Mapping cell-cell fusion at single-cell resolution

Andrea L. Gardner [1], Lan Zheng[1], Kennedy Howland[1], Andrew Saunders[1], Andrea Ramirez[1], Patrik Parker[1], Chisom Iloegbunam[1], Daylin Morgan[1], Tyler A. Jost[1], Amy Brock[1]

[1]Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA

Abstract

Cell-cell fusion is a tightly controlled process in the human body known to be involved in fertilization, placental development, muscle growth, bone remodeling, and viral response. Fusion between cancer cells results first in a whole-genome doubled state, which may be followed by the generation of aneuploidies; these genomic alterations are known drivers of tumor evolution. The role of cell-cell fusion in cancer progression and treatment response has been understudied due to limited experimental systems for tracking and analyzing individual fusion events. To meet this need, we developed a molecular toolkit to map the origins and outcomes of individual cell fusion events within a tumor cell population. This platform, ClonMapper Duo ('CMDuo'), identifies cells that have undergone cell-cell fusion through a combination of reporter expression and engineered fluorescence-associated index sequences paired to randomly generated nucleotide barcodes. scRNA-seq of the indexed barcodes enables the mapping of each set of parental cells and fusion progeny throughout the cell population. In triple-negative breast cancer cells CMDuo uncovered subclonal transcriptomic hybridization and unveiled distinct cell-states which arise in direct consequence of homotypic cell-cell fusion. CMDuo is a platform that enables mapping of cell-cell fusion events in high-throughput single cell data and enables the study of cell fusion in disease progression and therapeutic response.

Introduction

Intratumoral heterogeneity (ITH) is a defining feature of many tumor types which negatively affects prognosis[1–3]. Subclonal cancer cell populations within tumors may have unique properties which alter tumor aggressiveness and contribute to treatment failure. Consequently, ITH is especially problematic in cancers such as triple-negative breast cancer (TNBC) which do not have targeted treatment options and for which broad-spectrum chemotherapies remain the standard of care[4–6]. To develop more effective treatments for TNBC there is an urgent need to better understand the molecular drivers of ITH and tumor evolution.

Cell-cell fusion is a form of whole genome doubling (WGD), a known driver of ITH[7–13]. Novel methods have been developed to reconstruct WGD-aware subclonal phylogenies from patient data and more than 45% of breast cancers analyzed in the TCGA dataset show evidence of prior WGD[7–12]. While there is robust data to support WGD in tumor evolution and ITH generation, these phylogenetic reconstruction methods assume asexuality of tumor cells which may not be fully accurate[13,14]. Though understudied, the ability of cancer cells to fuse with other cells in the tumor microenvironment has been recognized for over a century[13,15–24], with recent studies suggesting roles for cell-cell fusion in cancer stemness, metastatic potential, karyotype changes, and ITH[13,17,18,21,25,26]. Recent advancements in live-cell imaging have generated increased interest in cell-cell fusion, but no tools exist to study and track cell-cell fusion events in high-throughput.

To meet this need, ClonMapper Duo ("CMDuo") was developed to allow clonally-resolved tracking of cell-cell fusion at single-cell resolution. In this study, we constructed high-resolution maps of cell-cell fusion events within two TNBC cell lines, identified unique cell states driven by cell-cell fusion, and linked transcriptomic features of parents to progeny at single-cell resolution.

Results

ClonMapper Duo is a two-color, two-index expressed DNA barcoding platform for tracking cell-cell fusion at single-cell resolution

CMDuo barcoded cells contain DNA barcodes that are transcribed as both polyadenylated transcripts and as CRISPR-compatible sgRNAs (Fig. 1a), allowing assignment to individual cells in scRNA-seq. To track cell-cell fusion using common live-cell imaging systems, CMDuo was engineered to co-express barcodes with nuclear-localized histone H2B-GFP ("CMDuo-green") or H2B-mCherry ("CMDuo-red") (Fig. 1a). Adding nuclear localization to the fluorescent signal allows precise cell counting and quantification from live cell imaging, and specific localization to H2B keeps fluorescence associated with parental nuclei in a shared cytoplasmic space until nuclear fusion takes place[27] (Fig. 1b-c). The identification of fusion cells in barcoded cell populations is a challenge as cell doublets, common in many scRNA-seq preparations, or cells with multiple viral integrations also appear as multi-barcoded cells in scRNA-seq[28,29]. Given the rarity of fusion events compared to the rate of doublets in scRNA-seq[13,28,30], potential fusion cells are likely filtered out and lost from analysis. To overcome this challenge, CMDuo was designed to confidently assign individual cells as fusion progeny in scRNA-seq using a combination of FACS enrichment on fluorescent identity and engineered fluorescence-associated index sequences paired to each random barcode sequence. For each fluorescent protein, four unique 5-nt barcode index sequences were designed by maximizing Hamming distance between all pairs[31] (Supp. Fig. 1, Supp. Table 1). With the complete CMDuo platform, parental cells will express either GFP with a single CMDuo-

green barcode or mCherry with a CMDuo-red barcode, while fusion cells will express both GFP and mCherry and contain both a CMDuo-green and CMDuo-red barcode (Fig. 1b-c).

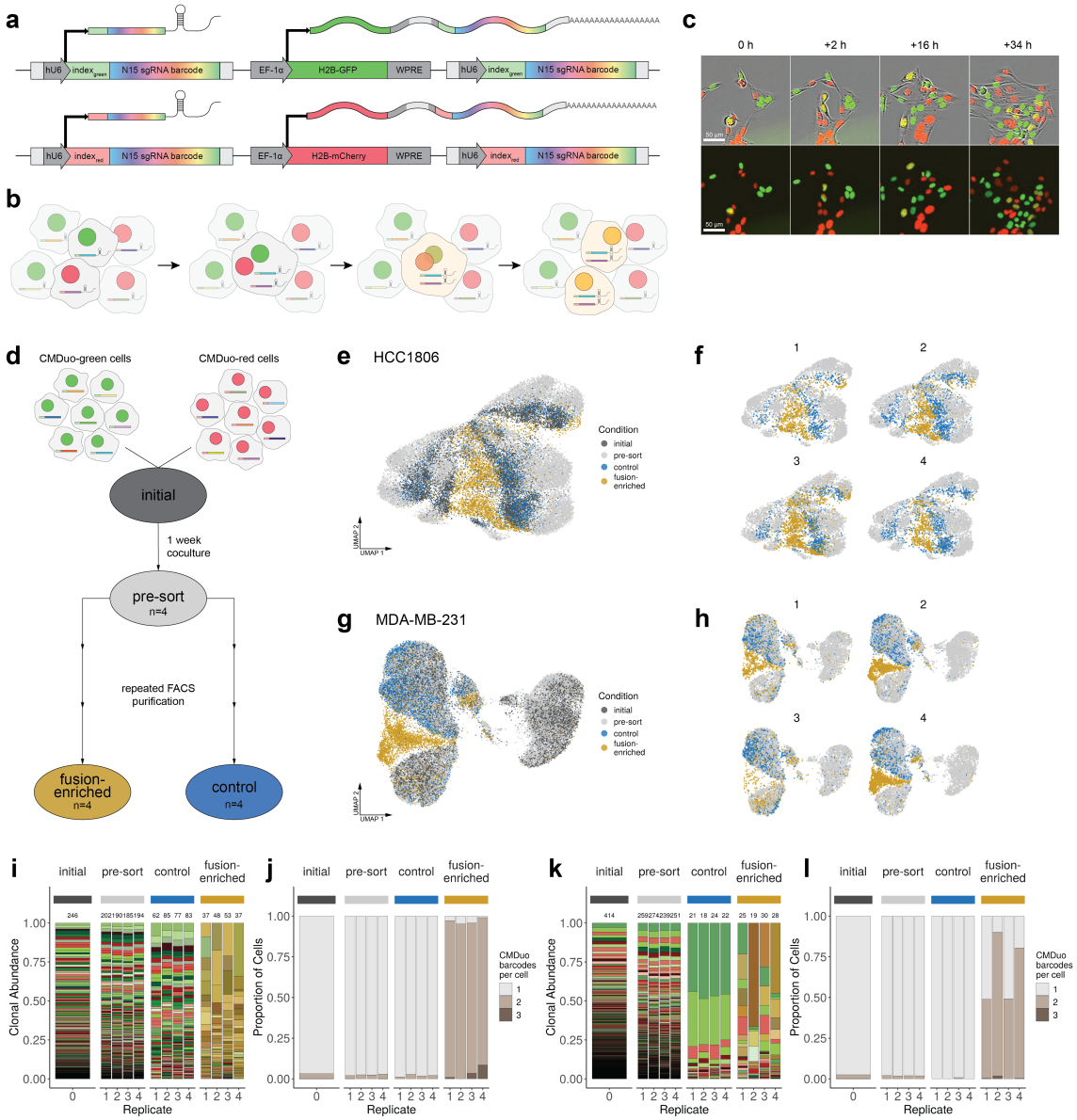

Figure 1. ClonMapper Duo is a two-color, two-index expressed DNA barcoding platform which enables tracking of cell-cell fusion events in scRNA-seq and live-cell imaging. (a) Schematic showing the core features of the CMDuo system. Cells with the integrated transgene express a barcode composed of a degenerate N(15) sequence with a 5-nt index sequence corresponding to the expressed fluorescent protein. The vector expresses the barcode as both a CRISPR-compatible sgRNA and a polyadenylated transcript. (b) Schematic following barcoded cells through cell-cell fusion, genomic recombination and division. Cells generated from cell-cell fusion of CMDuo-green and CMDuo-red cells will express barcodes and barcode-

associated fluorescent proteins from both parental cells. (c) 10X live-cell imaging capture of a spontaneous cell-cell fusion event between CMDuo-red and CMDuo-green HCC1806 cells (additional images in Extended Data. Fig. 1). (d) 1:1 mixtures of homotypic CMDuo-green and CMDuo-red barcoded HCC1806 or MDA-MB-231 cells were cocultured in 5 independent replicates. One replicate was analyzed by scRNA-seq ("initial") 2 days after plating. The remaining 4 replicates stayed in coculture for 1 week ("pre-sort") before undergoing multiple rounds of FACS sorting to enrich for double-positive fusion events ("fusion-enriched") or single positive ("control") cells from each replicate for each cell line. (e) UMAP representation of scRNA-seq for all cells in the HCC1806 data set colored by sample and (f) shown for each replicate. (g) UMAP representation of scRNA-seq for all cells in the MDA-MB-231 data set colored by sample and (h) shown for each replicate. (i) Distribution of clones and (j) the number of barcodes assigned to each cell within each sample for HCC1806. Here a unique green, red, or gold/brown was assigned to each CMDuo-green, CMDuo-red, or fusion clone, respectively. (k) Distribution of clones and (l) the number of barcodes assigned to each cell within each sample for MDA-MB-231.

