res.* 28, 8 (Aug. 2018), 1217–1227.

[26] Rudolf S N Fehrmann, Juha M Karjalainen, Małgorzata Krajewska, Harm-Jan Westra, David Maloney, Anton Simeonov, Tune H Pers, Joel N Hirschhorn, Ritsert C Jansen, Erik A Schultes, Herman H H B M van Haagen, Elisabeth G E de Vries, Gerard J te Meerman, Cisca Wijmenga, Marcel A T M van Vugt, and Lude Franke. 2015. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* 47, 2 (Feb. 2015), 115–125.

[27] Ruli Gao, Shanshan Bai, Ying C Henderson, Yiyun Lin, Aislyn Schalck, Yun Yan, Tapsi Kumar, Min Hu, Emi Sei, Alexander Davis, Fang Wang, Simona F Shaitelman, Jennifer Rui Wang, Ken Chen, Stacy Moulder, Stephen Y Lai, and Nicholas E Navin. 2021. Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nat. Biotechnol.* 39, 5 (May 2021), 599–608.

[28] Alice Giustacchini, Supat Thongjuea, Nikolaos Barkas, Petter S Woll, Benjamin J Povinelli, Christopher A G Booth, Paul Sopp, Ruggiero Norfo, Alba Rodriguez-Meira, Neil Ashley, Lauren Jamieson, Paresh Vyas, Kristina Anderson, Åsa Segerstolpe, Hong Qian, Ulla Olsson-Strömberg, Satu Mustjoki, Rickard Sandberg, Sten Eirik W Jacobsen, and Adam J Mead. 2017. Single-cell transcriptomics uncovers distinct molecular signatures of stem cells in chronic myeloid leukemia. *Nat. Med.* 23, 6 (June 2017), 692–702.

[29] Michel V Hadjihannas, Martina Brückner, Boris Jerchow, Walter Birchmeier, Wolfgang Dietmaier, and Jürgen Behrens. 2006. Aberrant Wnt/beta-catenin signaling can induce chromosomal instability in colon cancer. *Proc. Natl. Acad. Sci. U. S. A.* 103, 28 (July 2006), 10747–10752.

[30] Emily M Hatch, Andrew H Fischer, Thomas J Deerinck, and Martin W Hetzer. 2013. Catastrophic nuclear envelope collapse in cancer cell micronuclei. *Cell* 154, 1 (July 2013), 47–60.

[31] Aniek Janssen, Geert J P L Kops, and René H Medema. 2009. Elevating the frequency of chromosome mis-segregation as a strategy to kill tumor cells. *Proc. Natl. Acad. Sci. U. S. A.* 106, 45 (Nov. 2009), 19108–19113.

[32] Mihriban Karaayvaz, Simona Cristea, Shawn M Gillespie, Anoop P Patel, Ravindra Mylvaganam, Christina C Luo, Michelle C Specht, Bradley E Bernstein, Franziska Michor, and Leif W Ellisen. 2018. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* 9, 1 (Sept. 2018), 3588.

[33] Hiroyasu Konno, Shota Yamauchi, Anders Berglund, Ryan M Putney, James J Mulé, and Glen N Barber. 2018. Suppression of STING signaling through epigenetic silencing and missense mutation impedes DNA damage mediated cytokine production. *Oncogene* 37, 15 (April 2018), 2037–2051.

[34] Gennady Korotkevich, Vladimir Sukhov, and Alexey Sergushichev. 2019. Fast gene set enrichment analysis. *bioRxiv* (2019). https://doi.org/10.1101/060012

[35] L Lau, E E Gray, R L Brunette, and D B Stetson. 2015. DNA tumor virus oncogenes antagonize the cGAS-STING DNA-sensing pathway. , 568–571 pages.

[36] Alvin J X Lee, David Endesfelder, Andrew J Rowan, Axel Walther, Nicolai J Birkbak, P Andrew Futreal, Julian Downward, Zoltan Szallasi, Ian P M Tomlinson, Michael Howell, Maik Kschischo, and Charles Swanton. 2011. Chromosomal instability confers intrinsic multidrug resistance. *Cancer Res.* 71, 5 (March 2011), 1858–1870.

