OREGON HEALTH & SCIENCE UNIVERSITY
SCHOOL OF MEDICINE – GRADUATE STUDIES

From Method to Biological Discovery: Decoding Cancer Heterogeneity

By

Konstantin W. Queitsch

A DISSERTATION

Presented to the Program in Biomedical Sciences
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

September 2025

Table of Contents

I: List of Figures

II: List of Tables

III: List of Abbreviations

Abbreviation	Term
ATAC	Assay for Transpose-Accessible Chromatin
BCR	B-cell Receptor
CHC	Circulating Hybrid Cell
CNV	Copy Number Variant
CRC	Colorectal Cancer
FL	Follicular Lymphoma
GATK	Genome Analysis Toolkit
MAD	Median Absolute Deviation
THC	Tumor Hybrid Cell
TSS	Transcription Start Site
TSSe	Transcription Start Site enrichment
sciMET	high-throughput single-cell DNA methylation assay
scRNA-seq	single-cell RNA Sequencing
scWGS	single-cell Whole Genome Sequencing
SNV	Single Nucleotide Variant
VCF	Variant Call File
V(D)J	Variable, diversity, and joining gene segments

IV: Acknowledgements

I would like to begin by acknowledging my PhD mentor Andrew Adey. You offered me a fantastic mix of independence and support. Thank you for your guidance, patience in forgiving the occasional slip-ups at the bench, and for giving me the opportunity to acquire an entirely new computational skillset from scratch. I only deleted a reference directory once! Your lab was truly the perfect environment for me to pursue my scientific passions. I want to thank the whole Adey lab. On day one, Ruth and Brendan, you took me under your wing. Thank you for teaching me the ropes of single-cell assays and hands-on technology development. I apologize to Ruth for enjoying Brendan's tangents as much as I did. Andy and Sonia – thank you both for always being a friendly face willing to engage with my questions. Kevin – my senior graduate student in the lab – you've been a good friend, and I appreciate you taking the time to troubleshoot my code when I hit walls. Lauren, thank you for being brilliant, funny, and sharing in my indulgence for fantastic pastries. Ben and Sam, you're more recent arrivals; but, having you join brightened up the lab. I wish you both the best as you complete your own PhD journeys.

My last two years at OHSU were defined by two big collaborations. Thank you, Ashley Anderson for being the best collaborator I could have asked for. I am inspired by how you literally do it all and I am proud of the body of work we have produced together. Thank you, Melissa Wong, not just for being a fantastic co-mentor on this project, but also for joining my thesis committee. I greatly appreciate it. To my whole DAC committee – Laura, Gürkan, Hisham – I appreciate the guidance, conversation, and understanding. Although the last year has been slow going, I want to thank Matthew Stern and Tanya Shree for being fantastic collaborators. I think our project is close to fruition and I will do everything I can to help us get over the finish line. Have faith. I would further like to thank Amanda McCullough. You stepped in as an additional academic mentor, personal sounding board, and journal club course director. Taking a step back, much of the work that I have done in recent years would be impossible without the donation of biopsy specimens -

blood, pancreas, brain tissue, and more. I need to acknowledge the selfless individuals who make this type of research feasible. And thank you to the mice too.

Shon Green, to this day I doubt I would have committed to a PhD journey if it was not for your mentorship and warm approach to science. You taught me the true meaning of interdisciplinary collaboration, the joy that comes from teamwork, surmounting shared obstacles, and celebrating shared victories. John Stamatoyannopoulos, you helped me move scientific concepts out of the classroom and conceive of contributing to technologies that reshape how we investigate the world that surrounds us, thank you. Stanley Fields, it is not lost on me that as I finally move to submit my dissertation, your long career is being celebrated. Thank you for teaching me to ask the correct questions from my first internship and providing kind, clear-eyed counsel ever since.

I have received nothing but support from family and friends. To both my parents, thank you for your patience as I have worked to find my way here today. You have in so many ways been tremendous role models, sources of great joy and comfort. I am eternally grateful for having the two of you in my corner. I look forward to paying that forward with my younger brothers. Anton, Leo - y'all are good too. To Breland and Ryan, thank you for showing up when you did. It made all the difference in the world.

V: Abstract

Single-cell multi-omic technologies are pushing the envelope on the interrogatable questions for the life science community. They are crucial for deciphering molecular biology in heterogeneous tissues, where bulk assays lack the ability to resolve discrete differences between individual cell types and cell states. This dissertation presents the development of an improved approach for single-cell whole genome sequencing that achieves the quantitative metrics of gold-standard assays—including uniform, high genome coverage and high cell throughput—while relying solely on commercially available reagents, chemistry, and instrumentation. This democratization of scWGS technology eliminates the need for specialized equipment or reagents that have limited access to these powerful techniques. The method additionally enables enhanced experimental design flexibility through sample multiplexing and opens the door to dual-omics assessment, capturing both chromatin accessibility and whole genome information from the same nuclei.

I demonstrate the application and further development of this method in two distinct cancer biology systems. First, in cell fusion hybrids derived from murine colorectal carcinoma cells and primary macrophages, and hybrids isolated from human cancer patients, I employ complementary single-cell transcriptomic and epigenetic assays to reveal how these remarkable cells maintain dual tumor and macrophage characteristics. This work identified RUNX1 as a key transcription factor driving metastatic progression and characterized distinct phenotypic states these hybrid cells acquire. Second, I present the ongoing development of a V(D)J-seq coassay for human follicular lymphoma that simultaneously captures genome-wide alterations and endogenous clonal identifiers in the B-cell receptor. This approach hijacks the immune system's natural molecular barcoding to track subclonal evolution with unprecedented resolution, offering new insights into therapeutic resistance and disease progression.

The combined body of work advances the field by introducing a novel, readily adaptable single-cell genomics platform and demonstrating its potential to investigate cancer development

and evolution. By bridging the gap between methodological innovation and biological discovery, this dissertation establishes a foundation for more accessible and comprehensive cancer genomics research, ultimately contributing to the development of precision oncology approaches.

Introduction

PART I: Genomics and Epigenetics

The remarkable fidelity with which cellular identities are established and maintained throughout development and homeostasis belies the underlying complexity of regulatory mechanisms that govern gene expression. Despite sharing identical genomic sequences, cells within a multicellular organism exhibit vastly different transcriptional profiles, morphologies, and functional capabilities. Currently, it is understood that cellular diversity emerges through the coordinated actions including the activation of diverse transcription factors[1], chromatin modifications[2,3], three-dimensional nuclear organization[4–8], and regulatory element accessibility[9–16]—collectively forming what Conrad Waddington conceptualized in the mid-20th century as the "epigenetic landscape" [17,18].

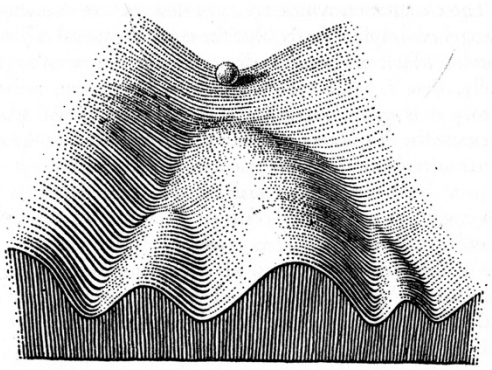


FIGURE 4
Part of an Epigenetic Landscape. The path followed by the ball, as it rolls down towards the spectator, corresponds to the develop–mental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold.

Figure 0: Seminal Waddington Epigenetic Landscape from 1957.[19]
It dpects cellular differentiation as a ball rolling down branching valleys, where each path represents a distinct developmental fate.

Waddington's metaphor depicted cellular differentiation as a ball rolling down a hillside of branching valleys, where each valley represents a distinct cell fate and the topography reflects the regulatory forces constraining cellular development.[19] This metaphor has proven remarkably prescient, with modern molecular biology revealing how Waddington's intuitive landscape emerges from the concrete interplay of multiple regulatory mechanisms. Transcription factor networks establish the basic architecture of cell-type-

specific expression programs[1], while chromatin modifications create heritable marks that reinforce transcriptional states[2,3]. Transcriptional states refer to the patterns of gene expression that define a cell's identity and function, maintained through the coordinated activity of regulatory elements, transcription factors, and chromatin architecture. The three-dimensional organization of chromatin brings distant regulatory elements into physical proximity, enabling coordinated regulation of gene expression across large genomic regions[4–8]. DNA methylation provides an additional layer of epigenetic memory, allowing cells to maintain stable repression of lineage-inappropriate genes[20–22]. Together, these mechanisms create a robust regulatory system capable of generating and maintaining cellular diversity.

The completion of the human genome draft in 2001 represented a transformative milestone in biology[23], yet it also highlighted the limitations of a purely sequence-centric view of heredity. The initial publication focused primarily on euchromatic regions, excluding much of the noncoding DNA that Susumu Ohno had dismissively termed "junk" in 1972[24]. By the turn of the millennium, an incontrovertible body of literature was clarifying this characterization as a profound misnomer. Repetitive transposons, introns, telomeres, and vast heterochromatin regions—previously dismissed if their functionality was not immediately apparent—were increasingly recognized as integral components of genome regulation. Epigenetic modifications to the genome added further layers of complexity that mere sequence information did not capture.

The decades following the initial genome publication witnessed extraordinary technological advances that transformed our ability to map regulatory landscapes. In the late 2000s and early 2010s landmark ENCODE publications utilized cutting-edge technologies to reveal novel molecular mechanisms in the genome's complex regulatory landscape. Furthermore, it connected them to the progression of human disease[4,9–11]. Systematic analyses revealed that disease-associated genetic variants are predominantly enriched in cell-type-specific regulatory elements, with DNase I hypersensitivity mapping demonstrating that common variants linked to autoimmune, metabolic, and neurological disorders localize to enhancers and promoters active in disease-relevant cell

types[10,25]. The insight that genetic risk for disease can operate through disruption of tissue-specific gene regulation rather than protein-coding changes provided a molecular framework for understanding how regulatory dysregulation drives pathogenesis. However, these foundational ENCODE studies relied predominantly on bulk assays that averaged signals across millions of cells, obscuring the cellular heterogeneity fundamental to both normal tissue function and disease progression. This limitation proves particularly problematic for studying complex tissues like tumors, brain, and immune organs, where rare cell populations—including stem cells, metastatic precursors, or therapy-resistant clones—often drive disease outcomes despite comprising a small fraction of the total tissue.

In cancer, this regulatory architecture becomes particularly relevant as both inherited risk variants and somatic mutations converge on the same developmental pathways and cell-type-specific regulatory networks. Tumorigenesis can be conceived as a fundamental disruption of Waddington's epigenetic landscape. Rather than uncontrolled proliferation alone, cancer cells exhibit profound alterations in cell fate regulation, with oncogenic transformation often involving the destabilization of lineage-specific transcriptional programs. The topography of cellular identity becomes distorted through dysregulation of transcription factors, chromatin remodeling complexes, and epigenetic modifiers, creating aberrant cellular states that promote malignant behavior. Critically, bulk sequencing approaches cannot resolve whether these alterations arise uniformly across tumor cells or represent distinct subpopulations with divergent regulatory programs. Either at the onset, or over the course of the progression of the disease, subclonal populations with distinct genomic and epigenomic diversity can emerge. These populations may be invisible to bulk analysis yet potentially determine therapeutic response and resistance. The heterogeneity within tumors is not limited to the cancer’s subclonal diversity. Solid tumors may additionally be composed of immune cells, fibroblasts, and endothelial tissues (i.e. tumor microenvironment). With the potential to be as heterogeneous as any other tissue, molecular understanding of tumorigenesis and tumor development is greatly enabled by the advent of single-cell technologies in the last decade.

Understanding the full spectrum of genomic diversity within a tumor is crucial to precision oncology.

The primary objective of this thesis is to push the envelope on existing single-cell methodologies available for whole genome sequencing, which will be discussed in the next section of this introduction. Further, it aims to apply this novel method to heterogeneous tissues, specifically cancers. Finally, by pairing this method in conjunction with single-cell transcriptomic and single-cell chromatin accessibility assays it aims to elucidate transcriptional diversity in various cancer systems and identify potential regulatory pathways of interest.

PART II: scWGS Assay Development

Motivation

In solid tumors, the cell heterogeneity and clonal diversity resulting from the dynamic, unstable processes of tumorigenesis and tumor evolution is well-established[12–16]. Tumorigenesis – and shortly after – is 'punctuated' by bursts of genomic instability[12] resulting in subclone emergence. These bursts of instability quell to a stable state of ongoing copy number expansion[13]. These diverse subclonal populations have distinct genomic profiles, potentially with mutations that directly impact the tumor's survival, proliferation, and response to therapeutic intervention. Single cell whole genome sequencing (scWGS) is a powerful tool for characterizing the genetic heterogeneity of these cell populations and for studying cancer development[14]. Principally, single-cell assays sidestep the limits of bulk cell strategies that obscure rare variants[14,15]. Assessing the heterogeneity in the tumor genome via copy number variation (CNV) allows for the identification of subclonal populations, understanding a cancer's progression, and determination of potential prognostic mutations[16,26]. A single-cell approach is also worth consideration for impure biopsy samples – such as a liquid biopsy in which the circulating tumor cells represent a small fraction of the total cells present – as it enables discrimination between healthy and malignant cells.

In 2020, the available scWGS technologies were not suitable for most researchers seeking to study such heterogeneous cell populations. Moreover, these technologies were often inaccessible to most researchers, as they are dependent on costly, custom reagents and commercially unavailable instrumentation. Older technologies confronted detection limitations. DOP-PCR[27] suffered from high allelic dropout rates and limited genome coverage, while MDA[28] exhibited extreme amplification bias despite better coverage. MALBAC[29] and LIANTI[30] provided intermediate solutions but remained limited primarily by cell throughput and a high cost per cell, reducing the applicability of the assay. The state of the field motivated filling a technological gap.

While scWGS methods remained limited, single-cell assays for RNA-seq and ATAC-seq had flourished on commercial platforms. The 10x Genomics Chromium platform, widely available

in core facilities, utilized Tn5 transposase-based library preparation for scATAC-seq. The company leveraged their platform for a scWGS kit until 2020[31], following a 2018 patent decision that found for Bio-Rad Laboratories, leaving a persistent gap in the field. Similarly, the Takara ICell8, a second commercially available high-throughput platform, offers potential for increased cell throughput at lower costs[32] for scATAC library generation, when paired with the Adey lab's single-cell combinational indexing strategies[33].