ClonMapper Duo allows clonal population tracking and confident assignment of multi-barcoded fusion cells

To track spontaneous homotypic cell-cell fusion from parental clones to progeny in triple negative breast cancer cells, the CMDuo platform was applied as in Fig. 1d for two cell lines (HCC1806, MDA-MB-231) with four parallel replicates each. Homotypic cocultures were created by seeding an equal proportion of CMDuo-green and CMDuo-red barcoded cell libraries across replicate plates ("initial"). 1 week after coculture and expansion ("presort"), cells in each replicate were FACS sorted into two subsets double-positive (GFP^+ $mCherry^+$) fusion-enriched cells and single-positive (GFP^+ or $mCherry^+$) control cells. As no surface markers of fusion events are known to assist in discriminating fusion events from red-green doublets, multiple rounds of FACS were needed to obtain high purity fusion samples. FACS enrichment for fusion and control cells was repeated until fusion-enriched populations were > 50% pure. From fixation of the "initial" cells to the time at which "fusion-enriched" and parallel "control" cells were fixed for scRNA-seq, HCC1806 cells underwent 3 enrichment sorts over 35 days and MDA-MB-231 underwent 4 enrichment sorts over 50 days to achieve the final "fusion-enriched" and "control" samples for scRNA-seq analysis with the purity of each population summarized in Supp. Table 2.

In the initial population of HCC1806 cells, 246 unique clones (136 green and 110 red) were identified with a Shannon diversity index of 5.1 (evenness=0.92). This diversity drops minimally among initial, pre-sort, and control samples in the HCC1806 cells, with fewer barcodes and lower diversity observed in the fusion-enriched samples (Fig. 1i,k, Supp. Table 3). In the initial MDA-MB-231 cell library, 414 unique clones (197 green, 216 red, and 1 whose index sequence could not be confidently assigned) were identified with a Shannon diversity index of 5.4 (evenness=0.90). Unlike the HCC1806 cells, the MDA-MB-231 cells showed a strong loss of diversity in the control cells which had undergone repeated rounds of FACS in parallel with the fusion cells (Fig. 1i,k). The loss of diversity in the MDA-MB-231 cells may be due to clonal growth rate differences over the longer experimental timeline or stronger subclonal responses to repeated cell sorting. The

number of barcodes assigned to each cell within each sample was quantified (Fig. 1j,l) and for all samples, the proportion of cells assigned multiple barcodes by scRNA-seq is well-matched to the proportion of fusion cells quantified via flow cytometry (Supp. Table 4, Supp. Fig. 2)

High-resolution mapping of cell-cell fusion events

To map parental clones to fusion progeny, cells in the fusion-enriched sample containing both red and green barcodes were assigned as confident fusion events when at least 5 clonal cells containing the same barcode set were found in a single replicate. Fusion events will be annotated as outlined in Fig. 2a. 39.4% (97/246) of clones participated in at least one detected fusion event in the HCC1806 population, and 7.7% (32/414) in the MDA-MB-231 population (Fig. 2b-f, shown by replicate for both populations in Extended Data. Fig. 2). To test whether the clones found in the fusion population occur more often than expected by random chance, a binomial test was conducted, weighting the probability of selection by abundance in the initial population. 5.3% (13/246) and 2.1% (9/414) of clones in HCC1806 and MDA-MB-231 appeared more often than expected by random chance in the fusion clones (Fig. 2g-h). These results reject the null hypothesis of random selection and suggest that certain clones are more fusogenic or, if fusion is occurring randomly, that certain clones have innate features which better prime progeny for survival after fusion.

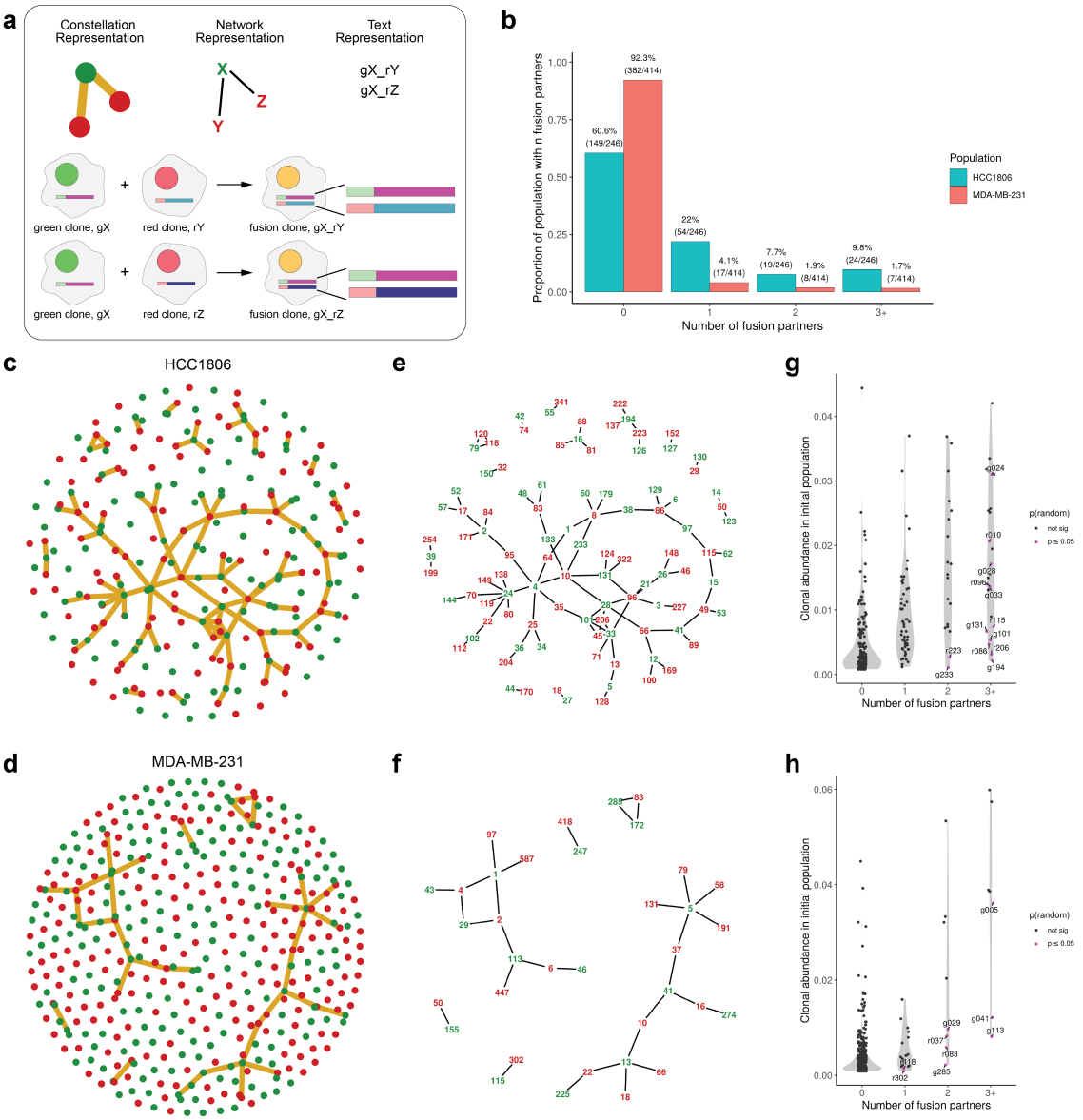

Figure 2. High-resolution mapping of cell-cell fusion events. (a) Legend showing an overview of different representations of fusion events, here we show how a fusion event between green clone 'gX' and two different red clones 'rY' and 'rZ' can be represented as constellation plots, network plots, and in text. (b) Proportion of clones participating in fusion events in each cell line. (c) Constellation plot showing all clones with edges representing a fusion event between clones in the HCC1806 and (d) MDA-MB-231 populations. (e) Network representation highlighting the barcode identifiers for clones involved in cell-cell fusion events for HCC1806 and (f) MDA-MB-231. (g) Quantification of the number of fusion partners for each clone, highlighting clones which participated in fusion events more often than random chance given abundance in the initial population for the HCC1806 and (h) MDA-MB-231 populations.