[37] Hae-Ock Lee, Yourae Hong, Hakki Emre Etlioglu, Yong Beom Cho, Valentina Pomella, Ben Van den Bosch, Jasper Vanhecke, Sara Verbandt, Hyekyung Hong, Jae-Woong Min, Nayoung Kim, Hye Hyeon Eum, Junbin Qian, Bram Boeckx, Diether Lambrechts, Petros Tsantoulis, Gert De Hertogh, Woosung Chung, Taeseob Lee, Minae An, Hyun-Tae Shin, Je-Gun Joung, Min-Hyeok Jung, Gunhwan Ko, Pratyaksha Wirapati, Seok Hyung Kim, Hee Cheol Kim, Seong Hyeon Yun, Iain Bee Huat Tan, Bobby Ranjan, Woo Yong Lee, Tae-You Kim, Jung Kyoon Choi, Young-Joon Kim, Shyam Prabhakar, Sabine Tejpar, and Woong-Yang Park. 2020. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. *Nat. Genet.* 52, 6 (June 2020), 594–603.

[38] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P Mesirov, and Pablo Tamayo. 2015. The Molecular Signatures Database Hallmark Gene Set Collection. , 417–425 pages.

[39] Michael I Love, Wolfgang Huber, and Simon Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 12 (2014), 550.

[40] Karen J Mackenzie, Paula Carroll, Carol-Anne Martin, Olga Murina, Adeline Fluteau, Daniel J Simpson, Nelly Olova, Hannah Sutcliffe, Jacqueline K Rainger, Andrea Leitch, Ruby T Osborn, Ann P Wheeler, Marcin Nowotny, Nick Gilbert, Tamir Chandra, Martin A M Reijns, and Andrew P Jackson. 2017. cGAS surveillance of micronuclei links genome instability to innate immunity. *Nature* 548, 7668 (Aug. 2017), 461–465.

[41] Sarah E McClelland, Rebecca A Burrell, and Charles Swanton. 2009. Chromosomal instability: a composite phenotype that influences sensitivity to chemotherapy. *Cell Cycle* 8, 20 (Oct. 2009), 3262–3266.

[42] Antje Menssen, Alexey Epanchintsev, Dmitri Lodygin, Nousin Rezaei, Peter Jung, Berlinda Verdoodt, Joachim Diebold, and Heiko Hermeking. 2007. c-MYC Delays Prometaphase by Direct Transactivation of MAD2 and BubR1: Identification of Mechanisms Underlying c-MYC-Induced DNA Damage and Chromosomal Instability. , 339–352 pages.

[43] Louisa Nelson, Anthony Tighe, Anya Golder, Samantha Littler, Bjorn Bakker, Daniela Moralli, Syed Murtuza Baker, Ian J Donaldson, Diana C J Spierings, René Wardenaar, Bethanie Neale, George J Burghel, Brett Winter-Roach, Richard Edmondson, Andrew R Clamp, Gordon C Jayson, Sudha Desai, Catherine M Green, Andy Hayes, Floris Foijer, Robert D Morgan, and Stephen S Taylor. 2020. A living biobank of ovarian cancer ex vivo models reveals profound mitotic heterogeneity. *Nat. Commun.* 11, 1 (Feb. 2020), 822.

[44] Martin A Nowak, Natalia L Komarova, Anirvan Sengupta, Prasad V Jallepalli, Ie-Ming Shih, Bert Vogelstein, and Christoph Lengauer. 2002. The role of chromosomal instability in tumor initiation. *Proc. Natl. Acad. Sci. U. S. A.* 99, 25 (Dec. 2002), 16226–16231.

[45] Irene Papatheodorou, Pablo Moreno, Jonathan Manning, Alfonso Muñoz-Pomer Fuentes, Nancy George, Silvie Fexova, Nuno A Fonseca, Anja Füllgrabe, Matthew Green, Ni Huang, Laura Huerta, Haider Iqbal, Monica Jianu, Suhaib Mohammed, Lingyun Zhao, Andrew F Jarnuczak, Simon Jupp, John Marioni, Kerstin Meyer, Robert Petryszak, Cesar Augusto Prada Medina, Carlos Talavera-López, Sarah Teichmann, Juan Antonio Vizcaino, and Alvis Brazma. 2020. Expression Atlas update: from tissues to single cells. *Nucleic Acids Res.* 48, D1 (Jan. 2020), D77–D83.

[46] Sidharth V Puram, Itay Tirosh, Anuraag S Parikh, Anoop P Patel, Keren Yizhak, Shawn Gillespie, Christopher Rodman, Christina L Luo, Edmund A Mroz, Kevin S

Emerick, Daniel G Deschler, Mark A Varvares, Ravi Mylvaganam, Orit Rozenblatt-Rosen, James W Rocco, William C Faquin, Derrick T Lin, Aviv Regev, and Bradley E Bernstein. 2017. Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell* 171, 7 (Dec. 2017), 1611–1624.e24.