Previously published methods from the Adey lab illustrated that nuclei, even after in situ nucleosome disruption, remain viable for Tn5 transposase-based protocols designed for scATAC. The nucleosome disruption allows for uniform library coverage by disturbing the DNA architecture of the genome and making all of it accessible to the transposase[34]. These advances suggested that existing commercial platforms suitable for scATAC could potentially be adapted for accessible, high-throughput scWGS. In Chapter 1, I will present a method to achieve uniform coverage of the entire genome, at scale, by pairing the in situ nucleosome disruption methods of the Adey lab[33] with the commercially available high-throughput scATAC-seq assay by 10x Genomics.

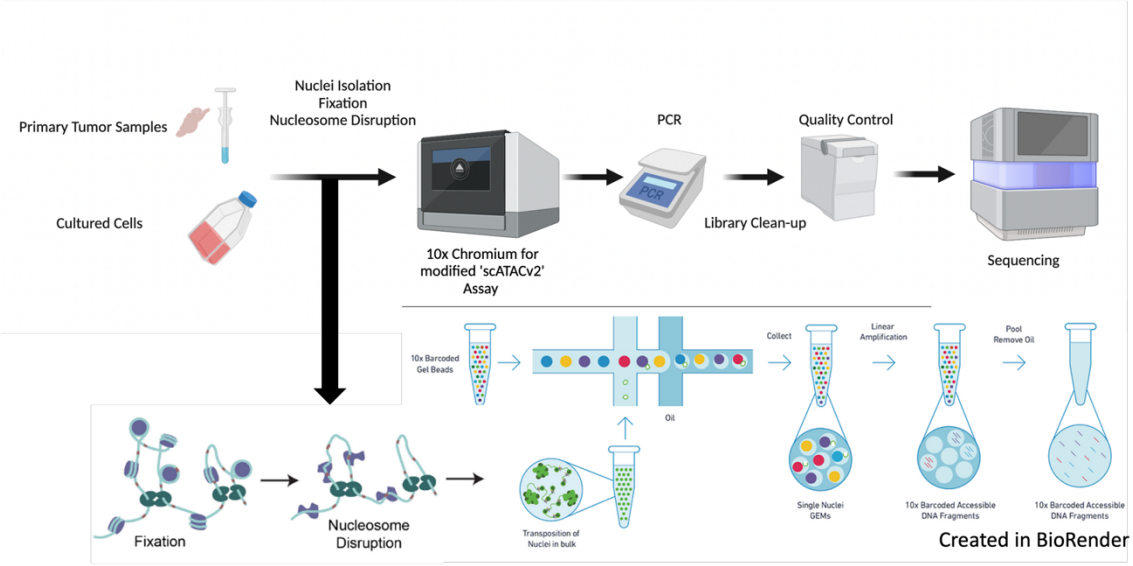


Figure 1: Visualization of scWGS on 10x Genomics Chromium.
A schematic for integration the Adey lab nucleosome disruption techniques with the 10x workflow.

Application and Analysis

The primary quantitative metrics for assessing the scWGS method developed were as follows: potential genome coverage/unique molecules, potential cell throughput/recovery rate, and transcription start site enrichment (TSSe) scores. The number of unique passing reads obtained per nuclei is leveraged as a proxy for potential physical coverage of the genome achieved. These reads must uniquely align to the genome and not be PCR duplicates. The number of unique reads translates to the percentage of the genome directly assayed at nucleotide resolution when taking fragment length into account. The minimum goal of the novel assay was to meet benchmarks for physical coverage from the previously published, lower-throughput 96-well assays – with a ballpark goal of 1e6 unique reads per nuclei – but many more nuclei at a time. TSSe serves a proxy for the success of the nucleosome disruption protocols. Unperturbed chromatin organization would favor transposase access near the TSS, resulting in more reads closely proximal to the site relative to the rest of the genome. Typical TSSe for ATAC data is 5-7 or higher. However, if uniform genome accessibility is successfully achieved, this bias is ablated and a TSSe ratio of 1 or less is expected. Doublet collision rate, sequencing saturation, and MAD Scores were also considered in analysis. Doublet collision rate measures the frequency at which multiple cells are captured within a single droplet and processed as one unit, typically ranging from 0.8-8% depending on cell loading concentration[29]. These events are detected through abnormal read distributions or conflicting genomic profiles and must be identified to prevent artificial "hybrid" cell profiles from confounding downstream analysis. Sequencing saturation, calculated as the fraction of duplicate reads among total reads, indicates whether the library has been exhaustively sampled. A high saturation, typically over 80%[35], suggests diminishing returns from additional sequencing, while low saturation indicates that deeper sequencing would yield more unique genomic coverage. The Median Absolute Deviation (MAD) score provides a robust measure of read depth uniformity across genomic bins that is less sensitive to outliers than standard deviation. In the context of CNV analysis[36], lower MAD scores indicate more consistent coverage across the genome, enabling more

reliable copy number calls, while high MAD scores suggest technical noise that could obscure true biological variation.

RIDDLER[37] was employed to analyze CNV in the proof-of-concept samples. RIDDLER, like other published CNV callers, bins the genome into windows, assesses the read counts within each window, and employs methods to reduce common bias due to amplification, mappability, and GC content. The robust framework allows for further parameters to be considered. AB chromatin domain compartments were included as an additional pertinent factor for the proof-of-concept samples.

For the proof of concept experiments the method was applied to immortalized cell lines, K562 and GM12878, and PDAC patient-derived cell lines. The former served as basic biological samples for optimization and comparison, and the latter served to illustrate the assay's ability to resolve CNV alterations in tumor samples. Subsequently, I applied the scWGS method on murine hybrid fusion cells formed from MC38 CRC and macrophage. There is evidence that the fusion of cancerous cells with leukocytes, specifically macrophages, leads to the development of hybrid cells with uniquely fused genomic profiles[38]. With properties related to both their epithelial and hematopoietic origins[39], hybrid-fusion cells exhibit increased metastatic behavior and represent a possible therapeutic target[38]. Moreover, due to their unique properties, hybrid fusion cells can be readily detected in the peripheral blood of cancer patients and be used as an additional biomarker for tumor diagnosis[40]. A precise scWGS method would be ideally suited to identifying the genomic material retained and lost during hybrid-fusion cells development.

PART III: V(D)J-seq Assay Development

Motivation

Follicular lymphoma (FL) accounts for an estimated 30% of all lymphomas and is the second most prevalent non-Hodgkin lymphoma[41]. The generally incurable malignancy arises in germinal center B-cells[42]. B-cells are lymphocytes responsible for humoral immunity, which they achieve by recognizing antigens via their surface immunoglobulin receptors and differentiating into antibody-secreting plasma cells or memory B-cells upon activation[43]. Germinal centers are specialized microstructures within secondary lymphoid organs where activated B-cells undergo rapid proliferation, somatic hypermutation, and affinity maturation to generate high-affinity antibodies against specific antigens[44,45]. FL can transform into more aggressive diffuse large B-cell lymphoma (DLBCL), with approximately 30 - 40% of patients experiencing relapse[46]. Determining the molecular underpinnings of the disease is highly desired, with the hope that it will inform future standards of care.

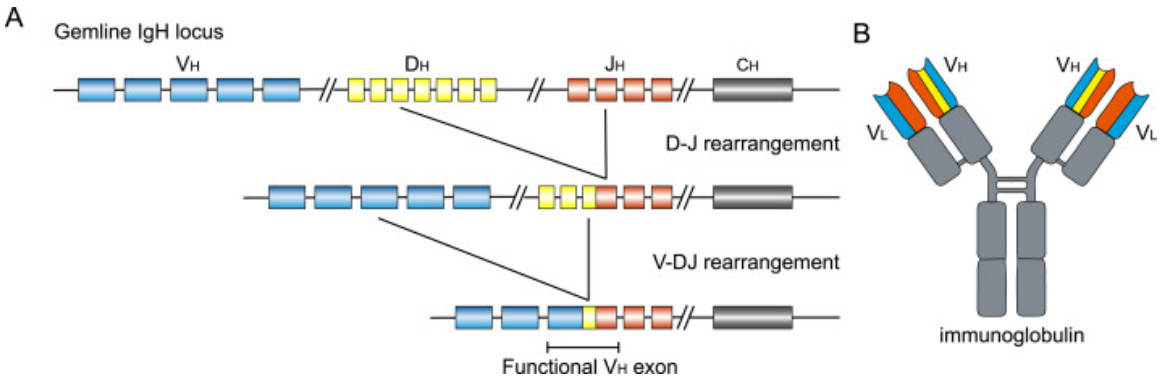


Figure 2: Visualization of V(D)J Recombination.
V(D)J recombination is responsible for the tremendous diversity in antigens B-cells can recognize[47].

The adaptive immune system's ability to recognize diverse antigens relies fundamentally on V(D)J recombination, a process that generates the extraordinary diversity of B-cell receptors[48,49]. During B-cell development, the variable (V), diversity (D), and joining (J) gene segments undergo random recombination to create unique heavy and light chain combinations. This process, combined with junctional diversification at the recombination sites, can theoretically generate over

10^{11} different BCR sequences. Following successful recombination, B-cells undergo clonal expansion and affinity maturation within the germinal center through somatic hypermutation, fine-tuning their antigen specificity. Each B-cell clone that exits the germinal center carries a unique V(D)J molecular "barcode" that serves as both its functional identity for antigen recognition and a permanent molecular barcode of its developmental history.

A hallmark of early FL development involves the acquisition of de novo glycosylation motifs in the B-cell receptor (BCR) due to aberrant somatic hypermutation[50]. In FL, this ongoing hypermutagenic process continues to modify the already unique V(D)J sequences, creating additional layers of molecular diversity within tumor populations, with a higher frequency[51] than in the rest of the already plastic cancer genome[52–54]. This aberrant hypermutation presents a unique opportunity to exploit V(D)J sequences as high-resolution molecular identifiers for tracking subclonal populations within the disease, as each malignant clone harbors its distinct recombination signature while acquiring tumor-specific mutations.

Prior work conducted by collaborator, Dr. Tanya Shree, established that FL exhibits detectable site-to-site heterogeneity in patients[55]. This heterogeneity extends beyond BCR evolution to differential tumor gene expression and cell-surface proteins[52], which raises the question of whether the distinct populations explain the observed differential responses to systemic therapies among patients. This knowledge gap motivated the further optimization of the previously described high throughput scWGS assay into a coassay that additionally captures the rearranged V(D)J region as clonal identifier. Once established, the coassay would allow for an in-depth genomic profiling of a patient's disease, a characterization of subclonal populations, and the tracking of subclonal responsiveness to administered therapeutic interventions in longitudinal studies. Furthermore, recently published sciMET protocols[56] could be readily integrated in place of the scWGS assay to capture additional epigenetic information, potentially revealing genome-wide regulatory mechanisms driving the cancer progression. sciMET is a single-cell methylation sequencing method developed by the Adey Lab that combines sciATAC-seq style combinatorial

indexing with bisulfite conversion, or enzymatic conversion in sciMETv3, to simultaneously profile chromatin accessibility and DNA methylation from the same single cell. Genome-wide methylation information may be of particular interest as dysregulation of genes related to DNA methylation is associated with worse prognosis in FL and appears to be subject to positive selection during relapse[57,58]. In a final addendum to this thesis, I will present a coassay seeking to capture both tagmented genomic fragments for scWGS and specifically the V(D)J region. Moreover, both types of fragments will be indexed such that after demultiplexing fragments from both assays can be attributed to the original individual nuclei.

Application and Analysis

The computational analysis pipeline for the coassay leverages established bioinformatic tools adapted to handle the unique dual-modality nature of the data, where genomic alterations and clonal lineage information must be integrated at single-cell resolution. The pipeline addresses three key analytical challenges: accurate demultiplexing of combinatorially indexed libraries, robust detection of copy number variations and single nucleotide variants from low-coverage scWGS data, and reconstruction of B-cell clonal lineages from V(D)J sequences. Together, these tools enable the linking of specific genomic alterations to distinct B-cell clones, providing unprecedented resolution of tumor heterogeneity and clonal evolution in FL.

Analysis of the coassay follows a two-pronged approach. First, scWGS and V(D)J amplified reads are attributed to individual cells via unidex (https://github.com/adeylab/unidex) demultiplexing, leveraging the three levels of indexing introduced during library preparation.

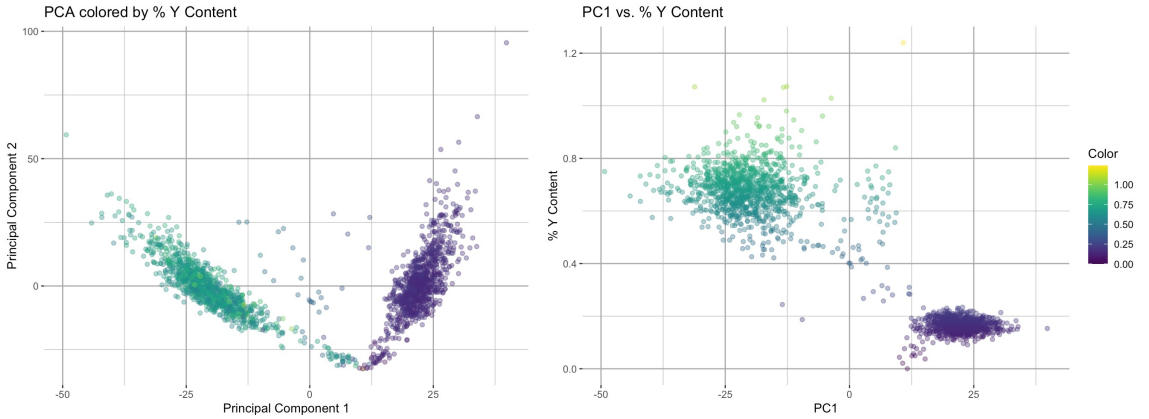


Figure 4: Example PCA Clustering from Proof-of-Concept B-Cell scWGS Experiments. scWGS libraries were processed for CLL cells from two individuals (one male and one female), sequenced, and analyzed via GATK and vcfR. The samples readily separated on percentage of Y content as a proof of concept for the computational and analysis pipeline.

The RIDDLER framework, previously described in Part II, is utilized to interrogate potential CNV in distinct FL clonal populations. The unbiased approach of the scWGS assay allows for the characterization of novel CNV across FL patient subclones. Previously identified CNVs in FL proved complementary to mutation data, demonstrating alterations to the S1PR1 and S1PR2 drove disease progression[59].