174 **Parental clones are enriched for actin and actomyosin-related pathways**

175 To investigate whether there are transcriptomic features of parental clones which
176 predict their occurrence in fusion progeny, parental clones which appeared more often in
177 fusion progeny than expected by random chance (Fig. 2g-h, labeled points) to clones
178 which were not found in any fusion progeny for each cell line using a mixed-model
179 analysis. At a log2FC cutoff of 0.2 and adjusted p-value (p-adj) <= 0.05, there were 224
180 upregulated and 324 downregulated genes identified in HCC1806 and 2 up and 1 down
181 in MDA-MB-231. Considering that the minimal number of significant genes in the MDA-
182 MB-231 sample may be driven by lower sample size due to the lower rate and/or capture
183 of fusion events in this cell line, we asked if any of the significant genes from the HCC1806
184 overlap with non-significant genes with a log2FC cutoff of 0.2 in MDA-MB-231 and find 37
185 potential markers (Supp. Table 5, Extended Data Fig. 3a-b). Although no significant genes
186 were shared between cell lines, gene set enrichment analysis identified enrichment for
187 actin and actomyosin-related pathways in both cell lines (Extended Data. Fig. 3c-d).
188

189 **Fusion progeny hybridize parental transcriptomic states and activate new**
190 **programs**

191 Next, we investigated the transcriptomic fate of fusion progeny by first asking if
192 fusion clone clones maintain clone-specific transcriptomic patterns from their parental
193 clones. Marker genes for parental clones in each parent-parent-progeny set were
194 calculated and then their expression was analyzed for all clones within the matched
195 parent-parent-progeny set, finding that fusion clones hybridize marker genes from their
196 parental clones in all examples (Fig. 3a-i, Extended Data Fig. 4-5). To quantify
197 transcriptomic similarity, the Pearson correlation coefficient was calculated for each
198 parent-parent and parent-fusion pair in each parent-parent-fusion set. This revealed that
199 parental clones in both populations are more similar to their fusion progeny than they are
200 to each other (for HCC1806 (n=27) $PCC_{parent\text{-}parent}$ = 0.87, 95% CI [0.82, 0.91], $PCC_{parent\text{-}$
201 $fusion}$ = 0.91, 95% CI [0.87, 0.91], and for MDA-MB-231 (n=12), $PCC_{parent\text{-}parent}$ = 0.88, 95%
202 CI [0.85, 0.95], $PCC_{parent\text{-}fusion}$ = 0.92, 95% CI [0.88, 0.95] (Extended Data. Fig. 6). These
203 results show fusion progeny inherit and hybridize clone-specific transcriptomic patterns
204 from their parental clones.

205 To investigate if novel cell states are activated after cell-cell fusion, differential
206 expression analysis was performed between for each fusion clone against both of its
207 parental clones. Here, in each fusion clone there is a set of genes that become highly
208 active in fusion clones compared to the parental clones. This suggests that in addition to
209 the hybridization of parental genes, the process of cell-cell fusion generates novel cell
210 states.
211
212

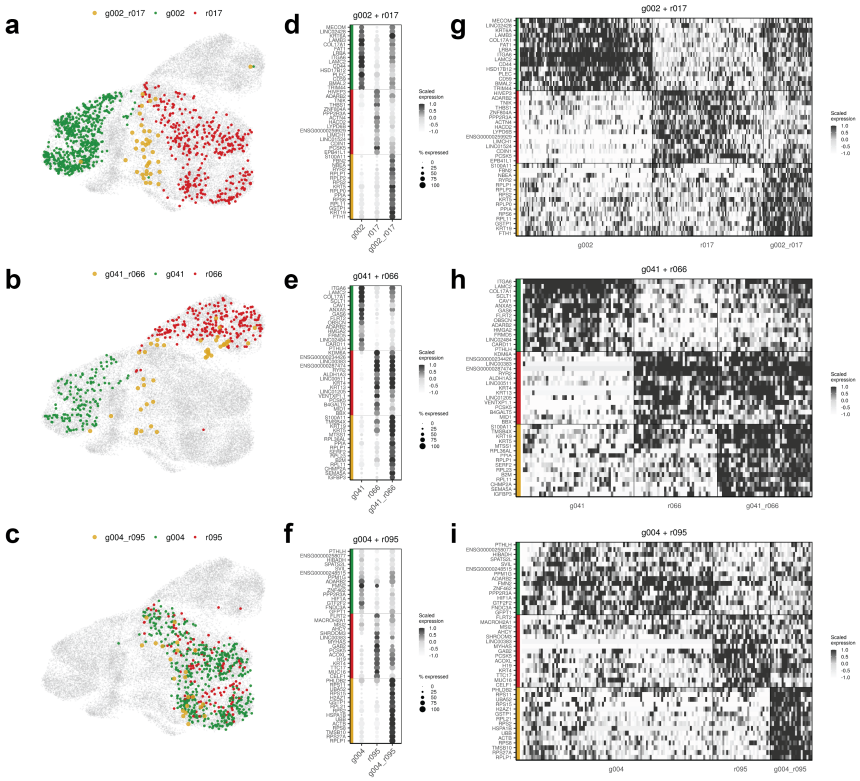

Figure 3. Fusion progeny hybridize parental transcriptomic states and activate new programs . Clone-to-clone analysis highlighting examples of 3 distinct fusion examples in the HCC1806 population: (top) g002+g017, (center) g041+r066, and (bottom) g004+r095. Showing for each (a-c) UMAP representations of each parent-parent-fusion set (d-f) average scaled expression of clone-specific marker genes for each clone in the parent-parent-fusion set where the intensity of the dot represents scaled expression and the size of the dot represents the number of cells of that clone expressing that gene, (g-i) scaled expression of clone-specific marker genes for each clone visualized across cells from each clone.

There is a distinct cell state achieved through cell-cell fusion independent of parental clone transcriptomic identity

Finding that fusion progeny not only hybridize parental cell states, but also activate distinct transcriptomic states at clonal resolution, we next sought to characterize general transcriptomic features of fusion progeny across the population. Cells were clustered on gene expression using the Louvain algorithm and clusters sharing highly similar distributions of barcodes were collapsed (Supp. Fig. 3). This resulted in 7 distinct barcode-informed clusters for the HCC1806 population (2 of 8 Louvain clusters collapsed), and 7 for the MDA-MB-231 population (0 of 7 Louvain clusters collapsed) (Fig. 4a-d, Extended Data. Fig. 7a-d). Marker genes were calculated for each cluster (Fig. 4i, Extended Data Fig. 7i) and are listed in Supp. Tables 6-7.

For each cell line, fusion clones were found in all clusters present in the parallel sorted controls (Fig. 4a-e, Extended Data Fig. 7a-e) but the highest number of cells and

236 unique clones was identified in Cluster F (54.3% of fusion cells and 33.1% of unique fusion
237 clones in HCC1806 and 64.3% of fusion cells and 28.6% of unique fusion clones in MDA-
238 MB-231) (Fig. 4e, Extended Data Fig. 7e). Parental cells were nearly absent in cluster F
239 (0.1% of cells in both cell lines) (Fig. 3f, Extended Data Fig. 7f).
240 Given the distribution of fusion clones across different transcriptomic clusters, we
241 asked if these different cell states were deterministically generated as a function of the
242 transcriptomic identities of parental clones. We find that the parental transcriptomic cluster
243 identity can be passed to fusion progeny if fusion occurs between clones from the same
244 cluster, but the inverse is not true, there is a transcriptomic cluster that cannot be
245 confidently mapped back to parental clone state. While the cluster F cell state is most
246 often achieved through cluster-mismatched clones, it can also be achieved through fusion
247 of any combination of parental cell states within the population (Fig. 4g-h, Extended Data
248 Fig. 7g-h).
249 To determine features that distinguish the fusion-enriched cluster F cell state from
250 other cell states, we identified overlapping differentially expressed genes and enriched
251 gene sets across both cell lines. 466 differentially expressed genes (206 up, 260 down) at
252 abs(log2FC) > 0.2 and p-adj <= 0.05 were identified in cluster F in both cell lines (Extended
253 Data 8a) (genes at abs(log2FC) > 0.35 shown in Table 1. all shared significant genes in
254 Supp. Tables 8-9), showing enrichment for pathways involved in chromatin remodeling,
255 DNA replication/repair, and downregulation of pathways involved in activation of immune
256 response (Extended Data Fig. 8c,e). This significant decrease in interferon signaling and
257 immune signaling pathways in cluster F cells may indicate a key mechanism of survival
258 for fusion progeny within tumor environments.
259
260

Cluster F vs. other clusters	Up	H2AC14, NRG1, HTR7, H4C3, H3C1, DIAPH2, H2AC11, FLRT2, H1-2, NAV3, DOCK4, DOCK10, TMEM117, ADAMTS1, SLC22A3, ADAMTS6, PDE4D, MECOM, SLC7A11, H3C3, STAC, GPR39, INSYN2B, KYNU, IMMP2L, GPAT3, COL13A1, H2BC18, FUT8, PRKCA, SCFD2, LINC01605, MIR924HG, DYNC1I1, NHSL1, NAALADL2, NALCN, H2AC4, NEGR1, SMYD3, CENPP, CRPPA, SERPINE2, ENSG00000287621, SYT1, FAM111B, CNTNAP3P2, DYNC2H1
	Down	L1CAM, ATF3, CREBRF, TCF7L1, TXNIP, ENSG00000259001, IRF1, HEG1, DUSP5, DNAH11, MIR23AHG, CERCAM, TENT5A, PLA2R1, RAB11FIP1, KDM5B, NEDD9, TUFT1, HBP1, OPTN, LINC00370, DNAJC6, HIVEP2, C6orf132, BCAS4, SCD, ARL6IP1, RNF19B, RAB11FIP5, TRIM65, IFIT1, ZNFX1, PECR, SLC22A5, RIPK4, SREBF1, MGAT4A, TNFAIP2, HIP1R, MXI1, SLC1A3, ARID4A, ELOVL7, TOB1, APOL6, CCNG2, KCNQ1OT1, TDRD7, AURKA, MATN2

261 **Table 1.** Top differentially expressed genes in Cluster F in both cell lines (abs(log2FC) > 0.35, p-adj <=0.05).
262