[47] Julia Rohrberg, Daniel Van de Mark, Meelad Amouzgar, Joyce V Lee, Moufida Taileb, Alexandra Corella, Seda Kilinc, Jeremy Williams, Marie-Lena Jokisch, Roman Camarda, Sanjeev Balakrishnan, Rama Shankar, Alicia Zhou, Aaron N Chang, Bin Chen, Hope S Rugo, Sophie Dumont, and Andrei Goga. 2020. MYC Dysregulates Mitosis, Revealing Cancer Vulnerabilities. *Cell Rep.* 30, 10 (March 2020), 3368–3382.e7.

[48] Konstantina Rowald, Martina Mantovan, Joana Passos, Christopher Buccitelli, Balca R Mardin, Jan O Korbel, Martin Jechlinger, and Rocio Sotillo. 2016. Negative Selection and Chromosome Instability Induced by Mad2 Overexpression Delay Breast Cancer but Facilitate Oncogene-Independent Outgrowth. *Cell Rep.* 15, 12 (June 2016), 2679–2691.

[49] Stefano Santaguida, Amelia Richardson, Divya Ramalingam Iyer, Ons M'Saad, Lauren Zasadil, Kristin A Knouse, Yao Liang Wong, Nicholas Rhind, Arshad Desai, and Angelika Amon. 2017. Chromosome Mis-segregation Generates Cell-Cycle-Arrested Cells with Complex Karyotypes that Are Eliminated by the Immune System. *Dev. Cell* 41, 6 (June 2017), 638–651.e5.

[50] Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* 33, 5 (May 2015), 495–502.

[51] Juan-Manuel Schvartzman, Pascal H G Duijf, Rocio Sotillo, Courtney Coker, and Robert Benezra. 2011. Mad2 is a critical mediator of the chromosome instability observed upon Rb and p53 pathway inhibition. *Cancer Cell* 19, 6 (June 2011), 701–714.

[52] Juan-Manuel Schvartzman, Rocio Sotillo, and Robert Benezra. 2010. Mitotic chromosomal instability and cancer: mouse modelling of the human disease. *Nat. Rev. Cancer* 10, 2 (Feb. 2010), 102–115.

[53] Laura Senovilla, Ilio Vitale, Isabelle Martins, Maximilien Tailler, Claire Pailleret, Mickaël Michaud, Lorenzo Galluzzi, Sandy Adjemian, Oliver Kepp, Mireia Niso-Santano, Shensi Shen, Guillermo Mariño, Alfredo Criollo, Alice Boilève, Bastien Job, Sylvain Ladoire, François Ghiringhelli, Antonella Sistigu, Takahiro Yamazaki, Santiago Rello-Varona, Clara Locher, Vichnou Poirier-Colame, Monique Talbot, Alexander Valent, Francesco Berardinelli, Antonio Antoccia, Fabiola Ciccosanti, Gian Maria Fimia, Mauro Piacentini, Antonio Fueyo, Nicole L Messina, Ming Li, Christopher J Chan, Verena Sigl, Guillaume Pourcher, Christoph Ruckenstuhl, Didac Carmona-Gutierrez, Vladimir Lazar, Josef M Penninger, Frank Madeo, Carlos López-Otín, Mark J Smyth, Laurence Zitvogel, Maria Castedo, and Guido Kroemer. 2012. An immunosurveillance mechanism controls cancer cell ploidy. *Science* 337, 6102 (Sept. 2012), 1678–1684.

[54] Jason M Sheltzer. 2013. A transcriptional and metabolic signature of primary aneuploidy is present in chromosomally unstable cancer cells and informs clinical prognosis. *Cancer Res.* 73, 21 (Nov. 2013), 6401–6412.

[55] Jason M Sheltzer and Angelika Amon. 2011. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends Genet.* 27, 11 (Nov. 2011), 446–453.

[56] Jason M Sheltzer, Julie H Ko, John M Replogle, Nicole C Habibe Burgos, Erica S Chung, Colleen M Meehl, Nicole M Sayles, Verena Passerini, Zuzana Storchova, and Angelika Amon. 2017. Single-chromosome Gains Commonly Function as Tumor Suppressors. *Cancer Cell* 31, 2 (Feb. 2017), 240–255.

[57] Judith E Simon, Bjorn Bakker, and Floris Foijer. 2015. CINcere Modelling: What Have Mouse Models for Chromosome Instability Taught Us? *Recent Results Cancer Res.* 200 (2015), 39–60.

[58] Shushu Song, Peike Peng, Zhaoqing Tang, Junjie Zhao, Weicheng Wu, Haojie Li, Miaomiao Shao, Lili Li, Caiting Yang, Fangfang Duan, Mingming Zhang, Jie Zhang, Hao Wu, Can Li, Xuefei Wang, Hongshan Wang, Yuanyuan Ruan, and Jianxin Gu. 2017. Decreased expression of STING predicts poor prognosis in patients with gastric cancer. *Sci. Rep.* 7 (Feb. 2017), 39858.