As a complement, the widely adopted Genome Analysis Toolkit (GATK) is employed for SNV calling and haplotype determination[60]. Variant call files (VCFs) generated from GATK are further processed using CellSNP-lite[61], which attributes the SNV information to individual cell barcodes. The variant data is further visualized with VCFR[62] which enables dimensionality reduction and clustering of distinct clonal populations in R (Fig. 4).

Change-O proved effective for V(D)J analysis and visualization, following evaluation of existing computational pipelines[60,63,64]. The well-documented and maintained toolkit demonstrated superior flexibility in handling unexpected read structures, which is invaluable for assessing reads with combinatorially indexed barcodes. Change-O identifies clonal lineage relationships by first grouping sequences that share the same V and J gene assignments and similar CDR3 junction lengths. Next, the tool hierarchically clusters them based on nucleotide similarity within the V segment, using mutational hotspot-aware distance models that reflect the biology of somatic hypermutation. Sequences exceeding a data-driven similarity threshold are assigned to the same clonal lineage, inferring common ancestry from a single naïve B-cell rearrangement event. Within each of these defined clones, Change-O reconstructs phylogenetic lineage trees via maximum-parsimony methods (Fig. 5). This clonal grouping and tree-building enables robust identification of clonal populations, generation of lineage trees, and a detailed assessment of somatic hypermutation motifs. Using the corresponding indexed barcodes to pair the two assays conducted in the same nuclei, SNVs observed in the scWGS data can be more precisely resolved, by taking advantage of the higher mutability of V(D)J regions.

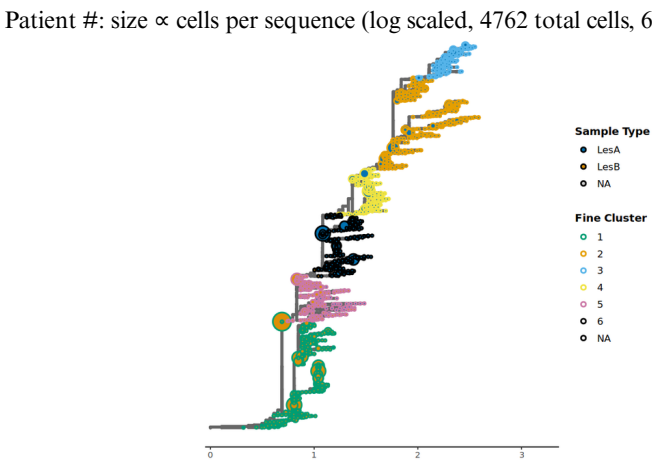


Figure 5: Example Lineage Trees by Change-O.
The application of Change-O clonal clustering and lineage reconstruction on a test dataset identified six distinct clonal lineages, demonstrating as a proof of concept the ability to resolve V(D)J sequences arising from a common B-cell progenitor.

Chapter 1: Accessible high-throughput single-cell whole genome sequencing with paired chromatin accessibility

This section is based on an article previously published as: Queitsch K, Moore TW, O'Connell BL, Nichols RV, Muschler JL, Keith D, Lopez C, Sears RC, Mills GB, Yardımcı GG, Adey AC. Accessible high-throughput single-cell whole-genome sequencing with paired chromatin accessibility. Cell Rep Methods. 2023 Nov 20;3(11):100625. doi: 10.1016/j.crmeth.2023.100625. Epub 2023 Nov 1. PMID: 37918402; PMCID: PMC10694488.

Authors collaborating in this work and affiliations

Konstantin Queitsch[1], Travis Moore[2,3], Brendan L. O'Connell[1,2,3], Ruth V. Nichols[1], John L. Muschler[3,4, 5], Dove Keith[5], Charles Lopez[3], Rosalie Sears[1,2,3,5], Gordon B. Mills[3,6], Gurkan Yardimci[2,3], Andrew C. Adey[1,2,3,7] *

1. Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA
2. Cancer Early Detection Advanced Research Center, Oregon Health & Science University, Portland, OR, USA
3. Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA
4. Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA
5. Brenden-Colson Center for Pancreatic Care, Oregon Health & Science University, Portland, OR, USA
6. Department of Cell, Developmental and Cancer Biology, Oregon Health & Science University, Portland, OR, USA
7. Knight Cardiovascular Institute, Oregon Health & Science University, Portland, OR, USA

* Lead Contact for correspondence: adey@ohsu.edu

Author contributions

A.C.A. conceptualized the method. A.C.A., B.L.O. and K.Q. designed experiments, K.Q. performed all experiments with assistance from B.L.O., R.V.N., G.Y., T.M. devised RIDDLER, T.M. performed copy number calling analysis, A.C.A., B.L.O., R.V.N. and K.Q. performed analysis, J.L.M., D.K., C.L., G.B.M. and R.S. managed the banked patient-derived samples.

Abstract

Single-cell whole-genome sequencing (scWGS) enables the assessment of genome-level molecular differences between individual cells with particular relevance to genetically diverse systems like solid tumors. The application of scWGS was limited due to a dearth of accessible platforms capable of producing high-throughput profiles. We present a technique that leverages nucleosome disruption methodologies with the widely adopted 10× Genomics ATAC-seq workflow to produce scWGS profiles for high-throughput copy-number analysis without new

equipment or custom reagents. We further demonstrate the use of commercially available indexed Tn5 transposase complexes from ScaleBio for sample multiplexing, reducing the per-sample preparation costs. Finally, we demonstrate that sequential indexed tagmentation with an intervening nucleosome disruption step allows for the generation of both ATAC and WGS data from the same cell, producing comparable data to the unimodal assays. By exclusively utilizing accessible commercial reagents, we anticipate that these scWGS and scWGS+ATAC methods can be broadly adopted by the research community.

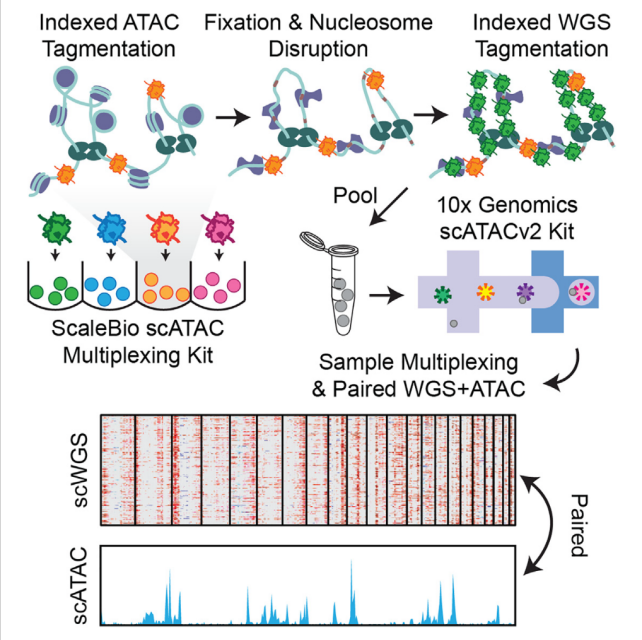


Figure 6: Graphical Abstract for Chapter 1.

Main Text

Single-cell whole genome sequencing (scWGS) has proven invaluable for the identification of tumor subpopulations and in advancing our understanding of clonal dynamics and tumor evolution[14,65–69], as well as in the study of somatic genomic copy number alterations in healthy tissues, including the brain[70]. Despite the value of scWGS, there is a dearth of commercially available options that provide enough cell throughput (i.e. power) to fully catalogue and characterize clonal populations within a tumor sample. This can be particularly problematic when rare, possible therapy-resistant, subpopulations are present within a tumor that may elude detection using existing methodologies. To compound these challenges, assay availability was further

reduced when the widely-available 10x Genomics Chromium, permanently discontinued its scWGS product.

We previously developed several technologies that leverage *in situ* disruption of nucleosomes such that nuclei remain intact and can be processed through *in situ* tagmentation methods typically used for single-cell ATAC-seq. The nucleosome disruption, which is performed by fixation and then treatment with a detergent, enables genome-wide access by the Tn5 transposase, thus producing random coverage across the entire genome. In prior technologies, combinatorial barcoding was used to enable single-cell profiling[71]; however, these methods require boutique reagents in the form of at least 96 tagmentation complexes with unique DNA barcodes, making the technology difficult to access for the majority of potential users. Here, we leverage similar nucleosome disruption methods, *in situ* tagmentation and cell barcoding, using off-the-shelf kits, with the widely accessible Chromium instrument from 10x Genomics. We further leverage commercially available indexed tagmentation reagents (24-plex), designed for use in scATAC sample multiplexing, to achieve multiplexing of scWGS samples in individual lanes of a Chromium chip. The same indexed tagmentation reagents lend themselves to a novel double tagmentation technique that enables the production of both ATAC and WGS data from the same single cells, that can additionally capitalize on the beforementioned sample multiplexing capabilities.

Single cell whole genome sequencing using the 10x Chromium scATAC workflow

To achieve high-throughput single-cell whole genome sequencing (scWGS) using off-the-shelf reagents, we reasoned that the nucleosome disruption technology which we had developed for combinatorial indexing based scWGS could be deployed within the single-cell ATAC-seq technology workflow that is commercially available from 10x Genomics (Figure 7A). Key to this workflow is the successful disruption of nucleosomes to enable genome-wide access for the Tn5 transposase to fragment DNA and append adapters (tagmentation), and that the nuclei remain intact

during the process for subsequent processing using droplet fluidics to index each library fragment with a corresponding cell barcode.

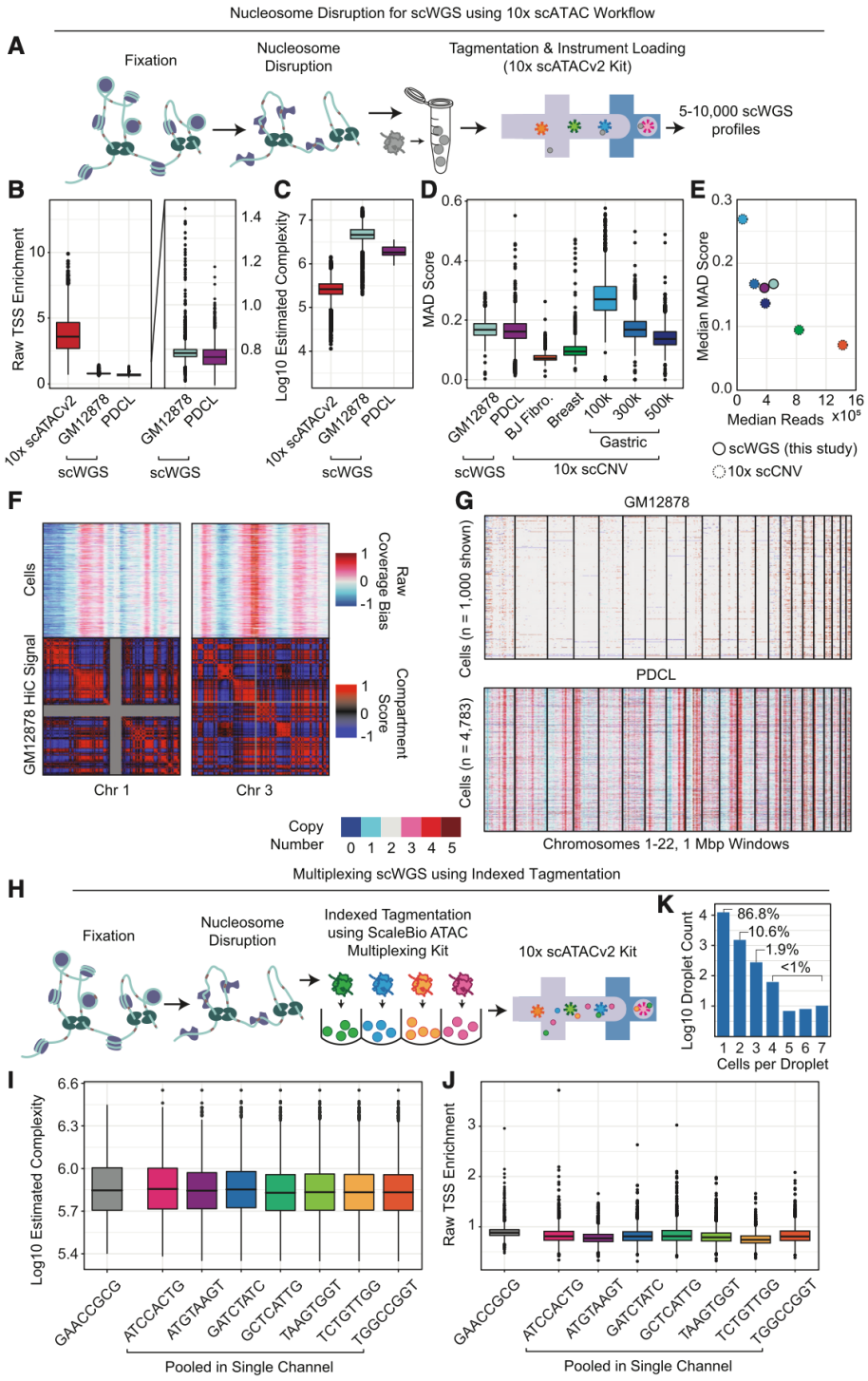


Figure 7: Single-cell WGS using the 103 Genomics scATAC kit.
(A) Workflow diagram. (B) Raw TSS enrichment for each condition. PDCL,patient-derived cell line. Boxes indicate 25th and 75th percentile, whiskers represent 90th and 10th percentile. These boxplot settings are used for all subsequent figures. (C) Estimated library complexity. (D) Comparison of MAD scores to assess coverage biases between our technique and the discontinued 103 scCNV kit. (E) Median MAD scores as a function of median read depth. Colors are as in (D). (F) Association between raw coverage bias and HiC compartments. (G) RIDDLER copy-number calls for GM12878 and PDCL samples. (H) Multiplexing workflow diagram. (I) Estimated library complexity. (J) Raw TSS enrichment for each indexed sample. (K) Cells per droplet assessed by the number of unique tagmentation barcodes associated with each droplet barcode.