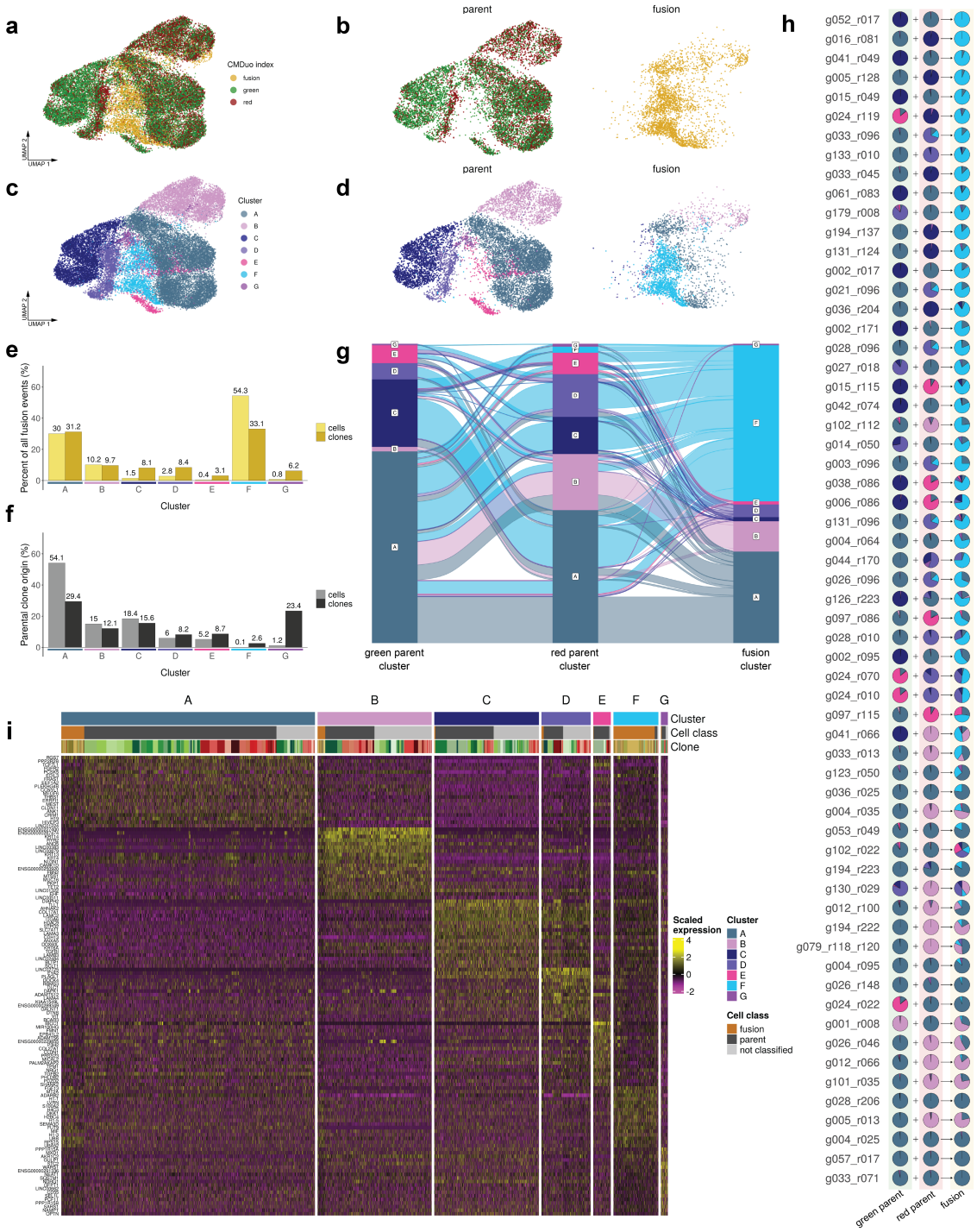

Figure 4. CMDuo reveals a distinct transcriptomic state driven by cell-cell fusion. UMAP representations of all barcoded cells colored by CMDuo index (a-b) or (c-d) transcriptomic cluster in the HCC1806 population (see Extended Data Fig. 7 for parallel analysis on MDA-MB-231). (e) Percent of all confident fusion events in each transcriptomic cluster showing percent of all fusion cells (light) and percent of all fusion clones (dark). (f) Origin of parental cells in each transcriptomic cluster showing percent of parental cells (light) and percent of parental clones (dark). (g) Alluvial plot summarizing the relationships between parental transcriptomic clusters

and fusion progeny transcriptomic cluster. (h) Pie chart representation of the proportion of each clone in each cluster for each set of parent-parent-fusion clones. (i) Heatmap showing marker gene expression for each cluster.

Fusion progeny within parental clusters have transcriptomic signatures distinct from cluster F fusion cells

While the highest number of fusion cells were found in the cluster F, fusion cells in both cell lines were identified in other transcriptomic clusters. To first identify gene expression patterns altered in fusion clones across all clusters, fusion cells were compared to parallel control cells within each cell line. 117 differentially expressed genes were found (78 up, 39 down) (significant genes at abs(log2FC) > 0.35 shown in Table 2, all significant genes in Supp. Tables 10-11), and there is a pattern of an enrichment for E2F targets, reactive oxidative species, oxidative phosphorylation, and DNA repair across all fusion cells (Extended Data Fig. 9).

Given that fusion clones could achieve multiple cell states by maintaining a state similar to their parental identity or activating cluster F programs (Fig. 4h, Extended Data Fig. 7h, Supp. Figs. 4-5), we next asked if fusion cells in cluster F are significantly different from fusion cells in other clusters and revealed striking differences. 585 differentially expressed genes shared in both cell lines (246 up, 339 down) were identified between fusion cells in cluster F and fusion cells in parental clusters (Fig. 5a-b) (abs(log2FC) > 0.2, p-adj <=0.05) (significant genes at abs(log2FC) > 0.5 shown in Table 3). Many of the top shared genes in cluster F fusion cells are histone genes, suggesting ongoing chromatin remodeling, supported by strong enrichment for nucleosome organization by GSEA (Fig. 5d). Compared to cluster F fusion cells, fusion cells found in their parental clusters show enrichment for proteotoxic stress and viral response related pathways (Fig. 5d). These fusion cells also show higher expression of mitotic genes (CCNB1, AURKA, PLK1, NEK2, KIF20A, CKAP2, CDC20, DEPDC1, CKS2, SGO2) and a higher proportion of cells in G2/M compared to cluster F (Extended Data Fig. 10g-h), which may suggest a more proliferative cell state.

Fusion cells vs. control cells	Up	MT2A, ADIRF, NAALADL2, COL13A1, ADAMTS6, KIZ-AS1, HCFC1R1, HMGN5, NDUFC1, IRF2BPL, SPTLC3, H4C3, CUTA, NR2F2-AS1, NDUFB1, RIBC2, NRM, ARHGDIB, ENSG00000255094, FTL
	Down	CADPS2, PDE3A, PRR5L, TBC1D4, PPP1R16A, AASS, BAHCC1, MAFK, HEG1, LRP5L

Table 2. Top differentially expressed genes in fusion cells versus parallel control cells across all clusters (abs(log2FC) > 0.35, p-adj <=0.05).

Fusion cells in cluster F vs. fusion cells in other clusters	Cluster F	ANXA10, NRG1, H1-4, ENSG00000249631, H3C3, H1-2, H2AC11, HTR7, H4C2, H2BC17, H2AC12, MECOM, ADAMTS1, PDE4D, H2BC12, NAV3, NALCN, H3C1, H3C2, H2BC14, DIAPH2, INSYN2B, GLRB, H2BC13, MIR924HG, H2AC21, TMEM117, H2BC21, IMMP2L, LINC01605, CPA6, SMYD3, H2AC14, WWOX, INTS6L, CENPP, TNIK, TARID, H2BC18, STAC, H2AC13, LINC00342, ENSG00000249738
	Other	LINC00862, CCNB1, AURKA, PLK1, NEK2, CREBRF, KIF20A, IFIT1, CKAP2, CDC20, ZNFX1, AGR2

Table 3. Top differentially expressed genes in fusion cells in cluster F vs fusion cells in other clusters (abs(log2FC) > 0.5, p-adj <=0.05).

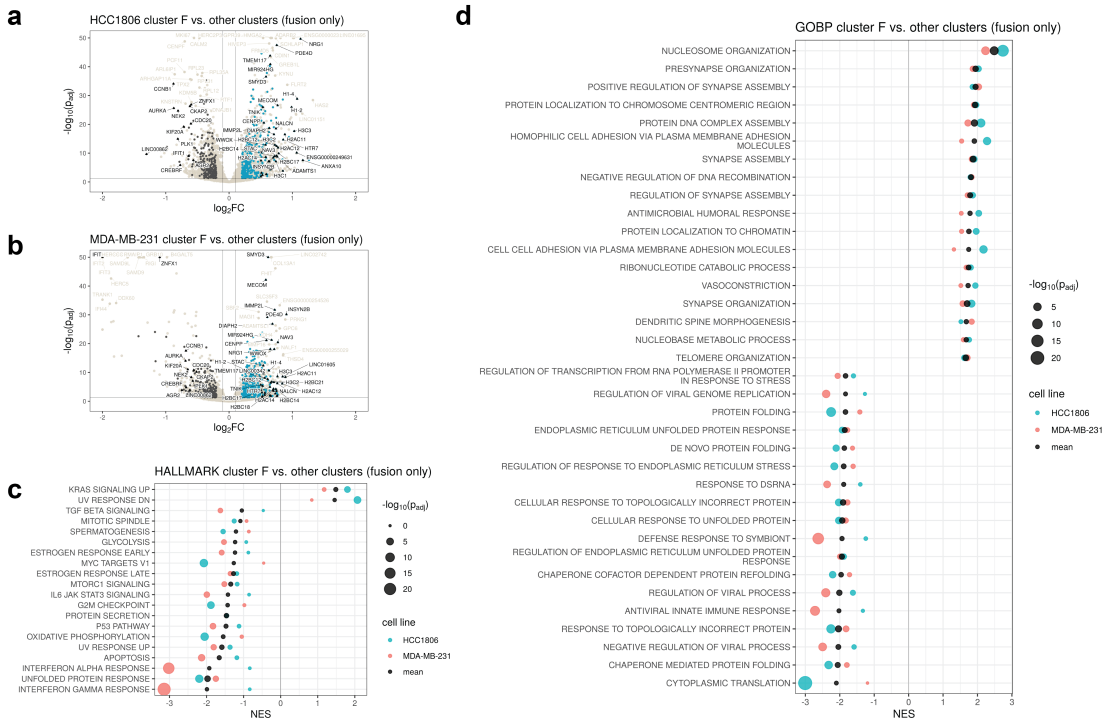

Figure 5. Differential expression and gene set enrichment analysis of fusion cells in cluster F versus fusion cells in other clusters. (a) Volcano plot of differentially expressed genes in cluster F fusion cells (blue, right) vs. other cluster fusion cells (dark grey, left) for the HCC1806 and (b) MDA-MB-231 populations, highlighting the top genes differentially expressed in both cell lines. (c) Hallmark enrichment analysis showing pathways up or down in both cell lines, with positive NES scores matched to cluster F fusion cells and negative NES scores for fusion cells in other clusters. (d) GO biological process enrichment analysis showing pathways up or down in both cell lines.