[59] Yunfan Sun, Liang Wu, Yu Zhong, Kaiqian Zhou, Yong Hou, Zifei Wang, Zefan Zhang, Jiarui Xie, Chunqing Wang, Dandan Chen, Yaling Huang, Xiaochan Wei, Yinghong Shi, Zhikun Zhao, Yuehua Li, Ziwei Guo, Qichao Yu, Liqin Xu, Giacomo Volpe, Shuangjian Qiu, Jian Zhou, Carl Ward, Huichuan Sun, Ye Yin, Xun Xu, Xiangdong Wang, Miguel A Esteban, Huanming Yang, Jian Wang, Michael Dean, Yaguang Zhang, Shiping Liu, Xinrong Yang, and Jia Fan. 2021. Single-cell landscape of the ecosystem in early-relapse hepatocellular carcinoma. *Cell* 184, 2 (Jan. 2021), 404–421.e16.

[60] Sarah L Thompson, Samuel F Bakhoum, and Duane A Compton. 2010. Mechanisms of Chromosomal Instability. , R285–R295 pages.

[61] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, 2nd, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S Genshaft, Travis K Hughes, Carly G K Ziegler, Samuel W Kazer, Aleth Gaillard, Kellie E Kolb, Alexandra-Chloé Villani, Cory M Johannessen, Aleksandr Y Andreev, Eliezer M Van Allen, Monica Bertagnolli, Peter K Sorger, Ryan J Sullivan, Keith T Flaherty, Dennie T Frederick, Judit Jané-Valbuena, Charles H Yoon, Orit Rozenblatt-Rosen, Alex K Shalek, Aviv Regev, and Levi A Garraway. 2016. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 6282 (April 2016), 189–196.

[62] Thomas B K Watkins, Emilia L Lim, Marina Petkovic, Sergi Elizalde, Nicolai J Birkbak, Gareth A Wilson, David A Moore, Eva Grönroos, Andrew Rowan, Sally M Dewhurst, Jonas Demeulemeester, Stefan C Dentro, Stuart Horswell, Lewis Au, Kerstin Haase, Mickael Escudero, Rachel Rosenthal, Maise Al Bakir, Hang Xu, Kevin Litchfield, Wei Ting Lu, Thanos P Mourikis, Michelle Dietzen, Lavinia Spain, George D Cresswell, Dhruva Biswas, Philippe Lamy, Iver Nordentoft, Katja Harbst, Francesc Castro-Giner, Lucy R Yates, Franco Caramia, Fanny Jaulin, Cécile Vicier, Ian P M Tomlinson, Priscilla K Brastianos, Raymond J Cho, Boris C Bastian, Lars Dyrskjøt, Göran B Jönsson, Peter Savas, Sherene Loi, Peter J Campbell, Fabrice Andre, Nicholas M Luscombe, Neeltje Steeghs, Vivianne C G Tjan-Heijnen, Zoltan Szallasi, Samra Turajlic, Mariam Jamal-Hanjani, Peter Van Loo, Samuel F Bakhoum, Roland F Schwarz, Nicholas McGranahan, and Charles Swanton. 2020. Pervasive chromosomal instability and karyotype order in tumour evolution. *Nature* 587, 7832 (Nov. 2020), 126–132.

[63] Seng-Ryong Woo, Mercedes B Fuertes, Leticia Corrales, Stefani Spranger, Michael J Furdyna, Michael Y K Leung, Ryan Duggan, Ying Wang, Glen N Barber, Katherine A Fitzgerald, Maria-Luisa Alegre, and Thomas F Gajewski. 2014. STING-dependent cytosolic DNA sensing mediates innate immune recognition of immunogenic tumors. *Immunity* 41, 5 (Nov. 2014), 830–842.

[64] Tianli Xia, Hiroyasu Konno, Jeonghyun Ahn, and Glen N Barber. 2016. Deregulation of STING Signaling in Colorectal Carcinoma Constrains DNA Damage Responses and Correlates With Tumorigenesis. *Cell Rep.* 14, 2 (Jan. 2016), 282–297.

[65] Bassem I Zaki, Arief A Suriawinata, Alan R Eastman, Kristen M Garner, and Samuel F Bakhoum. 2014. Chromosomal instability portends superior response of rectal adenocarcinoma to chemoradiation therapy. *Cancer* 120, 11 (June 2014), 1733–1742.