We previously described two versions of the nucleosome disruption technique: Lithium Assisted Nucleosome Disruption (LAND) and a method that relies on crosslinking followed by sodium dodecyl sulphate detergent treatment (xSDS)[72]. We found that LAND, which leverages the chaotropic salt, Lithium diiodosalycylate (LIS), was highly variable in its ability to retain nuclear integrity, often resulting in the rupture of all nuclei, likely due to the salt readily crashing out of solution at room temperature, resulting in variable effective concentrations[72]. We therefore focused on the xSDS approach, which we previously optimized for other combinatorial indexing-based technologies and have applied to a variety of tissues[71]. For the proof of principle, we employed the control lymphoblastic cell line, GM12878. Half of the cells were preprocessed for nucleosome disruption, and the control half was undisturbed, prior to being carried through the standard 10x Genomics scATACv2 workflow. Subsequently, we nucleosome disrupted cells from a patient-derived pancreatic cancer cell line, PDCL1, and assayed them in the same manner. Sequenced libraries were taken through barcode deconvolution and cell calling based on a unique read threshold and assessed for key metrics (Methods).

We first assessed the ability of our nucleosome disruption technique to ablate the chromatin accessibility signal by assessing transcription start site enrichment (TSSe), which is a standard measurement for ATAC-seq signal evaluation. As with our previous nucleosome disruption techniques for whole genome as well as DNA methylation techniques[73,74], the TSSe was reduced to below 1 for both the GM12878 control line (median = 0.783092) and the PDCL (median =

0.685304) using a raw TSS enrichment value that calculates the ratio of reads within a 500 bp window centered on TSSs over reads within 250 bp windows ± 1 kbp of TSSs which produces a TSSe ≤ 1 for shotgun WGS data[73] (Figure 7B, Methods). This is in contrast to the scATAC control which exhibited a median raw TSSe of 6.02, which is consistent with high-quality scATAC preparations using this calculation method[73]. We next assessed the complexity of each library by estimating the total possible unique observed sequence reads if libraries were sequenced to saturation (Methods), enabling a direct comparison that is not impacted by variation in raw sequencing depth. This revealed a high complexity library for the scATAC control, resulting in a median estimated unique read count of 260,846. The scWGS preparations produced a higher complexity due to the ability to tagment and sequence the entire genome, with median estimates of 4,603,746 and 1,950,670 for the GM12878 control and PDCL, respectively (Figure 7C). The reduced complexity of the aneuploid PDCL may be due to the increased DNA content, which alters the effective ratio of tagmentation complexes and DNA present or possibly the increased gDNA within the nucleus reduces the ability of tagmentation complexes to penetrate throughout the entire volume.

Copy number calling of single-cell WGS data typically involves binning the genome into windows and assessing the read counts for each cell in each of the windows followed by methods to mitigate amplification or other biases – often correlating with GC content[75,76]. We assessed counts across non-overlapping 1 Mbp windows and first assessed the Median Absolute Deviation (MAD) scores, a measurement of overall bias in WGS data[76], and compared our values to released datasets produced using the now-discontinued 10x Genomics scCNV kit including a diploid fibroblast cell line and aneuploid breast cancer tissue sequenced to high depth, and then an aneuploid gastric cancer cell line (MKN-45) sequenced at increasing levels of coverage[77] (Figure 7D). At the sequencing depth we obtained for the GM12878 and PDCL experiments, we observed slightly greater median MAD score than the 10x Genomics scCNV datasets (Figure 7E). We next assessed global coverage and noticed that the primary bias in window counts correlated with

chromatin domains (Figure 7F). We suspect that this is due to the tagmentation process occurring *in situ*, where tagmentation efficiency is likely correlated with proximity to the nuclear periphery. We therefore included chromatin domain as a factor in copy number assessment using RIDDLER, a versatile copy number calling tool that leverages robust Poisson regression to reliably detect outlier windows in each cell. This produced the expected copy number neutral profiles for the GM12878 cell line and expected copy number aberrant profiles for the PDCL line (Figure 7G).

Multiplexing samples for scWGS using indexed tagmentation reagents

The ability to produce large cell count scWGS datasets shifts the primary cost burden of any given study to the sequencing depth that is required. Cell throughput from a single 10x Genomics fluidics channel can reach an excess of 10,000 cells, which results in a prohibitive sequencing cost for any individual sample for most scWGS studies. Furthermore, an assessment of the clonal diversity of most samples can typically be deconvolved with far fewer profiles[71]. We therefore leveraged the scATAC multiplexing kit, commercially available from ScaleBio, that uses indexed tagmentation to pre-index samples prior to loading onto the chip. This can be used for standard loading as well as 'superloading', where multiple pre-indexed nuclei can be processed within the same droplet[78]. The classic approach is to simply load the desired number of cells of each sample into an individual fluidics channel; however, this results in greater library preparation costs, with each sample requiring its own set of reagents.

We carried out two preparations leveraging indexed tagmentation reagents using the GM12878 control cell line. The first leveraged a single tagmentation complex with 5,000 nuclei loaded into a single channel to serve as a non-multiplex control using the indexing reagents. The second pooled approximately 3,000 nuclei from each of seven indexed tagmentation reactions which was also loaded onto a single channel (Figure 7H). After index demultiplexing (Methods) the single index channel produced 3,678 cell profiles and the multiplex channel produced 16,280, with a mean of 2,326 ± 352 cells per tagmentation index (median 2,289), with equivalent estimated

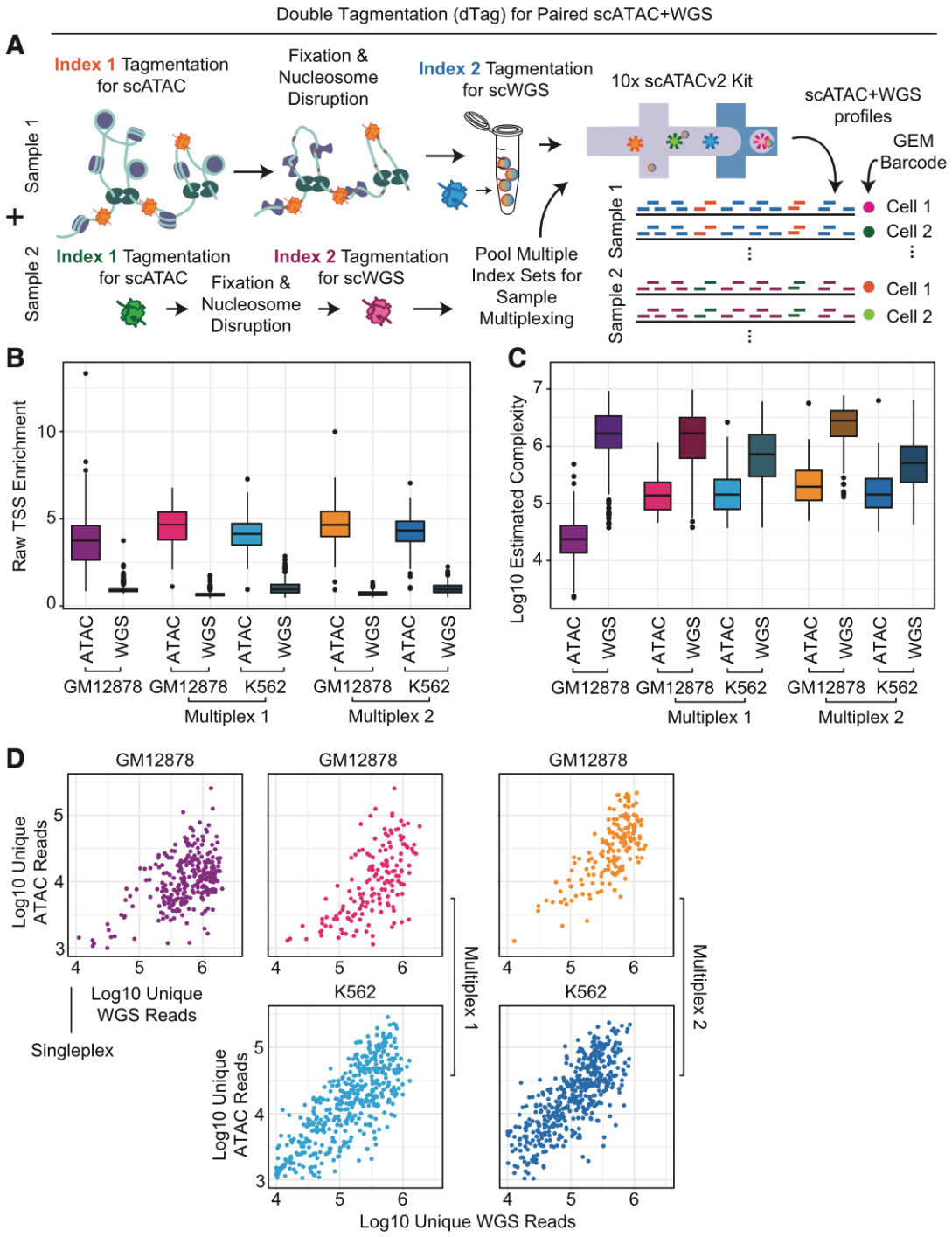


Figure 8: Double-tagmentation for paired ATAC+WGS.
(A) Workflow diagram. (B) Raw TSS Enrichment for each modality for each sample. (C) Estimated library complexity. (D) Log10 unique reads for each cell for ATAC vs WGS modalities.

complexity and TSS enrichment (Figure 7I, BJ). We then assessed the distribution of cell counts across droplets which revealed that the large majority (86.8%) of cells were the only cell within the droplet with the remaining droplets containing two (10.6%) or more (2.6%) pre-indexed cells (Figure 7K).

Double tagmentation with indexed complexes enables scWGS + scATAC from the same cell

The ability to perform indexed tagmentation and deconvolution within a single droplet, raises the possibility of leveraging multiple indexes per cell to encode separate properties. We reasoned that separate indexed tagmentation complexes could be used in succession to profile both chromatin accessibility and genomic DNA from the same cells and developed a double tagmentation (dTag) workflow (Figure 8A). Nuclei are first isolated according to scATAC workflows and carried through tagmentation across several indexed tagmentation complexes using the commercially-available scATAC multiplexing kit. Importantly, the ATAC tagmentation is performed prior to cell fixation, as in the standard scATAC workflow, thus avoiding any potential biases associated with tagmenting fixed DNA. Replicate wells are then pooled to achieve enough nuclei to perform fixation and nucleosome disruption followed by a second round of tagmentation with a second, distinct tagmentation index. dTag nuclei can then be pooled with other samples that leverage other sets of tagmentation indexes and loaded onto a 10x Genomics scATAC channel for droplet encapsulation and bead-barcoding followed by cleanup, PCR amplification, and sequencing.

We carried out an initial preparation containing only GM12878 cells that were tagmented using a single index for the ATAC component across multiple reactions and then a single index for the whole genome component. After the second tagmentation approximately 2% of the post nucleosome-disrupted nuclei recovered, all of which were loaded onto a single fluidics channel. After sequencing and index deconvolution, we assessed the TSSe for each matched component of each cell which produced a median of 3.75 for the ATAC component and 0.89 for the WGS

component (Figure 8B), matching what we observed with single-modality assays (median of 6.02 and 0.78 for ATAC and WGS, respectively). Similarly, the estimated total unique library molecules also reflected the balance of fewer ATAC reads compared to WGS at a median of 23,676 and 1,652,516, respectively (Figure 8C).

We next assessed the ability to multiplex samples using distinct index sets within the same scATAC fluidics channel by processing both GM12878 and K562, a human chronic myelogenous leukemia cell line, in parallel. To improve post-WGS tagmentation nuclei survival we also assessed increased formaldehyde fixation conditions prior to nucleosome disruption after the initial ATAC tagmentation reaction (1% and 1.25% for Multiplex 1 and Multiplex 2, respectively, versus 0.75% for the initial experiment). For each ATAC tagmentation we leveraged 5 sets of indexes for each sample which were then pooled for fixation, nucleosome disruption, and WGS tagmentation. The increased formaldehyde concentration resulted in improved nuclei survival rates (approximately 5-fold for 1% and ~2-fold for 1.25% fixation), though still lower than WGS tagmentation alone. For each preparation, we loaded nuclei, pooling the same fixation conditions for the GM12878 and K562 preparations within each channel. Loaded nuclei counts were limited due to the costs associated with sequencing the WGS component, targeting approximately 3,000 total nuclei loaded per channel. Passing cell profiles were comparable between conditions for each cell line with GM12878 producing 165 and 169 cells and K562 producing 332 and 392 cells for the 1% and 1.25% fixation conditions, respectively (Methods). Yields were notably lower than expectations based on the loaded count; however, a clear and distinct population of cell barcodes that contained a high count of WGS and ATAC reads was observed (Figure 8D), suggesting that the low cell count was due to an underestimation of the loaded concentration as opposed to a high failure rate.

Consistent with previous preparations, WGS reads associated with each sample produced a raw TSSe centered near 1, indicating ablation of the chromatin accessibility signal (Figure 8B). Each preparation also produced comparable estimated total unique library molecules with a median over 1 million for each GM12878 preparation (1,684,065 and 2,797,486) and somewhat less for

the K562 cells (714,419 and 507,716) (Figure 8C), which is also consistent with our scWGS standalone datasets that produced reduced coverage for cells that contain greater amounts of gDNA. Notably for both TSSe and library complexity the fixation percentage did not appear to have any impact, making the higher nuclei yields of the 1% fixation condition favorable overall.

To evaluate the WGS portion of the libraries, we reasoned that inclusion of the ATAC reads for each cell is warranted for genomic copy number calling, as WGS-alone assays do not exclude accessible regions, they just do not enrich for them. Using the combined data, we first assessed the MAD score coverage uniformity producing values that fell within the same range (median <0.3, Figure 9A) as the standalone scWGS datasets and the 10x Genomics scWGS datasets using the discontinued kit (Figure 7D,E). We next evaluated how many cells were required to achieve genome-wide physical coverage by randomly sampling cells and aggregating reads to achieve the mean fold-coverage of the genome and percent of the genome covered (Figure 9B). Consistent with complexity observations, coverage was reduced for aneuploid cells (PDCL and K562) versus the diploid GM12878 cell line. For GM12878 conditions other than our initial dTag singleplex experiment prior to optimization, cell counts required to reach 5-fold genome coverage, which is sufficient for cluster-based genotyping, was 20 versus 70. Between GM12878 conditions, the standalone produced the highest fold-coverage and was able to achieve 30-fold coverage at 80 cells, which is considered sufficient for de novo variant calling.