Discussion

Cell-cell fusion in cancer confounds our current understanding of tumor evolution as an asexual process[14,32], but it can no longer be ignored. Genome doubling through cell-cell fusion may be a mechanism by which cells are able to offset the negative effects of mutational load and slow Muller's Ratchet[33,34]. There are well-studied cell types in the human body known to participate in fusion in normal and pathological contexts, but these events can be more readily classified due to maintenance of multiple nuclei. In this study and numerous others, multinucleation after cell-cell fusion was observed as a transient cell state producing single-nucleated fusion progeny visually similar to the parental cells (Fig. 1c, Extended Data Fig. 1). Without the use of red and green fluorescent markers, fusion cells may be indistinguishable from parental cells. This leads us to hypothesize that the rate of cell-cell fusion in cancer and in normal tissues may be higher than currently measurable. CMDuo now enables high-resolution *in vitro* investigation into the causes and consequence of cell-cell fusion in cancer and other tissues.

Prior studies investigating cancer cell-cell fusion have largely used heterotypic, often genotypically mismatched, and sometimes species mismatched cells[13,35–41]. Studies performed in these model systems have allowed assessment of parental contribution to the phenotypic and karyotypic state of fusion progeny, but many of the reported signatures have focused on inherited features from parental cell types. For example, it has been reported that fusion between immune and cancer cells produces tumor hybrid cells with leukocyte-specific pathways[23,42–44], but the extent to which the physical process of cell-cell fusion may have further amplified or muted the acquired features and the role of subclonal phenotypes within each parental population has not yet been explored. With the resolution achievable using the CMDuo platform, we chose to investigate homotypic cell-cell fusion to decouple transcriptomic changes induced by the process of cell-cell fusion itself from changes driven by response to mismatched genotypes and states inherited from parental cells.

In the cell populations tested here, a small percentage of parent clones (Fig. 2b,g-h) with enrichment for actin and actomyosin-related pathways (Extended Data Fig. 3c-d) appeared more often in fusion progeny than expected by random chance. This may represent a phenotype which enhances survival and proliferation after fusion, or may suggest active actin restructuring toward a more fusogenic cell state[45,46].

This study specifically enriched for and investigated cells which proliferate after cell-cell fusion, but cell-cell fusion does not always lead to a proliferative phenotype. Fusion progeny often do not survive fusion events[30], they may remain multinucleated, or may adopt a giant polyaneuploid cell state. Multinucleated fusion cells and giant polyaneuploid cancer cells can become senescent or undergo eventual ploidy reduction to generate proliferative progeny[47–49]. In this timescale of this experiment, it is reasonable to assume that some of the fusion progeny collected and characterized at the final timepoint may have arisen from transition through a giant cell state. Further work is needed

370 to quantify the occurrence of different cell-cell fusion outcomes in these populations and
371 to link each outcome to parental cell states.
372 In this study, lineage tracing with ClonMapper Duo revealed that fusion progeny
373 inherit and hybridize transcriptomic patterns from their parental clones and activate distinct
374 transcriptomic programs in response to cell-cell fusion. We identified two distinct cell
375 states achieved through homotypic cell-cell fusion often co-occurring within each fusion
376 clone. The most common state achieved by cell-cell fusion was transcriptomically distinct
377 and not correlated to parental cell state. This fusion-driven cell state was characterized by
378 pathways related to chromatin remodeling, potentially suggesting a more epigenetically
379 plastic cell state. Additionally, compared to other cells in the population, fusion progeny in
380 the fusion-driven cell state had a marked decrease in interferon signaling suggesting a
381 more immune privileged cell state which may have therapeutic implications. Fusion cells
382 that adopted a transcriptomic signature more closely matched to their parental state were
383 enriched for both oxidative phosphorylation and glycolysis and showed signatures
384 correlated to an increased rate of mitosis. These fusion cells also showed enrichment for
385 viral response pathways, which we hypothesize is due to presence of cytosolic DNA from
386 micronuclei-mediated chromothripsis[50–52].
387 The relationship between the two identified fusion cell states is currently unknown.
388 Fusion cells may be bistable in both states or fusion cells may pass through the fusion-
389 enriched cluster F state before stabilizing into a phenotype more closely matched to the
390 initial population. Given the mitotic signatures in the fusion cells outside of cluster F, the
391 latter hypothesis appears supported by work from Miroshnychenko et al. who showed an
392 increase in the growth rate of fusion clones over passaging[13]. Longitudinal studies are
393 needed to investigate the stability of these cell states and assess the degree of
394 interconversion.
395 Application of the CMDuo platform in the context of this study has revealed novel
396 insights into cancer cell-cell fusion, but it is not without shortcomings. To prevent gene
397 signatures associated with response to antibiotics from skewing transcriptomic results, we
398 chose to engineer the CMDuo system without antibiotic selection features. While we were
399 able to achieve relatively pure fusion samples for all HCC1806 replicates and 2 of the 4
400 MDA-MB-231 replicates during the experiment (Supp. Table 2, Supp. Table 4), the cells
401 needed to undergo multiple rounds of cell sorting and recovery to reach this level of purity.
402 The inclusion of FACS controls in this study assisted with regressing out changes in gene
403 expression due to cell sorting, but the repeated sorting process itself was a stronger
404 selective pressure than expected in the MDA-MB-231 model. In a prior study, two highly
405 distinct transcriptomic clusters were described within the MDA-MB-231 cell line,
406 characterized by high or low expression of the surface marker, ESAM[53]. Here, the ESAM-
407 low subpopulation was depleted in both the fusion and control samples (Supp. Fig. 6, Fig.
408 1h, Extended Data Fig. 7d) and the same 2 clones were selected for in control samples
409 across all four replicates (Fig. 1k). There may be significant fusion occurring in the ESAM-
410 low MDA-MB-231 subpopulation, but robust inferences cannot be made given the degree
411 of FACS-induced selection seen in the control samples and the low number of fusion and

control cells in this subpopulation. Antibiotic selection with the proper controls, may have been a more gentle and efficient method to enrich for double-barcoded fusion cells. Future iterations of the CMDuo system should include antibiotic selection as an optional feature.

ClonMapper Duo is a novel platform which was used to study homotypic cell-cell fusion in triple-negative breast cancer cells but can be used to study homotypic or heterotypic cell-cell fusion between any mammalian cell types *in vitro*. Additionally, this study highlighted the use of CMDuo for tracking cell-cell fusion, but it can be readily applied to lineage tracing studies with other goals. The fluorescence-associated barcodes allow confident assignment of population identity in scRNA-seq; different cell populations can be cocultured and accurately demultiplexed in scRNA-seq which will be useful for investigation of other cell-cell interactions. Further, the localization of the barcodes to the nucleus with H2B-GFP or H2B-mCherry allows precise counting of population size with live-cell imaging, which is useful for determining population growth rates and fitting data to mathematical models of cancer[54,55].

Without biomarkers, the clinical relevance of cell-cell fusion in cancer remains unknown[35]. In this study, we applied a novel barcoding platform and unveiled signatures of clonal population before and after cell-cell fusion. Future studies into the mechanistic role of the identified gene signatures are required, and these studies may lead to the discovery of biomarkers for assessing fusogenic potential and the ability to quantify the prevalence of fusion in patient samples. If cell-cell fusion is identified as a key component of tumor evolution in any cancer type, then targeting pathways involved in fusogenic potential or response to cell-cell fusion may represent novel treatment strategies.

Methods

Construction of the CMDuo plasmids

The H2B-GFP gene block amplifying was cloned from the FU-H2B-GFP-IRES-Puro plasmid (AddGene #69550) by performing a 1X Q5 PCR reaction. Briefly 25 µL Q5 High-Fidelity 2X Master Mix, 2.5 µL of 10uM forward primer, 2.5 µL of 10uM reverse primer, 0.5 ng of template DNA, and nuclease-free water to 50 uL were mixed then PCR amplified with the following settings: (1) 98C for 2 min, (2) 98 °C for 5 sec, (3) 72 °C for 30 sec, (4) 72 °C for 1 min, (5) Repeating steps 2-4 for 20 cycles, then (6) 72 °C for 2 min, and (7) 4 °C hold. The PCR product was isolated and purified using gel electrophoresis. H2B-mCherry was ordered as a gene block designed to match the resulting H2B-GFP gene block but with mCherry in place of GFP. To linearize and remove the BFP sequence from the ClonMapper vector backbone (CROPseq-BFP-WPRE-TS-hU6-BsmBI, AddGene # 137993), 3 ug of plasmid was digested in a 3X restriction enzyme reaction containing 15 uL Digestion rCutSmart Buffer (NEB, B6004), 3 µL BsiWI-HF (NEB, R3553), 3 µL MluI-HF (NEB, R3198), and nuclease-free water to 150 µL at 37 °C for 1 hour. The linearized