Using the combined data for each sample we then performed copy number calling using RIDDLER as detailed previously (Figure 9C). Visually, increased noise was observed for the dTag GM12878 copy number profiles when compared to the standalone assay (Figure 7G). To assess this in greater detail, we calculated the distribution of copy number calls against each integer for the dTag experiments compared to the GM12878 standalone scWGS assay (Figure 9D). This assessment revealed the proportion of copy number 2 calls was 91.7, 90.8, and 94.6 percent for the dTag multiplex 1, 2 and dTag singleplex experiments respectively. This represents a decreased proportion when compared to the 95.9 percent for the standalone scWGS workflow and the

theoretical expectation of 100 percent. The discrepant windows tended to fall in regions that exhibited chromatin domain signal bias, such as the q-arm of chromosome 1, suggesting that increased noise in the dTag scWGS dataset may reduce the efficacy of the bias reduction performed by RIDDLER.

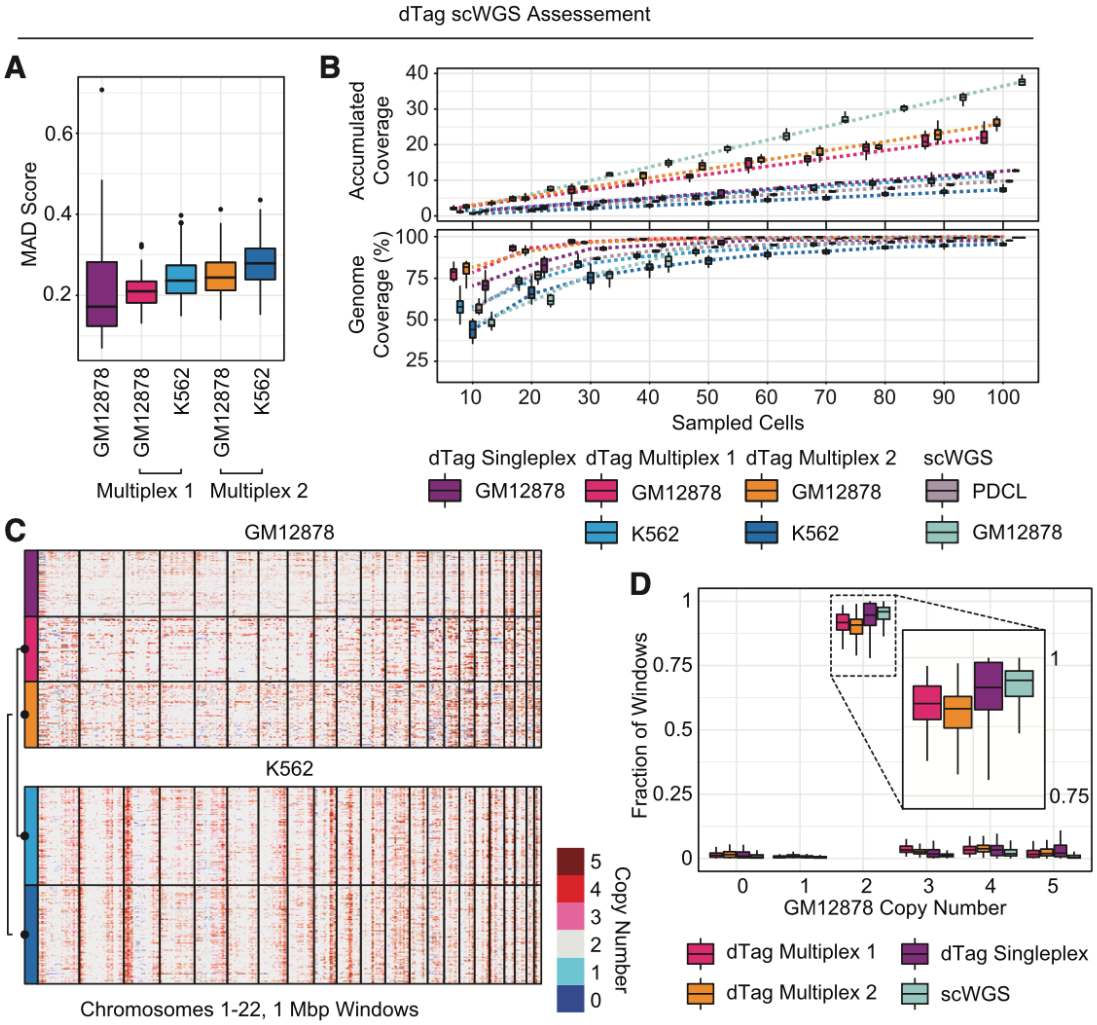


Figure 9: dTag WGS performance.
A) MAD scores for dTag scWGS data. (B) Accumulated fold-coverage (top) and percent of genome covered (bottom) for random subsampled cell increments. (C) RIDDLER copy number profiles for dTag scWGS cells. (D) Distribution of copy number calls for GM12878 compared between scWGS and dTag scWGS methods. CN=2 is highlighted.

Next, we evaluated the ATAC component of the dTag preparations by producing fragment files and evaluating performance alongside the GM12878 scATAC control preparation. TSS Enrichment as assessed using the ArchR software suite revealed comparable enrichment across conditions with a slight increase in enrichment for the dTag preparations [79] (Figure 10A). This observation may be due to the increased proportion of sub-nucleosomal fragments that occurs due to the second tagmentation for WGS that likely disrupts these fragments leaving only the end tagmentation events from the initial ATAC reaction (Figure 10B). Despite the difference in fragment size, the two cell types produced distinct profiles that separated cleanly on a UMAP projection with GM12878 cells from both the unimodal assay and dTag conditions co-embedding with one another (Figure 10C). When examining genomic signal tracks, clear ATAC peaks were observed across each experiment with distinct differences between the GM12878 and K562 cell lines (Figure 10D).

Finally, we assessed the ability to call peaks on pooled cell profiles from our dTag experiments and the standalone 10x scATAC dataset. Given that peak calling power correlates strongly with depth of sequencing, we downsampled the standalone 10x scATAC dataset to several increments covering a comparable range to our dTag scATAC datasets and performed peak calling on each sample as well as the merged reads from all dTag scATAC datasets. This produced slightly higher peak call counts using the standalone assay versus the dTag multiomic workflow (Figure 10E), with a similar trend of increased peak calls based on the total reads used in the peak calling. To further explore these calls, we assessed the percent overlap of peaks with public annotations of putative regulatory elements, including GeneHancer (GH) and ENCODE DNase Hypersensitivity Sites (DHS) (Figure 10F). This produced comparable percentages for the Multiplex 2 dataset versus the 10x scATAC standalone conditions with comparable coverage and a reduced percentage for other dTag conditions.

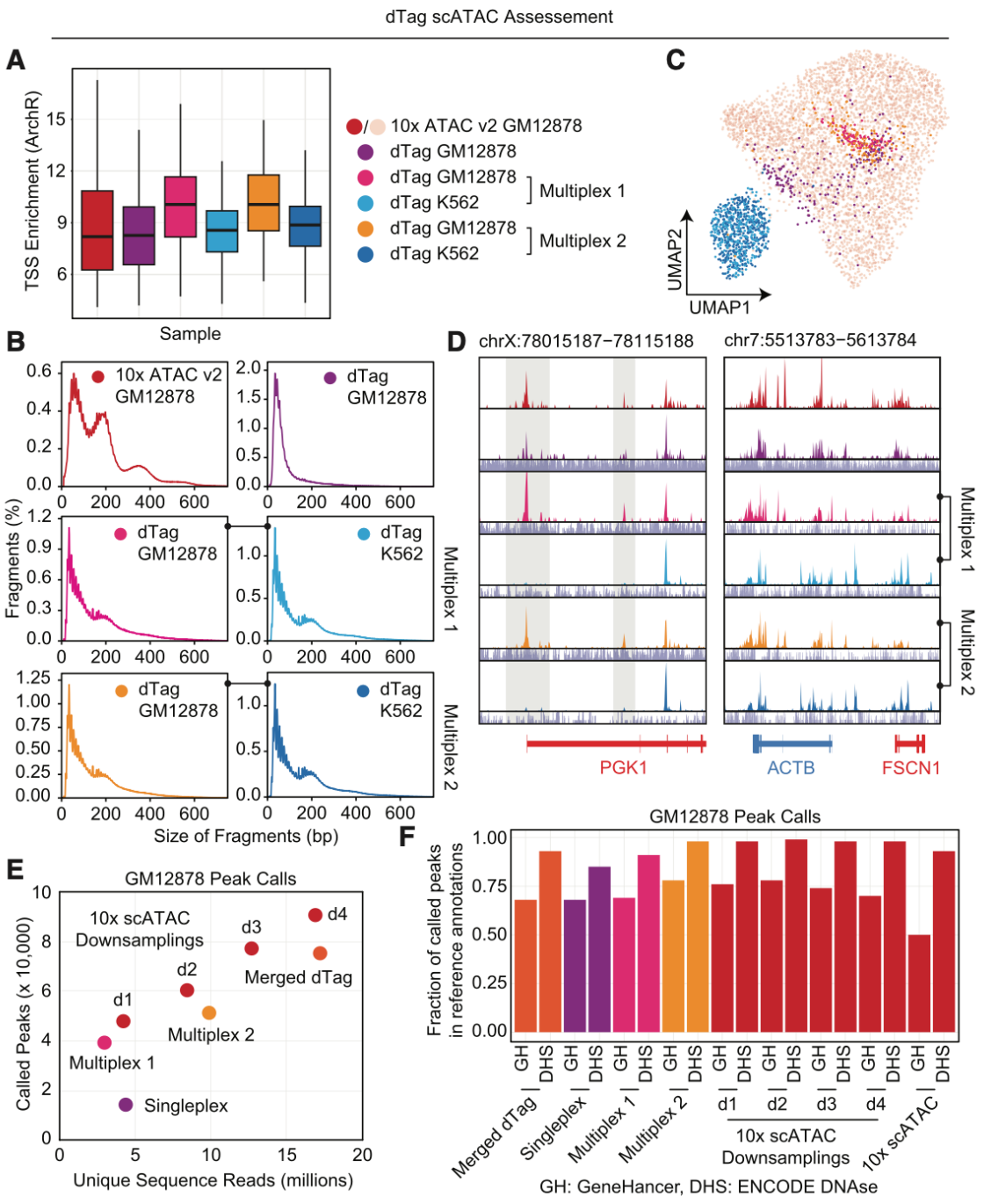


Figure 10: dTag ATAC performance.

(A) ArchR TSS Enrichment. (B) ATAC modality fragment size distribution. (C) UMAP of standalone scATAC (light red) and dTag scATAC conditions. Respective cell lines group together. (D) ATAC coverage tracks for housekeeping genes shows the same patterning for GM12878 for standalone and dTag assays, with clear differences in the K562 line highlighted. Below each track is the corresponding scWGS track. (E) Called peaks as a function of unique sequence reads for the dTag scATAC conditions as well as four downsampled datasets from the scATAC standalone assay. (F) Percentage of peak calls in GM12878 scATAC standalone and dTag datasets compared to GeneHancer annotated regions (GH) and ENCODE DNase Hypersensitivity sites (DHS).

Discussion

In this work, we detail the adaptation of our previously described nucleosome disruption technologies to enable the acquisition of single-cell whole genome sequencing data using off-the-shelf kits for the widely adopted 10x Genomics Chromium instrument. These techniques seek to fill the gap in high-throughput scWGS with the goal of profiling copy number alterations without the need to purchase new instrumentation, depend on bespoke enzymes, or resort to custom sequencing chemistry. The primary motivation was demonstrably achieved via alterations to the standard 10x Genomics workflows using readily-available reagents. The scWGS data produced is of high complexity, producing sufficient read depth for copy number calling genome-wide, and we expect that the ease of access to the technology will make it an appealing option for studies that rely on copy number assessment for tumor clonal structure and evolution analysis.

We next utilized a commercially available scATAC multiplexing kit (ScaleBio) to enable one to load multiple samples in a single chip lane without a loss in the quality of the data. This flexibility allows many more samples to be queried on a full chip; or alternatively, allows for a rationed consumption of 10x barcoding reagents driving down library preparation costs. The improved flexibility granted by pre-indexing also laid the foundation for the dTag protocol to achieve ATAC plus WGS from the same cells. The dTag workflow leverages an initial round of tagmentation with indexed Tn5 followed by nucleosome disruption and then a second tagmentation using differently indexed Tn5 complexes. We also demonstrate that separate samples can be multiplexed in the same 10x Genomics instrument channel if separate sets of indexes are utilized. Notably, the scWGS data generated by the dTag method matches the quality and complexity of the scWGS protocol we present; and the scATAC component matches closely to standalone 10x Genomics scATAC data. Taken together, dTag, while using exclusively commercially available reagents, promises researchers the ability to assess heterogeneous samples for copy number alternations in individual cells, and additionally use accessibility information to evaluate the putative functional impact on the epigenome within copy number altered loci.

Limitations of the Study

The scWGS standalone assay is capable of delivering genome-wide coverage sufficient for copy number calling to enable tumor subclone identification; however, we do observe a bias in raw coverage that correlates with broad chromosome compartments. We believe that this bias is due to the high-level structure of chromatin in the nucleus with so-called "A" and "B" compartments that are positioned in either the nuclear interior or nuclear periphery, respectively, and that the Tn5 enzyme has variable penetration into the core of the nucleus, resulting in the observed coverage bias. Fortunately, the broad chromatin compartment structure has been shown to be consistent across cell types[80], and can be readily accounted for during the copy number calling process. The other limitation to the scWGS approach is that it does not provide sufficient coverage for *de novo* single nucleotide variant calling due to the lack of a genome amplification step – something that is shared across all direct tagmentation scWGS techniques[81], restricting its application to copy number assessment or the genotyping of variants identified in an aggregate dataset, or variant calling within copy-number-defined clusters.

The primary challenge with the dTag technique is establishing ideal fixation and nucleosome disruption conditions to maximize nuclei recovery after the second tagmentation reaction. We noticed improved nuclei yields with a higher formaldehyde percentage without a reduction in data quality; however, the majority (~90%) of nuclei were lost at this step, suggesting additional optimization will be required for samples that do not have at least ~1 million nuclei to start with – something that is not a problem with most tumor tissue specimens, but can be tricky when handling biopsy-derived tissue that is often portioned for multiple assays. Furthermore, the low cell yield makes it challenging to achieve sufficient genome coverage for *de novo* variant calling on identified clusters, restricting utility to samples where paired bulk exome or whole genome data can be used to produce a reference variant set against which dTag cell clusters can be genotyped.

Table 1: Chapter 1 Materials		
Reagent / Resource	**Source**	**Identifier**
Chemicals, Peptides, and Recombinant Proteins		
Hepes, pH 7.5	Sigma-Aldrich	H4034
$MgCl_2$	Sigma-Aldrich	M8226
NaCl	Fisher Scientific	M-11624
IGEPAL	Sigma-Aldrich	I8896
Tween-20	Sigma-Aldrich	P7949
Formaldehyde	Fisher Scientific	PI28906
Glycine	Sigma-Aldrich	G8898-500G
UltraPure Sodium Dodecyl Sulfate	Invitrogen	15525-017
D-(+)-Glucosamine hydrochloride	Sigma-Aldrich	G1414-100G
Critical Commercial Assays		
Chromium Next GEM Single Cell ATAC Kit v2	10x Genomics	PN-1000390
Next GEM Chip H Single Cell Kit	10x Genomics	PN-1000162
Single Index Kit N, Set A, 96	10x Genomics	PN-1000212
Single Cell ATAC Gel Beads v2	10x Genomics	PN-2000210
scATAC Pre-Indexing Kit	ScaleBio	N/A
Qubit 1x dsDNA HS Assay Kit	Invitrogen	Q33231
High Sensitivity D1000 ScreenTape	Agilent	5067-5584
High Sensitivity D1000 Sample Buffer	Agilent	5067-5603
NextSeq2000 Kits	Illumina	20046811, 20046812
Deposited Data		
Raw and analyzed data	This paper	
Human Reference Genome NCBI Build 37, GRCCh37	Genome Reference Consortium	http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
10x Genomics scCNV kit released datasets	10x Genomics[14]	10xgenomics.com/resources/datasets
Experimental Models: Cell Lines		
K562	Tanner Lab	N/A
GM12878	Coriell Institute	Cat. No. GM12878

Patient-Derived Pancreatic Cancer Cell Line	Rosie Sears Lab (OHSU)	ST-00024058
Software		
unidex	Adey Lab	https://github.com/adeylab/unidex
bwa-mem (v0.7.15-r1140)	Li H, Durbin R[82]	https://github.com/lh3/bwa
scitools	Adey Lab	https://github.com/adeylab/scitools
RIDDLER	Gurkan Lab	https://github.com/yardimcilab/RIDDLER
ArchR	Granja JM, *Corces MR* et al.[79]	https://www.archrproject.com/
Other		
Cytiva HyClone RPMI 1640 Media	Fisher Scientific	SH3002701
Gibco DMEM Media, high glucose	ThermoFisher Scientific	11960044
Glutamax 100x	Life Technologies	35050061
HyClone Bovine Calf Serum, heat inactivated	VWR	SH30073.04HI
Penicillin/Streptomycin (10,000 U/ml)	Life Technologies	15-140-122
TrypLE Express Enzyme	ThermoFisher Scientific	12604013

Resource availability

Lead contact: Requests for protocols and reagents can be directed to Andrew Adey (adey@ohsu.edu).

Materials availability

Materials used in this study are commercially available.

Data and Code Availability

- Data are accessible through the NCBI Gene Expression Omnibus (GEO) under accession GSE243430.
- This study does not report original code
- Any additional information needed to reanalyze the data reported in this paper is available from the Lead Contact by request.

Experimental model and subject details

Preparation of cell lines

We cultured GM12878 and K562 in Roswell Park Memorial Institute Medium, supplemented with 1% additional glutamine, 1% penicillin/streptavidin, and 10% calf bovine serum. Standard humidified incubator conditions were kept at 37°C with 5% CO_2. We cultured the adherent Neuro2A cells in Dulbecco's Modified Eagle's Medium and dissociated them with TrypLE Express Enzyme, as required for cell passage. Otherwise, the culture conditions remained identical.

Preparation of PDCL

The PDCL (ST-00024058) was established from a dissociated human pancreatic ductal adenocarcinoma tumor metastasis and cultured for continuous propagation in culture medium containing ROCK inhibitor (Y-276320)[83]. Briefly, 100,000 viable, disaggregated tumor cells were plated to a 13mm diameter, collagen-coated well (Gibco, A11428–02) and passaged while subconfluent until reaching 85% confluence on a 10cm diameter dish, which was designated as passage 1. DNA was extracted from a third passage to validate by whole exome sequencing that the mutation profile in the PDCL matched the patient tumor DNA profile, including $Kras^{G12D}$ and $TP53^{R175H}$ mutations (detected at 45% and 100% frequency, respectively, in the PDCL). The PDCL also exhibited morphologies consistent with epithelial tumor cells and abundant KRT expression was detected by immunocytofluorescence using the monoclonal antibodies AE1/AE3, C-11, and Cam5.2. The tumor originated from a deidentified 60-year-old male, under the oversight and with the approval of the OHSU Institutional Review Board.

METHOD DETAILS

Nuclei Isolation

To isolate the nuclei, we collected a 2 mL suspension of cells in media. We spun the suspension down (5 minutes, 500xg, 4C) to remove the media supernatant and resuspended it in 1 mL of NIB-Hepes buffer (10 mM Hepes, pH 7.5, 3 mM $MgCl_2$, 10 mM NaCl, 0.1% IGEPAL (v/v), 0.1% Tween-20 (v/v)). We then incubated the cell suspension for 5 minutes on ice; and

subsequently, we spun down the sample (5 minutes, 500xg, 4C). We then resuspended the pellet in 1 mL NIB-Hepes, spun again, and resuspended in 1 mL once more before quantification.

Nucleosome Disruption

We diluted the samples to 1 million nuclei per mL in NIB-H buffer and then added 46.9 µL 16% formaldehyde (final concentration is ~0.75% formaldehyde). After pipette mixing, we fixed the sample at room temperature over 10 minutes, on an orbital shaker set to 50rpm. We next added 46.9 µL 2.5M glycine to quench the reaction, incubated for 5 minutes on ice, and then spun the suspension down (5 minutes, 500xg, 4°C). We next resuspended the sample in 970 µL NIB-H and then added 30 µL 10% SDS. We incubated the nuclei for 20 minutes at 37°C in this solution. We carefully spun down the samples (5 minutes, 500xg, 10°C), as the SDS can precipitate and taint the pellet if left cold for too long. Nuclei were then resuspended in 1 mL NIB-H buffer and quantified.

10x Tagmentation and Barcoding for scWGS

For the single cell whole genome sequencing protocol leveraging the 10x transposition chemistry, we took the preprocessed, nucleosome disrupted nuclei and diluted them in 10x Genomics 20x Nuclei Buffer to the desirable targeted cell recovery. We then proceeded with the 10x Genomics Single Cell ATAC v2 Kit according to their published protocol.

ScaleBio Tagmentation and 10x Barcoding for scWGS

For the scWGS protocol leveraging the ScaleBio transposition chemistry, we resuspended the preprocessed nuclei in PBS Buffer with BSA, such that there were 10,000 or 20,000 nuclei per µL. We also added 30 mM D-glucosamine to the PBS/BSA buffer, which will amount to final concentration of 10 mM during the tagmentation reaction, to improve the recovery of nuclei after the tagmentation[84]. In a 96-well plate, we added 5 µL of diluted nuclei to 5 µL of ETB3, prior to adding 5 µL of the indexed Scale TSM. We mixed the plate by shaking at 1400 rpm for a minute, briefly spun it down at 500xg, and then we incubated the plate at 37°C for one hour on a thermocycler. The lid should be set to 47°C. In the interim, we thawed the ScaleBio wash and loading buffers on ice.

After the tagmentation was completed, we removed the plate from the thermocycler, briefly spun the plate down at 500xg, and incubated it on ice for 5 minutes. We pooled the samples in an Eppendorf tube and filled the remaining volume with the ScaleBio wash buffer. For subsequent steps, using a spinning bucket centrifuge aids in retention of the nuclei pellet. We spun down the samples (3 minutes, 500xg, 4°C), discarded the supernatant, added another 1.5 mL of wash buffer, and spun again. We again removed all the supernatant and resuspended the pellet in 25 µL of ScaleBio's loading buffer. We quantified the nuclei and added the desired amount to be loaded in 15 µL loading buffer on the 10x Chromium. At this point multiple samples with unique Tn5 indexes can be pooled in the 15 µL of loading buffer, to be run on a single lane on the Chromium Chip for multiplexing.

With the quantified, transposed nuclei in hand, we continued to Step 2 (Gem Generation and Barcoding) of the 10x Genomics Single Cell ATAC v2 Kit protocol. We mixed our 15 µL of nuclei with 60 µL of the described 10x Master Mix and proceeded from here on out exactly as described by the 10x protocol. The one subsequent deviation is during Step 4.1.c (Sample Index PCR), where instead of using the provided Sample Index N, Set A Reagent – we used a ScaleBio S700 index primer compatible with the ScaleBio tagmentation.

ScaleBio Tagmentation and 10x Barcoding for DoubleTag

For the double tagmentation protocol leveraging the ScaleBio transposition chemistry, we proceeded through the ScaleBio Tagmentation approach described above on isolated, but not nucleosome disrupted, nuclei. After tagmentation, we washed twice with the ScaleBio wash buffer. We next resuspended the nuclei in NIB-H buffer and perform the nucleosome disruption protocol. Following nucleosome disruption, we repeated the ScaleBio tagmentation – being sure to use differently indexed Tn5 so that the ATAC and WGS fragments can be distinguished. After the second tagmentation, and subsequent the washing and resuspension in the loading buffer, we quantified the nuclei and added the desired amount to be loaded in 15 µL loading buffer to the 10x Chromium. As before, we continued from Step 2 (Gem Generation and Barcoding) of the 10x

Genomics Single Cell ATAC v2 Kit protocol, after mixing the 15 µL with 60 µL of the 10x Master Mix and loading the chip. Again, we would go on to use the ScaleBio S700 index primers in lieu of the 10x Sample Index N during Step 4.1.c (Sample Index PCR).

Library Quantification & Sequencing

We cleaned library preparations as described in the 10x Genomics Single Cell ATAC v2 Kit. We quantified generated libraries via the Qubit dsDNA High Sensitivity assay (Thermo Fisher Q32851). We then confirmed the molarity of the DNA via the Agilent Tapestation 4150 D5000 tape (Agilent 5067-5592), with preparations diluted to 2ng/µL based on the Qubit data if necessary. We sequenced library preparations using standard chemistry on the Illumina NextSeq2000 for 650pm using a P2-100 or P2-200 flow cell (Illumina Inc. 20046811, 20046812). ScaleBio tagmented libraries were sequenced as paired-end with 85 cycles for read 1, 125 cycles for read 2, 8 cycles for index 1, and 16 cycles for index 2. Standard 10x tagmentation libraries were sequenced as paired-end with 50 cycles for read 1 and read 2, but 8 cycles for index 1, and 16 index 2. Several libraries, for which we required great sequencing depth, were also sequenced on a NovaSeq S2 flowcell following manufacturer's instructions (Illumina Inc. 20028315).

QUANTIFICATION AND STATISTICAL ANALYSIS

Computational Analysis

Primary Sequence Data Processing

Sequence reads were demultiplexed using unidex (https://github.com/adeylab/unidex) which matches barcode regions to a whitelist, allowing for a hamming distance of 1 from the 10x bead barcode and index read 2, and then a hamming distance of 2 for index read 1 (sample index) or the first 8bp of read 2 which serves as the tagmentation index when indexed tagmentation is performed. For indexed tagmentation experiments the 8 bp index was trimmed, along with the next 20 bp of mosaic end sequence. Read names were then replaced with the error-corrected barcode combination. Demultiplexed and barcode-matched reads were then aligned to the human reference genome hg38 using bwa-mem (v0.7.15-r1140), then PCR-duplicate removed in a barcode-aware manner using scitools rmdup (https://github.com/adeylab/scitools) and then filtered to only contain

reads in cell barcodes reaching a minimum unique read count as determined by the inflection point on a knee plot.

MAD Scores

Median Absolute Deviation (MAD) scores were calculated as previously described[73,76]:

$$MAD\ score_i = median(|d - median(d)|);\ where: d = \frac{\frac{Y_{i,j}}{N_i B_j} - \frac{Y_{i,j+1}}{N_i B_{j+1}}}{\left(\sum_{j=i}^{n} \frac{Y_{i,j}}{N_i B_j}\right) \Big/ n}$$

Where $Y_{i,j}$ is the raw read count for the i^{th} cell of the j^{th} bin. N_i is the cell-specific scaling factor (total reads), and B_j is a bin-specific scaling factor (total reads in bin across all cells).

Coverage Accumulation

Coverage accumulation was carried out by randomly sampling cells without replacement and aggregating the coverage assuming paired 150 bp sequence reads, and only considering the physical coverage of the read if the insert size was less than 300 bp (i.e. when read pairs overlap, it is collapsed to a coverage of 1x for the overlapping bases). 150 bp was chosen as it represents what a typical sequencing instrument can produce if genomic coverage is the goal (as opposed to shorter reads if only copy number calling is desired).

CNV detection using RIDDLER

CNV detection of cells was performed using the RIDDLER framework. Reads were binned at 1 Mbp resolution, creating a matrix of cells by windows used as input. Zero values in the matrix were assessed as potential dropout events, treating each cell as a sample of reads over all windows. The probability of these windows having zero reads in the absence of CNVs was estimated, based on the total reads in the cell and average window distribution across all cells. Windows with a high estimated probability (0.95 or higher) of being dropout events were removed from the input data. This typically only occurred in low coverage windows of lower coverage cells.

Robust Poisson regression, detailed in Cantoni & Ronchetti (2001)[85] and implemented in the R package robustbase, was used to model the expected distribution of reads per window across cells. The use of robust regression as opposed to least squares regression reduces the influence of

outliers, making them easier to disentangle from the data. As covariates for the regression model, we used mappability, AB compartment, and median coverage. Mappability was scored using 50mer alignability[85], AB compartment scores were taken as the ENCODE GM12878 Hi-C 5 kbp genome compartments[86], and median coverage was computed using the reference cell line GM12878. Each covariate was averaged within each window to match the input dimensions, and the same covariates were used for each cell and each experiment. Regression models were computed for each chromosome individually, and used to predict an expected read count for each window.

CNVs were detected by comparing the observed window reads in each cell to the expected reads predicted by the model. Using the model prediction as the expected value, p-values were computed for each window using a negative binomial distribution. Dispersion values for the negative binomial distribution were estimated for each chromosome using the GM12878 cell line, to establish a baseline of expected variation. Window P-values were combined with neighboring windows up to 3 Mbps away to create aggregate P-values using an empirical Bates distribution. These aggregate P-values were then FDR corrected (Benjamini Hochberg 1995)[87], and thresholded for significance at a value of 0.025 for each tail. Windows that passed this threshold on the upper tail are labeled as CNV gains, and those that pass on the lower tail are labeled as CNV losses.