454 plasmid band was purified using gel electrophoresis. Gene blocks were then separately
455 ligated into the linearized ClonMapper backbone at a molar ratio of 2:1 in a 1X HiFi
456 Assembly reaction by mixing 0.0638 pmol PCR product, 0.0319 pmol linearized plasmid,
457 10 µL HiFi DNA Assembly Master Mix (NEB, E2621), and nuclease-free water to 20 µL at
458 50 °C for 1 hour. 2 µL of the finished reaction mixture was heat shocked into one tube of
459 competent e. coli (NEB, C3040) at 42 °C for 30 seconds, incubated in recovery media for
460 1 hour, then spread over ampicillin agar plates for overnight incubation. Colonies were
461 expanded and the insertion of the reporter sequence was confirmed with Sanger
462 sequencing.
463
464 **Generation of indexed barcode plasmid libraries**
465 Random indexed barcode library gene bocks for GFP and mCherry were made by mixing
466 equimolar ratios of each of their respective 4 random barcode templates (Supp. Table 1)
467 and then performing an extension reaction as previously published[56–58]. The random
468 indexed barcode gene blocks were inserted into their respective plasmid backbones using
469 a BsmBI-v2 golden gate reaction following the published procedure. Targeted barcode
470 sequencing was performed on the generated plasmid barcode libraries sequencing to
471 assess population diversity following the published procedure.
472
473 **Cell culture**
474 HCC1806 cells were cultured in RPMI with 2.05 mM L-glutamine (HyClone, SH30027.01),
475 supplemented with 10% FBS (Sigma, F0926-500mL). MDA-MB-231 cells were cultured in
476 high glucose DMEM (Sigma D5796) supplemented with 10% FBS (Sigma, F0926-500mL)
477 and 1X penicillin-streptomycin (ThermoFisher, 15140122). MDA-MB-231 cells were
478 passaged with 0.05% Trypsin-EDTA (ThermoFisher, 25300062) for 5 minutes and
479 HCC1806 cells were passaged with 0.25% Trypsin-EDTA (ThermoFisher, 25200056) for
480 12 minutes when cultures reached 70% confluence. MDA-MB-231 cells were obtained
481 from ATCC and barcoded within 5 passages. HCC1806 cells in the lab were STR verified
482 before use.
483
484 **Generation of barcoded cell libraries**
485 Lentiviral production and low MOI transduction (MOI=0.01) was performed according to
486 previously published methods for the ClonMapper system[56–58]. For each cell line, CMDuo-
487 green and CMDuo-red populations were generated in parallel from the same starting
488 population by sorting replicates of 500, 800, or 1000 cells for each cell line. Sorted cell
489 libraries were expanded and targeted barcode sequencing was performed to assess
490 number of unique clones and skew within each sorted population. Low cell number
491 libraries showing the highest evenness were chosen for the experiments. The HCC1806
492 populations selected for the scRNA-seq experiment came from expanded cell libraries of
493 500 CMDuo-green and 500 CMDuo-red sorted cells, and the MDA-MB-231 populations
494 came from 800 CMDuo-green and 800 CMDuo-red sorted cells.
495

496 **Cell sorting and sample collection**
497 For each cell line, five replicate plates were seeded, containing 1e6 cells each. One plate
498 was fixed for scRNA-seq 2 days after seeding ("initial"), the other 4 plates were expanded
499 in culture as independent replicates for one week before the first FACS enrichment for
500 fusion cells. At the time of sorting, 0.5e6 cells were fixed for Parse scRNA-seq fixation and
501 0.5e6 cells were pelleted for DNA and targeted barcode sequencing. These cells comprise
502 the "pre-sort" samples. The remaining cells were FACS sorted for fusion cells (both GFP^+
503 and $mCherry^+$ cells) and control cells (single GFP^+ or single $mCherry^+$). As FACS is
504 imperfect at double discrimination and there are no additional surface markers currently
505 known which can be added in to discriminate fusion events from red-green doublets,
506 multiple rounds of FACS were needed to obtain high purity fusion samples. Given that
507 multiple rounds of FACS may act as a selective pressure and cause transcriptomic shifts,
508 control cells (GFP^+ or $mCherry^+$) were sorted in parallel with fusion cells ($GFP^+mCherry^+$)
509 to correct for any selective pressures and transcriptomic shifts which may be driven by
510 cell sorting stress rather than fusion in downstream analysis. Pre-sort purity and cell yield
511 for each round of FACS is summarized in Supp. Table 2. Sorted cells were allowed to
512 recover from cell sorting and expanded before undergoing repeat rounds of cell sorting
513 until fusion populations were > 50% pure, then "fusion-enriched" and "control" cells were
514 prepped for scRNA-seq and DNA extraction. From fixation of the "initial" timepoint to the
515 time at which "fusion-enriched" and parallel "control" cells were fixed for scRNA-seq,
516 HCC1806 cells underwent 3 enrichment sorts over 35 days and MDA-MB-231 underwent
517 4 enrichment sorts over 50 days. The final purity of each sample fixed for scRNA-seq
518 assessed by flow cytometry is summarized in Supp. Tables 2, 4.
519
520 **Bulk targeted barcode library preparation**
521 DNA was extracted from frozen cell pellets using the Invitrogen PureLink Genomic DNA
522 kit (ThermoFisher, K182001) according to manufacturer's instructions. Amplification steps
523 were performed following published ClonMapper barcode amplification procedures[56–58].
524 Briefly, DNA was quantified with NanoDrop, then 1 µg of DNA was loaded into the first
525 PCR reaction. Amplified DNA was cleaned with SPRI beads, then 4 ng of cleaned stage
526 1 product was loaded into the next PCR reaction to add unique Illumina i7/i5 indices to
527 each sample. Stage 2 samples were cleaned with SPRI beads, then quantified Qubit using
528 the 1X dsDNA high-sensitivity assay (ThermoFisher, Q33231). Illumina indexed samples
529 were run on a NovaSeq 6000 (S1 flow cell, single read, 100 cycles) by the UT Genome
530 Analysis and Sequencing Facility, targeting 100e6 reads for all samples and biasing more
531 reads toward the earlier timepoint samples. ~14e6 reads were obtained for each initial
532 population, ~6e6 for each pre-sort sample, and ~3e6 for each control and fusion sample
533 in each cell line.
534
535 **Bulk targeted barcode sequencing analysis, barcode naming convention, and**
536 **creating a reference list for Parse CRISPR Detect alignment**

537 The results of targeted barcode sequencing performed on bulk DNA extracted from each
538 sample were combined into early and late timepoint groups for each cell line, with the early
539 group containing the initial and pre-sort samples, and the fusion-enriched and control
540 samples in the late group. The set of barcodes in each group was determined with
541 pycashier[59] (version 2024.1001) (https://github.com/brocklab/pycashier/), setting the
542 Levenshtein distance to 2 and the clustering ratio to 1. Post-processing on the pycashier
543 outputs was performed to remove noisy sequences detected in both cell lines and ensure
544 a minimum Levenshtein distance of 2 between all pairs of barcodes within each cell line.
545 If any similar barcodes were detected, they were collapsed into the more abundant
546 sequence. Barcode color assignment was performed by matching the first 5 nucleotides
547 of each detected sequence to the designed index sequences, (withGFP being one of the
548 following sequences: GACAA, GCTGT, ATCGC, ACGCA, and mCherry being: CATCC,
549 CGAGA, TAGTG, TCAAC) and allowing a Levenshtein distance of 1 to account for
550 sequencing errors. Unique nucleotide barcodes were assigned names according to their
551 fluorescence reporter and their relative abundance in the ranking abundance, i.e. g001
552 and r001 represent the most abundant CMDuo-green and CMDuo-red clones in the early
553 timepoint samples via targeted sequencing. Named barcodes and sequences were
554 formatted into a reference list for input to the Parse CRISPR Detect alignment pipeline (v
555 1.3.1), setting the barcode name as the 'Guide_Name', the detected sequence as both
556 the 'Guide_Sequence' and 'Target_Gene', and including flanking regions around the
557 barcode as the 'Prefix' (ATCTTGTGGAAAGGACGAAACACCG) and 'Suffix'
558 (GTTTTAGAGCTAGAAATAGCAAGTT).
559
560 **Parse scRNA-seq library prep**
561 Library prep was performed according to manufacturer guidelines outlined in CRISPR
562 Detect for Evercode WT v2 (User Manual Version 1.1, UM0025). As the WT kit can
563 process 100,000 cells, samples from both cell lines were prepared with the same kit. For
564 each cell line, the 'initial' sample was split over 4 wells targeting 13,000 cells total, each
565 of the four 'pre-sort' samples were split over 3 wells targeting 6,250 cells per sample, and
566 each of the sorted control and fusion-enriched samples received its own well, targeting
567 1500 cells per sample. 8 identical sublibraries were prepared, but one sublibrary was lost
568 during preparation, adjusting the expected cell number to 87,500 total and 11,375, 5,469,
569 and 1312 per sample per cell line, respectively. ClonMapper Duo gRNA barcodes were
570 amplified from sublibrary cDNA according to manufacture instructions for CRISPR
571 Sequencing Library Preparation in CRISPR Detect for Evercode WT v2 (User Manual
572 Version 1.1, UM0025). Whole transcriptome and CRISPR detect (CMDuo barcode)
573 libraries were amplified separately from cDNA and sequenced independently across 8
574 sublibraries each. Whole transcriptome and CRISPR Detect sublibraries were sequenced
575 at the UT Austin Genomic Sequencing and Analysis on a NovaSeq 6000 (S2 flow cell,
576 paired-end, 200 cycles). 500e6 reads were targeted for each WT sublibrary and 50e6
577 reads for each CRIPSR Detect sublibrary, producing 41,973 mean reads per cell across

transcriptomic libraries and 3,238 means reads per cell in the combined CRISPR Detect sublibraries.