To assign copy numbers (0, 1, 2, 3, 4, or 5+) to each window, we computed the log-likelihood of the observed reads using a corresponding multiplier of the model fitted reads (0, 0.5, 1, 1.5, 2, and 2.5 respectively) as the expected value. Since having an expected value of 0 produces errors in the negative binomial distribution, we instead used an expected value multiplier of 0.1 to assess the likelihood of the copy number 0 label. The copy number with the highest log-likelihood was then assigned to the window.

Analysis of scATAC Data

Aligned and duplicate-removed bam files were converted into fragment files using sinto 0.9.0 and then loaded into ArchR[16] to generate arrow files for each ATAC fragment file which were then compiled into an ArchR project. Iterative LSI, harmony integration, and UMAP projections

were performed using default parameters. Track plots were generated by selecting known housekeeping genes that are expected to be active in all cell lines.

Acknowledgements

This work was supported by an NCI IMAT R33 (R33CA269015) to A.C.A., an NIH Ruth L Kirschstein T32 Fellowship (5T32GM142619-02) to K.Q.

Declaration of Interests

A.C.A. is an author on a patent that covers one or more aspects of the nucleosome disruption technology utilized here. This potential conflict is managed by the OHSU office of research integrity.

Chapter 2: Cell fusion reprograms cancer cells and promotes RUNX1 mediated tumor-macrophage hybrid cell invasion in colorectal cancer

This section is based on an article to be published as: Anderson A, Queitsch K, Giske N, Jones J, Pang A, Zucker A, Huang G, Rounds C, Smith B, Swain J, Moore, A, Greer W, Tao K, Wu G, Bertassoni L, Franca C, Fischer J, Adey AC, Gibbs SL, Wong MH. Cell fusion reprograms cancer cells and promotes RUNX1 mediated tumor-macrophage hybrid cell invasion in colorectal cancer.

Authors collaborating in this work and affiliations

[1]Department of Cell, Developmental and Cancer Biology, Oregon Health and Science University (OHSU); Portland, OR, USA

[2]Department of Molecular and Medical Genetics, OHSU; Portland, OR, USA

[3]Department of Biomedical Engineering, OHSU; Portland, OR, USA

[4]Center Early Detection Advanced Research Center, Knight Cancer Institute, OHSU; Portland, OR, USA

[5]Department of Medical Informatics and Clinical Epidemiology, OHSU; Portland, OR, USA

[6]Department of Oral Rehabilitation & Integrative Biosciences, OHSU; Portland, OR, USA

[7]Knight Cancer Institute, OHSU, Portland, OR, USA

[#]Corresponding Author: Melissa H. Wong, wongme@ohsu.edu

Author contributions

Conceptualization: A.N.A., A.A., S.L.G., M.H.W., Methodology: A.N.A., K.Q., G.W., A.A., S.L.G., M.H.W., Formal analysis: A.N.A., K.Q., J.J., C.R., G.W., C.M.F., J.F., A.A, S.L.G, M.H.W, Data curation: A.N.A., K.Q., N.G., A.Z, J.J., A.P., G.H, A.I., B.J.S., J.S., H.F., W.G., C.M.F., J.F. Writing – original draft preparation: A.N.A., K.Q., S.L.G., M.H.W. Writing – reviewing and editing: A.N.A., K.Q., A.A., S.L.G., M.H.W. Resources: K.T., G.W., L.B., C.M.F., J.F., A.A., S.L.G., M.H.W. Supervision and project administration: A.A., S.L.G., M.H.W. Funding acquisition: A.N.A., K.Q., A.A., S.L.G., M.H.W.

I spearheaded the single-cell sequencing efforts including the experimental design for the RNA-seq, ATAC-seq, and scWGS experiments. Working closely with A.A., I executed all single-cell sequencing experiments, performed library preparation and sequencing, and conducted the subsequent data processing and analysis. I developed and optimized the protocols for hybrid fusion cells on the Takara ICell8. I implemented the computational pipelines for ICell8, SmartSeq, and 10x libraries, and performed the key analyses identifying the macrophage-tumor hybrid populations and their transcriptomic and accessibility signatures. I also contributed to data interpretation, figure generation (specifically Figures 13, 14, 17, 18, and 19), and manuscript preparation. For the purpose of this thesis, I drafted a Supplemental Contribution: scWGS in Hybrid Cells section describing the application of my novel scWGS method to hybrid cells; this work represents thesis-specific content that will not be included in the submitted manuscript.

Abstract

Cancer metastasis is a leading cause of cancer-related morbidity and mortality. One proposed mechanism involves the fusion of neoplastic cells with immune cells (e.g., macrophages), forming tumor-immune hybrid cells that acquire the functional ability to migrate and disseminate into peripheral blood. This study investigates hybrid cell heterogeneity and molecular mechanisms underlying dissemination in colorectal cancer (CRC). Using single-cell RNA sequencing, cyclic immunofluorescence and an *in vitro* model of CRC hybrid cells, we identified Runt-related transcription factor 1 (*Runx1)* significantly expressed in hybrid cells with migratory and invasive phenotypes. *Runx1* depletion downregulated protease expression and reduced chemotaxis and matrix invasion. $RUNX1^{+}$ hybrid cells are identified in primary tumor and peripheral blood of patients with CRC, with circulating levels correlating with disease stage. These findings indicate that RUNX1 is a key regulator of hybrid cell invasiveness, contributing to CRC metastases.

Main Text

Metastatic progression is a primary cause of cancer-related mortality, as disseminated tumor cells establish secondary lesions at distant sites that are resistant to first line therapies, impacting patient outcomes[88]. Over 20-25% of colorectal cancer (CRC) patients are diagnosed with metastatic disease at the time of detection and an additional 20-30% of patients with early-stage cancer harbor undetectable micro-metastatic disease that leads to disease recurrence[89,90]. Metastatic disease develops when molecular alterations and microenvironmental cues reprogram neoplastic cell behavior, enabling their escape from the primary tumor and migration to distant organs—a process known as the metastatic cascade[91,92]. During this progression, distinct tumor cell clones emerge, generating tumor heterogeneity, which complicates effective treatment response. Despite advances in understanding these processes, the molecular mechanisms underlying the acquisition of metastatic traits and therapeutic resistance remain incompletely defined.

One proposed mechanism by which neoplastic cells acquire metastatic potential is through fusion with immune cells, forming neoplastic-immune hybrid cells[93–107]. These hybrids retain both genotypic and phenotypic traits of their parent cells, combining the immune cell's migratory and immune-evasive properties with the tumor cell's proliferative and tumor-initiating capabilities[108,109]. In addition, hybrid cells express expanded cellular identities, including cancer stem cell features, indicating that fusion contributes to cellular reprogramming[95,96,99,110–130].[8,9,12,23-44] As a result, neoplastic-immune hybrid cells gain enhanced capacity to traverse the metastatic cascade. Hybrid cells are detectible in murine cancer models and human cancer patients in primary tumors, peripheral blood, and metastatic lesions across over fourteen different solid tumor types[131,132]. Notably, circulating hybrid cells (CHCs) are more prevalent in peripheral blood than circulating tumor cells (CTCs), which lack immune protein expression[131,133]. CHC levels correlate with disease burden and overall survival in cancers such as pancreatic, gastrointestinal, uveal melanoma, and head and neck malignancies, underscoring their potential as a non-invasive

biomarker[131–136]. Despite this, the genomic landscape, heterogeneity, and mechanisms of dissemination of hybrid cells remain poorly understood. As the predominant neoplastic cell type in circulation, uncovering the molecular drivers of hybrid cell behavior is critical to advancing cancer early detection strategies and developing therapeutic approaches to prevent or limit metastatic disease.

In this study we aim to define the transcriptomic, epigenetic and functional heterogeneity of hybrid cells in colorectal cancer (CRC) and to investigate the molecular mechanisms underlying their dissemination. We utilize single cell RNA sequencing (scRNA-seq) analyses of both previously generated datasets, and newly established datasets from hybrid cells sorted from patient matched primary tumors and peripheral blood to assess the phenotypes of disseminating hybrid cells. To complement transcriptomic analysis, we use highly multiplexed, whole slide immunofluorescence to evaluate hybrid cell phenotypes at the protein level in matched tumor and blood specimens. Our findings reveal that CRC hybrid cell phenotypes shift throughout the metastatic cascade: tumor-localized hybrids are enriched for migratory pathways, whereas CHCs display expression of immune-related and immune-evasive programs. Using an *in vitro* model of CRC cell fusion we demonstrate that hybrid cells gain transcriptional plasticity, a hallmark of metastasis-initiating cells[137]. In addition, these hybrids exhibit distinct gene expression profiles compared to both parental cells and neoplastic-immune cell doublets confirming their unique identity and altered cell phenotypes. Single cell derived hybrid clones exhibit properties ranging from “more tumor-like” to “more immune-like” evaluated at the gene, protein and phenotypic level. Notably, we identify the transcription factor, RUNX1 as significantly more accessible and overexpressed at the chromatin, gene and protein level in hybrid clones with enhanced migratory and invasive behavior. *Runx1* depletion downregulates protease expression and reduces migratory and invasive behavior. $RUNX1^{+}$ hybrid cells are detectable in human CRC primary tumors and in circulation, with increased prevalence in advanced stage CRC. Together, these results identify

RUNX1 as a key regulator of hybrid cell dissemination in CRC and highlight its potential as a therapeutic target to limit metastatic progression.

Results

Neoplastic-macrophage hybrid cell heterogeneity in CRC primary tumors and lymph node metastases

To gain an understanding of the diversity of neoplastic-macrophage hybrid cells within CRC primary tumors and lymph node metastases, we set out to identify hybrid cells within publicly available scRNA-seq datasets of CRC. This approach builds on our previous study that identified hybrid cells in a scRNA-seq dataset of uveal melanoma tumors[138]. We identified neoplastic-macrophage hybrid cells in a recently published dataset comprising 189 samples from 63 patients with CRC[139] (Figure 11A) based on the co-expression of monocyte/macrophage markers (*CD68, CD14*, *CD163* and/or *CD11b)* alongside one or more epithelial cell markers (E-cadherin (*ECAD)*, *EpCAM*, or pan-cytokeratin (defined as keratins 2-5, 7-8, 14-16, and 19); Figure 11B).

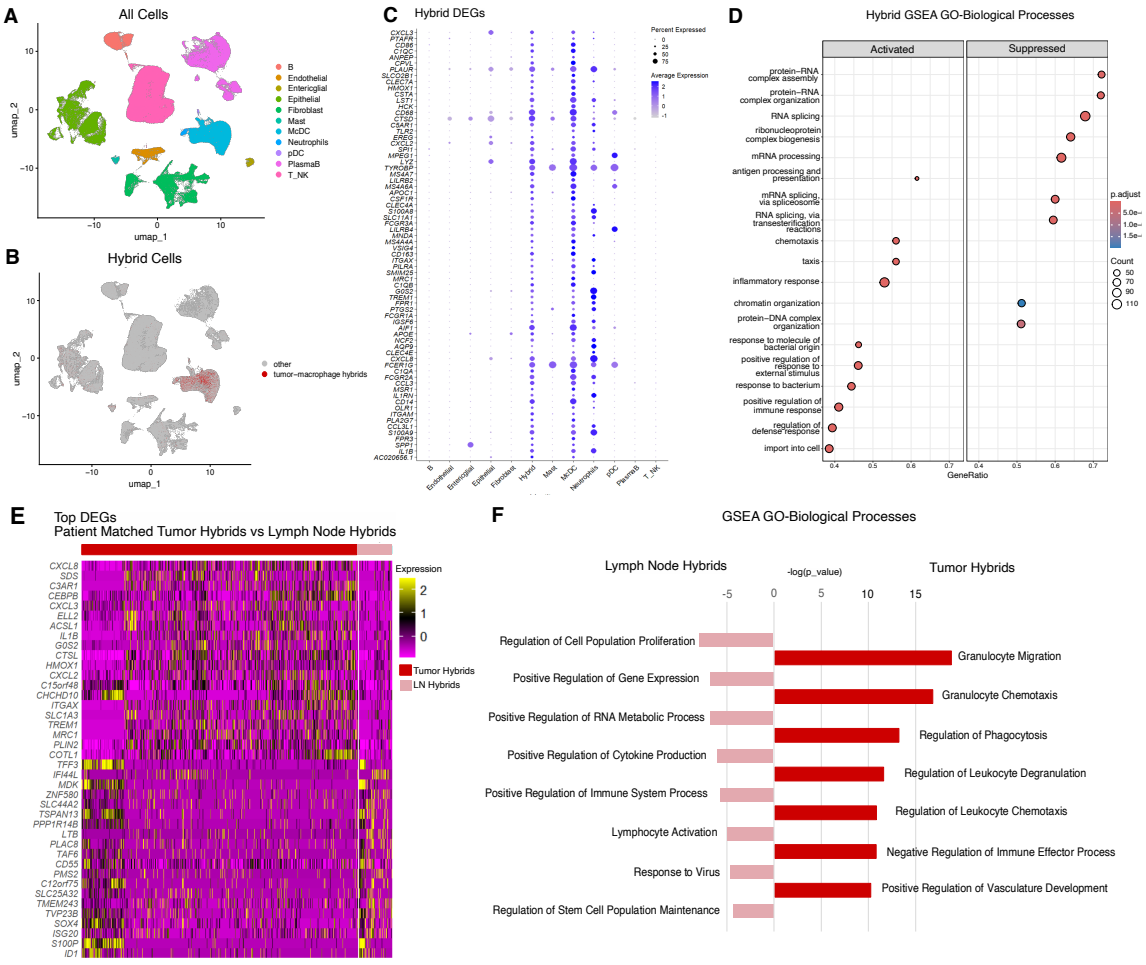


Figure 11: Tumor–macrophage hybrid cells upregulate immune evasive, migratory, and stem-like gene programs during colorectal cancer metastatic progression.
A) UMAP depicting annotated cell type clusters of 373,058 cells from scRNA-sequencing (n=63 patients).[54] B) Hybrid cells (red) identified by co-expression of macrophage markers (CD45, CD163, CD14) and one more epithelial markers (EPCAM, ECAD, panCK (KRT2–5, KRT7–8, KRT14–16, KRT19) overlaid onto the umap. C) Differential gene expression analysis of all identified hybrid cells vs all other sequenced cells. D) Gene set enrichment analysis (GSEA) Gene Ontology (GO)- biological processes of hybrid cells using differential gene expression output from panel C. E) Differential gene expression analysis of tumor hybrid cells vs lymph node hybrid cells identified in patient matched tumor and lymph node specimens (n=7 patients). F) GSEA GO-biological processes of tumor hybrid cells vs lymph node hybrid cells using differentially expressed gene output from panel E.