scRNA-seq genome annotation and sequence alignment

All sequence alignment and sample annotation steps were performed with Parse Bioscience pipeline (v 1.3.1) in a Linux environment running Python version 3.10. Genome annotation and reference files containing eGFP and mCherry sequences were concatenated to the most recent human genome build (GRCh38, release 111). A Parse reference genome was built from the concatenated genome and annotation files. Each Parse whole transcriptome sublibrary was aligned to the built genome, then aligned sublibraries were combined for analysis.

scRNA-seq barcode alignment

Alignment and sample annotation steps were performed with Parse Bioscience pipeline (v 1.3.1) in a Linux environment running Python version 3.10. Formatted pycashier outputs from independently run targeted barcode sequencing were used as inputs to the Parse CRISPR Detect pipeline, setting a minimum read threshold of 3 and a minimum transcript threshold of 3.

scRNA-seq quality control filtering and normalization

Quality control steps were performed on the unfiltered outputs of the Parse pipeline using Seurat (v5.1.0). Cells in retained for analysis had total RNA counts between 1,500-100,000 genes, total feature counts between 800-50,000 counts, and mitochondrial percent between 2-18%. Doublet detection was performed using scDblFinder[28] (v1.18.0) and cells classified as doublets were removed. Cells were split by cell line and data from each cell line was independently normalized using a log transformation. Variable features were identified using a variance stabilizing transformation. Cell cycle score was calculated for each cell and scaled data from cell cycle regression was used to identify transcriptomic clusters and generate UMAP projections, but downstream analysis was carried out on non-cell cycle regressed data. Louvain clustering was performed using the first 20 principal components to identify transcriptomic neighbors and at a clustering resolution of 0.18 in both cell lines. 8 clusters were identified for HCC1806 and 7 clusters for MDA-MB-231 before barcode-informed clustering correction.

scRNA-seq assignment of barcodes to cells

Barcodes were assigned to each cell using a probabilistic model accounting for sequencing noise in the data using the raw outputs from the Parse CRISPR detect pipeline. Assigned barcodes were first filtered to match the expected cell line, to have at least 3 reads per barcode in a cell, and to ensure that detected barcodes is present in at least 3 cells at this threshold within a sample. After this minimal filtering, the expected noise generated from each barcode was calculated based on overall barcode abundance in the population and the number of reads per cell. A negative binomial model was used

620 to determine the probability that a barcode within a cell is noise. Barcodes with a p-
621 adjusted value <= 0.05 were assigned to cells. Cells with multiple instances of green
622 barcodes or multiple instances of red barcodes were indicated with underscores, e.g.,
623 gXXX_gYYY. Once barcodes were assigned, clone color was assigned as green, red, or
624 fusion by presence of 'g' and/or 'r' in the assigned barcode name.
625
626 **scRNA-seq barcode-informed clustering correction**
627 Given the that most clonal lineages maintain transcriptomic similarity[57], we corrected for
628 potential over clustering from the Louvain algorithm by quantifying clonal abundance
629 within each cluster and combining Louvain clusters with similar clonal profiles. Distance
630 between vectorized clonal abundance in each cluster was quantified using the Canberra
631 method. Hierarchical clustering was performed to visualize similarity of barcode
632 distributions within clusters and create a dendrogram tree. The tree was cut at a minimum
633 distance of 200 in both cell lines. This resulted in no changes to the MDA-MB-231
634 population, retaining its 7 Louvain clusters, and one change to the HCC1806 population,
635 collapsing 2 of the 8 clustering into one, and producing 7 barcode-informed clusters for
636 downstream analysis in each cell line (Supp. Fig. 3).
637
638 **scRNA-seq cluster naming**
639 Barcode-informed clusters were re-named from numeric values to letters A-G in each data
640 set, setting cluster F in each to be the cluster with the highest quantity of fusion cells.
641 Across data sets, cluster F is the only cluster which has any informed similarity. Cluster
642 names A-E, and G were assigned to the remaining numerical clusters within each cell line
643 without preference.
644
645 **Cell classification**
646 Unique combination of at least one red and green barcode appearing in at least 5 cells
647 within a fusion-enriched replicate were classified as “fusion” clones. The parental
648 barcodes associated with each classified fusion clone were then used to classify cells as
649 “parental” clones.
650
651 **scRNA-seq differential expression analysis**
652 Cluster marker genes were identified using all cells in each data set using a Wilcoxon rank
653 sum test, assigning the highest significant log2FC values as maker genes for each cluster.
654 A mixed model approach with NEBULA[60] was used for all other differential expression
655 analyses to correct for underlying clonal composition, setting cell barcode as the random
656 effect (i.e. the subject id). Unless otherwise stated, a minimum of 5 cells per classified
657 clone was required to be included in each analysis. Output log fold change was converted
658 to log2FC and p-values were adjusted using a Benjamini-Hochberg correction. The
659 adjusted p-values, p-adjust, were used to determine significance. Differentially expressed
660 genes were defined at an absolute log2FC threshold of 0.2 and a p-adjust value <= 0.05.
661

662 **Gene set enrichment analysis**
663 Genes from NEBULA mixed model analysis were pre-ranked using log2FC and GSEA
664 was performed using fgsea[61] (v1.30.0) on the v2023.2 h.hallmarks and
665 c5.go.biological_process gene sets from the MSigDB[62,63]. Gene sets displayed in plots
666 show gene sets that were up or down in both cell lines and had a false discovery rate of
667 less than or equal to 0.25 in at least one cell line.
668
669 **Parent-progeny marker gene analysis**
670 For each parent-parent-fusion set with at least 20 cells in each fraction, parental marker
671 genes were calculated using a Wilcoxon rank sum test between cells of the green parent
672 clone and cells from the red parent clone. The top 15 differentially expressed genes for
673 each were considered parental marker genes and used for visualizations in Fig. 3. Clonal
674 similarity was calculated by pseudobulking gene expression for each clone, then
675 calculating pairwise Pearson correlation coefficient between clones in each parent-parent-
676 fusion set.
677
678
679 **Acknowledgements**
680
681 We are grateful for funding from the NIH which supported this project: NIH F31CA268833
682 (to A.L.G.), R01CA226258, R01CA255536, and U01CA253540 (to A.B.) We thank the
683 core facilities at UT Austin for access to their services and knowledge. Sequencing was
684 performed by the Genomic Sequencing and Analysis Facility at UT Austin, Center for
685 Biomedical Research Support (RRID: SCR_021713). Flow cytometry and FACS was
686 performed at the Center for Biomedical Research Support Microscopy and Imaging
687 Facility at UT Austin (RRID: SCR_021756).
688
689 **Data Availability**
690
691 Raw and processed single cell RNA-sequencing can be downloaded from GEO with
692 accession number GSE286213. Targeted barcode sequencing to generate a list of
693 barcodes in the samples can be downloaded from Gene Expression Omnibus (GEO) with
694 accession number GSE286212.
695
696
697 **Code Availability**
698
699 Code used to process data and generate figures in this study can be found at
700 www.github.com/brocklab/CMDuo-analysis
701
702
703

704 **References**

705

706 1. Marusyk, A., Janiszewska, M. & Polyak, K. Intratumor Heterogeneity: The Rosetta

707 Stone of Therapy Resistance. *Cancer Cell* **37**, 471–484 (2020).

708 2. Morris, L. G. T. *et al.* Pan-cancer analysis of intratumor heterogeneity as a

709 prognostic determinant of survival. *Oncotarget* **7**, 10051–10063 (2016).

710 3. Tasdogan, A. *et al.* Metabolic heterogeneity confers differences in melanoma

711 metastatic potential. *Nature* **577**, 115–120 (2020).

712 4. Jhan, J.-R. & Andrechek, E. R. Triple-negative breast cancer and the potential for

713 targeted therapy. *Pharmacogenomics* **18**, 1595–1609 (2017).

714 5. Stewart, R. L. *et al.* A Multigene Assay Determines Risk of Recurrence in Patients

715 with Triple-Negative Breast Cancer. *Cancer Res* **79**, 3466–3478 (2019).

716 6. Li, X. *et al.* Triple-negative breast cancer has worse overall survival and cause-

717 specific survival than non-triple-negative breast cancer. *Breast Cancer Res*

718 *Treat* **161**, 279–287 (2017).

719 7. Taylor, A. M. *et al.* Genomic and Functional Approaches to Understanding Cancer

720 Aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).

721 8. Kaufmann, T. L. *et al. MEDICC2: Whole-Genome Doubling Aware Copy-Number*

722 *Phylogenies for Cancer Evolution*.

723 http://biorxiv.org/lookup/doi/10.1101/2021.02.28.433227 (2021)

724 doi:10.1101/2021.02.28.433227.

725 9. Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human
726 cancer. *Nat Biotechnol* **30**, 413–421 (2012).
727 10. Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nat*
728 *Genet* **45**, 1134–1140 (2013).
729 11. Watkins, T. B. K. *et al.* Pervasive chromosomal instability and karyotype order in
730 tumour evolution. *Nature* **587**, 126–132 (2020).
731 12. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer.
732 *N Engl J Med* **376**, 2109–2121 (2017).
733 13. Miroshnychenko, D. *et al.* Spontaneous cell fusions as a mechanism of
734 parasexual recombination in tumour cell populations. *Nat Ecol Evol* **5**, 379–391
735 (2021).
736 14. Laplane, L. & Maley, C. C. The evolutionary theory of cancer: challenges and
737 potential solutions. *Nat Rev Cancer* (2024) doi:10.1038/s41568-024-00734-2.
738 15. Lu, X. & Kang, Y. Cell Fusion as a Hidden Force in Tumor Progression. *Cancer*
739 *Research* **69**, 8536–8539 (2009).
740 16. Searles, S. C., Santosa, E. K. & Bui, J. D. Cell-cell fusion as a mechanism of DNA
741 exchange in cancer. *Oncotarget* **9**, 6156–6173 (2018).
742 17. Zhang, L.-N. *et al.* Fusion with mesenchymal stem cells differentially affects
743 tumorigenic and metastatic abilities of lung cancer cells: ZHANG ET AL. *J Cell*
744 *Physiol* **234**, 3570–3582 (2019).

745 18. Gauck, D., Keil, S., Niggemann, B., Zänker, K. S. & Dittmar, T. Hybrid clone cells
746 derived from human breast epithelial cells and human breast cancer cells
747 exhibit properties of cancer stem/initiating cells. *BMC Cancer* **17**, 515 (2017).
748 19. Shabo, I. *et al.* Roles of cell fusion, hybridization and polyploid cell formation in
749 cancer metastasis. *WJCO* **11**, 121–135 (2020).
750 20. Wakeling, W. Efficient Spontaneous Fusion Between Some Co-cultured Cells,
751 Especially Murine Melanoma Cells. *Cell Biology International* **18**, 207–210
752 (1994).
753 21. Gast, C. E. *et al.* Cell fusion potentiates tumor heterogeneity and reveals
754 circulating hybrid cells that correlate with stage and survival. *Sci. Adv.* **4**,
755 eaat7828 (2018).
756 22. Miller, F. R., McInerney, D., Rogers, C. & Miller, B. E. Spontaneous fusion
757 between metastatic mammary tumor subpopulations. *J. Cell. Biochem.* **36**,
758 129–136 (1988).
759 23. Shabo, I. *et al.* Macrophage traits in cancer cells are induced by macrophage-
760 cancer cell fusion and cannot be explained by cellular interaction. *BMC Cancer*
761 **15**, 922 (2015).
762 24. Shabo, I., Stål, O., Olsson, H., Doré, S. & Svanvik, J. Breast cancer expression of
763 CD163, a macrophage scavenger receptor, is related to early distant recurrence
764 and reduced patient survival: Expression of Macrophage Receptors in Breast
765 Cancer. *Int. J. Cancer* **123**, 780–786 (2008).

766 25. Suetsugu, A. *et al.* Color-coded Live Imaging of Heterokaryon Formation and
767 Nuclear Fusion of Hybridizing Cancer Cells. *Anticancer Res* **36**, 3827–3831
768 (2016).
769 26. Dittmar, T. & Zänker, K. Tissue Regeneration in the Chronically Inflamed Tumor
770 Environment: Implications for Cell Fusion Driven Tumor Progression and
771 Therapy Resistant Tumor Hybrid Cells. *IJMS* **16**, 30362–30381 (2015).
772 27. Suetsugu, A. *et al.* Color-coded Live Imaging of Heterokaryon Formation and
773 Nuclear Fusion of Hybridizing Cancer Cells. *Anticancer Res* **36**, 3827–3831
774 (2016).
775 28. Germain, P.-L., Lun, A., Garcia Meixide, C., Macnair, W. & Robinson, M. D.
776 Doublet identification in single-cell sequencing data using scDblFinder.
777 *F1000Res* **10**, 979 (2022).
778 29. Zhang, Z. *et al.* Synthetic DNA barcodes identify singlets in scRNA-seq datasets
779 and evaluate doublet algorithms. *Cell Genomics* 100592 (2024)
780 doi:10.1016/j.xgen.2024.100592.
781 30. Sato, S., Rancourt, A. & Satoh, M. S. Cell fate simulation reveals cancer cell
782 features in the tumor microenvironment. *Journal of Biological Chemistry* **300**,
783 107697 (2024).
784 31. Buschmann, T. DNABarcodes: an R package for the systematic construction of
785 DNA sample tags. *Bioinformatics* **33**, 920–922 (2017).

786 32. Shultes, P. V., Weaver, D. T., Tadele, D. S., Barker-Clarke, R. J. & Scott, J. G. Cell-
787 cell fusion in cancer: The next cancer hallmark? *The International Journal of*
788 *Biochemistry & Cell Biology* **175**, 106649 (2024).
789 33. Muller, H. J. The relation of recombination to mutational advance. *Mutation*
790 *Research/Fundamental and Molecular Mechanisms of Mutagenesis* **1**, 2–9
791 (1964).
792 34. TRACERx Consortium *et al.* Interplay between whole-genome doubling and the
793 accumulation of deleterious alterations in cancer evolution. *Nat Genet* **52**, 283–
794 293 (2020).
795 35. Hass, R., von der Ohe, J. & Dittmar, T. Hybrid Formation and Fusion of Cancer
796 Cells In Vitro and In Vivo. *Cancers (Basel)* **13**, 4496 (2021).
797 36. Melzer, C., von der Ohe, J. & Hass, R. Enhanced metastatic capacity of breast
798 cancer cells after interaction and hybrid formation with mesenchymal
799 stroma/stem cells (MSC). *Cell Commun Signal* **16**, 2 (2018).
800 37. Miller, F. R., McInerney, D., Rogers, C. & Miller, B. E. Spontaneous fusion
801 between metastatic mammary tumor subpopulations. *J. Cell. Biochem.* **36**,
802 129–136 (1988).
803 38. Shabo, I. *et al.* Roles of cell fusion, hybridization and polyploid cell formation in
804 cancer metastasis. *WJCO* **11**, 121–135 (2020).

805 39. Shabo, I. *et al.* Macrophage traits in cancer cells are induced by macrophage-
806 cancer cell fusion and cannot be explained by cellular interaction. *BMC Cancer*
807 **15**, 922 (2015).
808 40. Wakeling, W. Efficient Spontaneous Fusion Between Some Co-cultured Cells,
809 Especially Murine Melanoma Cells. *Cell Biology International* **18**, 207–210
810 (1994).
811 41. Zhang, L.-N. *et al.* Fusion with mesenchymal stem cells differentially affects
812 tumorigenic and metastatic abilities of lung cancer cells: ZHANG ET AL. *J Cell*
813 *Physiol* **234**, 3570–3582 (2019).
814 42. Gast, C. E. *et al.* Cell fusion potentiates tumor heterogeneity and reveals
815 circulating hybrid cells that correlate with stage and survival. *Sci. Adv.* **4**,
816 eaat7828 (2018).
817 43. Parappilly, M. S. *et al.* Circulating Neoplastic-Immune Hybrid Cells Predict
818 Metastatic Progression in Uveal Melanoma. *Cancers* **14**, 4617 (2022).
819 44. Powell, A. E. *et al.* Fusion between Intestinal Epithelial Cells and Macrophages
820 in a Cancer Context Results in Nuclear Reprogramming. *Cancer Res* **71**, 1497–
821 1505 (2011).
822 45. Dittmar, T. & Hass, R. Intrinsic signalling factors associated with cancer cell-cell
823 fusion. *Cell Commun Signal* **21**, 68 (2023).
824 46. Rodríguez-Pérez, F. *et al.* Ubiquitin-dependent remodeling of the actin
825 cytoskeleton drives cell fusion. *Developmental Cell* **56**, 588-601.e9 (2021).

826 47. Pienta, K. J., Hammarlund, E. U., Brown, J. S., Amend, S. R. & Axelrod, R. M.
827 Cancer recurrence and lethality are enabled by enhanced survival and
828 reversible cell cycle arrest of polyaneuploid cells. *Proc Natl Acad Sci USA* **118**,
829 e2020838118 (2021).
830 48. Pienta, K. J., Hammarlund, E. U., Axelrod, R., Brown, J. S. & Amend, S. R. Poly-
831 aneuploid cancer cells promote evolvability, generating lethal cancer. *Evol Appl*
832 **13**, 1626–1634 (2020).
833 49. Chen, J. *et al.* Polyploid Giant Cancer Cells (PGCCs): The Evil Roots of Cancer.
834 *CCDT* **19**, 360–367 (2019).
835 50. Bakhoum, S. F. *et al.* Chromosomal instability drives metastasis through a
836 cytosolic DNA response. *Nature* **553**, 467–472 (2018).
837 51. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour
838 evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
839 52. Stephens, P. J. *et al.* Massive Genomic Rearrangement Acquired in a Single
840 Catastrophic Event during Cancer Development. *Cell* **144**, 27–40 (2011).
841 53. Gardner, A. L., Jost, T. A., Morgan, D. & Brock, A. Computational identification of
842 surface markers for isolating distinct subpopulations from heterogeneous
843 cancer cell populations. *npj Syst Biol Appl* **10**, 120 (2024).
844 54. Liu, L. *et al.* Minimization of thermodynamic costs in cancer cell invasion. *Proc.*
845 *Natl. Acad. Sci. U.S.A.* **110**, 1686–1691 (2013).

846 55. Johnson, K. E. *et al.* Cancer cell population growth kinetics at low densities
847 deviate from the exponential growth model and suggest an Allee effect. *PLOS*
848 *Biology* **17**, e3000399 (2019).
849 56. Al'Khafaji, A. M., Deatherage, D. & Brock, A. Control of Lineage-Specific Gene
850 Expression by Functionalized gRNA Barcodes. *ACS Synth Biol* **7**, 2468–2474
851 (2018).
852 57. Gutierrez, C. *et al.* Multifunctional barcoding with ClonMapper enables high-
853 resolution study of clonal dynamics during tumor evolution and treatment. *Nat*
854 *Cancer* **2**, 758–772 (2021).
855 58. Gardner, A., Morgan, D., Al'Khafaji, A. & Brock, A. Functionalized lineage tracing
856 for the study and manipulation of heterogeneous cell populations. *Methods in*
857 *Molecular Biology* (2021).
858 59. Morgan, D. & Brock, A. Pycashier: cash in on DNA barcode tags. *JOSS* **9**, 6891
859 (2024).
860 60. He, L. *et al.* NEBULA is a fast negative binomial mixed model for differential or
861 co-expression analysis of large-scale multi-subject single-cell data. *Commun*
862 *Biol* **4**, 629 (2021).
863 61. Korotkevich, G. *et al.* Fast gene set enrichment analysis. Preprint at
864 https://doi.org/10.1101/060012 (2016).
865 62. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set
866 Collection. *Cell Systems* **1**, 417–425 (2015).

867 63. Liberzon, A. *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*

868 **27**, 1739–1740 (2011).

869