After rigorous doublet discrimination and removal, we identified 5,592 hybrid cells, representing approximately 2% all sequenced cells, across all the tumor and lymph node samples (Figure 11B and Supplemental Figure 11). Differential gene expression (DEG) analyses revealed that hybrid cells expressed elevated levels of genes associated with migration and immune cell

function, including *SPI1*, *PLAUR*, *GRN*, *AIF1* and *TYROBP* (Figure 11C). Notably, *AIF1* and *TYROBP* were also upregulated in hybrid cells from uveal melanoma,[138] suggesting a conserved hybrid cell gene expression program across cancer types.

Overall, the transcriptional profiles of CRC hybrid cells more closely resembled that of macrophages rather than epithelial cells, as shown by both differential expression and UMAP visualization (Figure 11B,C). Gene set enrichment analysis using the Gene-Ontology biological pathways annotated dataset revealed significant enrichment for immune-related and chemotaxis-associated pathways (Figure 11E), consistent with previous findings in hybrid cells from uveal melanoma and prostate cancer[138,140,141].

To further evaluate hybrid cell identity, we directly compared hybrid cells to epithelial and monocyte/classical dendritic cell (McDC) populations. Compared to epithelial cells, hybrid cells showed increased expression of immune-related genes including *CD163, CD45, FPR1, CD64, FCGR3A, PLEK*, and complement genes *C1QA-C*. In contrast, compared to the McDC population, hybrid cells expressed higher levels of epithelial genes (*KRT8, KRT10, KRT19, KRT18, EpCAM*), as well as *TFF3, LGALS4*, and *ARG2,* supporting their dual tumor-macrophage identity beyond the markers used for hybrid cell classification (full DEG lists in Supplemental Files 5.2-5.3).

A unique feature of this scRNA-seq dataset was the inclusion of seven matched tumor and lymph node samples, which enabled us to evaluate hybrid cell gene expression across the metastatic cascade by comparing hybrids in the primary tumor to those disseminated to the lymph nodes. Within this subset, we identified 784 hybrid cells (0.85% total sequenced cells), including 700 hybrids from primary tumors and 84 hybrids from lymph node specimens.

Differential gene expression analysis revealed that tumor-localized hybrids upregulated *CXCLs 2,3,8* and *IL1B*, chemokines known to recruit neutrophils and other immune cells, and previously associated with pro-metastatic signaling (Figure 11E)[142,143]. In contrast, lymph node-

localized hybrids showed increased expression of LTB, a gene commonly enriched in lymphoid tissues[144].

Notably, we observed a subset of tumor-derived hybrid cells with gene expression profiles similar to top DEGs in lymph node-localized hybrids (Figure 11E, leftmost cluster). These cells expressed SOX4, a stem cell-associated transcription factor, and CD55, an immune-evasive marker enriched in cancer stem cells[145], suggesting that stem-like hybrid cells may possess enhanced metastatic potential.

Gene set enrichment analysis further supported functional divergence: tumor-resident hybrids were enriched for pathways involved in vasculature development, chemotaxis and cell migration, while lymph node-localized hybrids were enriched for immune signaling and stem cell-related pathways (Figure 11F). Together, these results indicate that hybrid cells within the primary tumor may be poised to metastasize by promoting vasculature remodeling and modulating the immune microenvironmental, whereas hybrid cells that successfully disseminate to and persist within lymph nodes adopt phenotypes associated with cell proliferation, immune adaptation and stemness.

Tumor-macrophage hybrid cell clones exhibit phenotypic heterogeneity

To investigate how cell fusion alters neoplastic cell phenotypes and drives increased dissemination of hybrid cells, we leveraged our previously established *in vitro* model of spontaneous fusion between the MC38-H2B-RFP (tumor) CRC cell line, and primary bone marrow-derived macrophages (macrophage) from β-actin-GFP C57BL/6 mice[108,131]. After four days of co-culture, double-positive GFP^+/RFP^+ hybrid cells were FACS-isolated and either subjected to single cell RNA sequencing (scRNA-seq), or expansion in culture with repeated purity FACS-isolation.

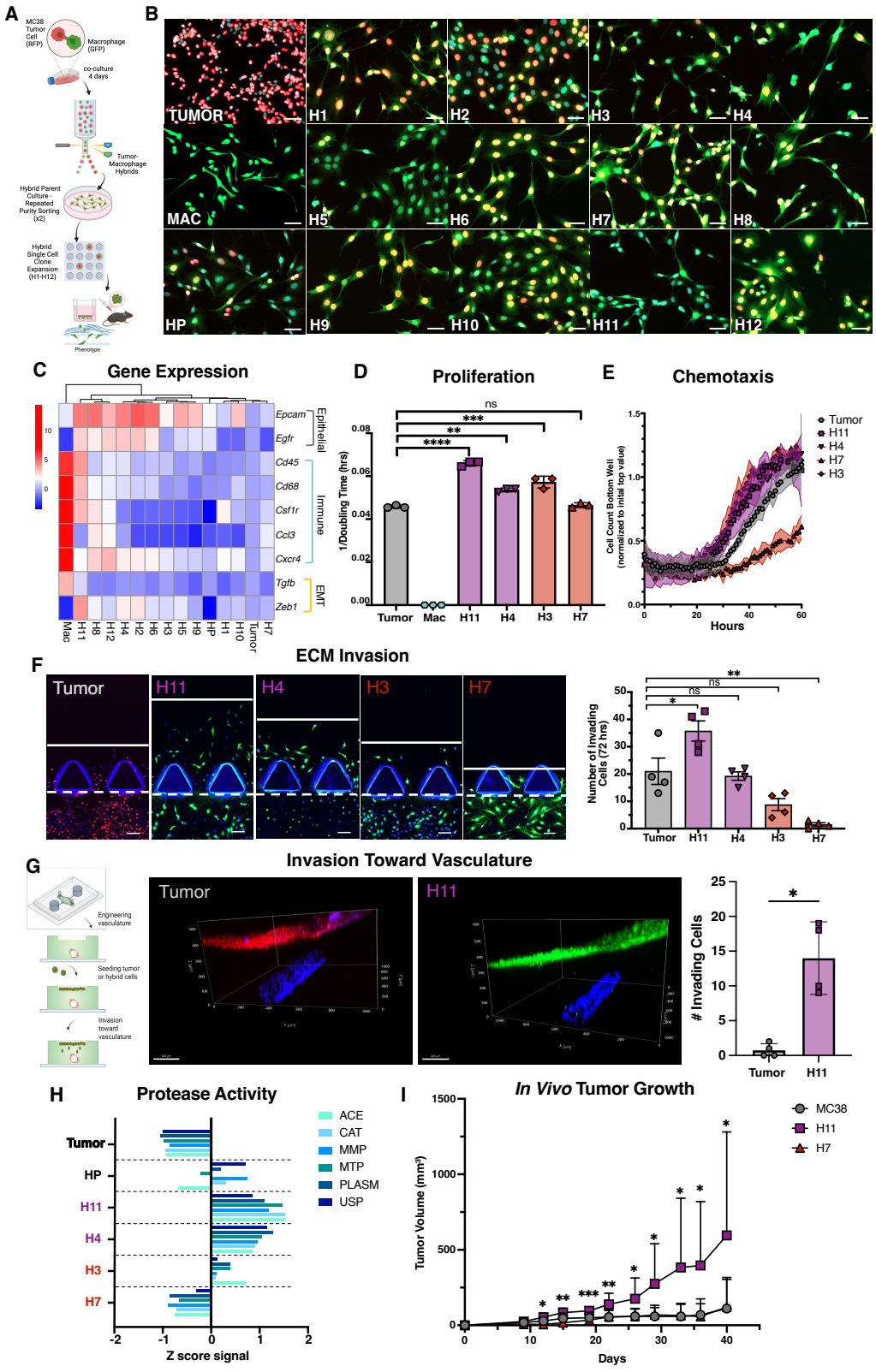

A
MC38 Tumor Cell (RFP)
Macrophage (GFP)
co-culture 4 days
Tumor-Macrophage Hybrids
Hybrid Parent Culture - Repeated Purity Sorting (x2)
Hybrid Single Cell Clone Expansion (H1-H12)
Phenotype
B
TUMOR
H1
H2
H3
H4
MAC
H5
H6
H7
H8
HP
H9
H10
H11
H12
C
Gene Expression
Epcam
Egfr
Cd45
Cd68
Csf1r
Ccl3
Cxcr4
Tgfb
Zeb1
Epithelial
Immune
EMT
Mac
H11
H8
H12
H4
H2
H6
H3
H5
H9
HP
H1
H10
Tumor
H7
D
Proliferation
1/Doubling Time (hrs)
ns
***
**
****
Tumor
Mac
H11
H4
H3
H7
E
Chemotaxis
Cell Count Bottom Well (normalized to initial top value)
Hours
Tumor
H11
H4
H7
H3
F
ECM Invasion
Tumor
H11
H4
H3
H7
Number of Invading Cells (72 hrs)
G
Invasion Toward Vasculature
Engineering vasculature
Seeding tumor or hybrid cells
Invasion toward vasculature
Tumor
H11
Invading Cells
H
Protease Activity
Tumor
HP
H11
H4
H3
H7
ACE
CAT
MMP
MTP
PLASM
USP
Z score signal
I
In Vivo Tumor Growth
MC38
H11
H7
Tumor Volume (mm³)
Days

Figure 12: Colorectal cancer hybrid cell clones exhibit phenotypic and functional heterogeneity.
A) Schematic of experimental design. B) Images of MC38xY01 hybrid cells, parental MC38 tumor cells, parental Y01-Actin-GFP bone marrow derived macrophages, and single cell clones generated from MC38xY01 hybrid cell population. (scale bar = 50uM) C) Hierarchical clustering of macrophage, epithelial, and EMT gene expression from hybrid single cell clones and parental tumor and macrophages (n=3 biological replicates, each with 3 technical replicates, averaged) D) Proliferation of hybrid single cell clones (data for all hybrid single cell clones in supplemental figure 2). E) Trans-well chemotaxis of hybrid single cell clones and tumor parent F) Hybrid cell invasion into collagen extracellular matrix (ECM) (imaging shown represents one section of whole well imaging, scale bar = 100uM) and quantification of hybrid cell invasion in ECM represented as total number of invading cells after 72hrs G) Invasion of MC38 tumor cells and H11 cells through collagen matrix toward vasculature (imaging shown represents one experimental replicate, scale bar = 200uM) and quantification of the number of invading cells I) Protease activity in hybrid single cell clones and tumor parent J,K) H11, H7 and MC38 tumor volume of subcutaneous tumors in female and male mice. (p-values: **** <0.0001, *** <0.001, **<0.01, *<0.05)

To assess hybrid cell heterogeneity, we used FACS to isolate single hybrid cells and place them into individual wells of a 96-well plate (Figure 12A). Twelve hybrid clones grown from single cells were selected for further evaluation based on their differences in morphology and expression of RFP and GFP (Figure 12B). Gene expression analysis using qRT-PCR revealed diverse expression of selected epithelial, immune, and epithelial-to-mesenchymal (EMT)-associated genes across hybrid cell clones (Figure 12C). While some clones expressed macrophage-associated genes (*Cd45, Cd68, Csf1r, Ccl3*), others did not. Notably, most hybrid clones harbored elevated expression of epithelial genes, *EpCAM* and *Ecad*, despite their absence in the parental tumor cell line.

Hierarchical clustering of the gene expression data identified distinct subgroups of clones: macrophage-like hybrids (e.g., Hybrid (H)4, H8, H11 and H12) and tumor-like hybrids (e.g., H1, H7, H9, H10). Functional assays revealed marked heterogeneity in proliferation and chemotactic responses among clones (Figure 12D and E). In ECM invasion assays, macrophage-like clones (H11, H4) possessed significantly greater matrix degradation and invasion (both in cell number and depth) compared to tumor-like clones (H3, H7) and the parent tumor MC38 cell line (Figure 12F

and G). Furthermore, protease expression directly correlated with ECM invasion across all hybrid clones (Figure 12H).

In co-culture with vasculature, H11 also showed enhanced invasion and directional migration toward vasculature compared to the MC38 tumor parent cell line (Figure 12I). *In vivo*, H11 exhibited significantly greater tumorigenic potential in C57Bl/6 mice, forming more tumors and growing at a faster rate than both the MC38 tumor parent and H7 cells (Figure 12J-K). Collectively, these findings demonstrate that cell fusion induces substantial phenotypic and functional heterogeneity, giving rise to hybrid subpopulations with distinct gene expression profiles, invasive capacities, and tumorigenic potential.

Tumor-macrophage hybrid cell lines display transcriptomic heterogeneity

To gain a comprehensive understanding of hybrid cell heterogeneity, we examined their transcriptomic and epigenetic landscapes at the single cell level using scRNA and single cell assay for transposase accessible chromatin (scATAC) sequencing. We generated scRNA libraries from three groups: 1) "freshly fused" hybrid cells FACS-isolated after 4 days of co-culture, 2) hybrid cells maintained in culture through repeated purity FACS-isolation and passaging in culture, and 3) the parental MC38 tumor and macrophage populations (Figure 13A).

A critical initial step in our studies was to distinguish hybrid cells from macrophage-tumor doublets, which can arise as artifacts of droplet-based library generation platforms. To establish the difference between hybrid cells and artificially generated doublets, we employed the Takara's ICELL8 platform, which allows for both cell dispensing and nanowell imaging, to generate artificial doublets (tumor + macrophage). These doublets were confirmed by manual annotation of imaged wells, then subjected to Smart-Seq-based scRNA-seq with nanowell-specific barcoding enabling precise cell identity assignment. UMAP analysis demonstrated that cell fusion hybrids clustered distinctly from artificially generated doublets and exhibited a unique gene expression signature, clearly separable from either parental cell type or the two artificially combined in the

macrophage-tumor doublets (Figure 13B). These findings validate the integrity of our single-cell -omics pipeline and confirmed that hybrid cell signatures are not artifacts of unresolved doublets.

We then used 10x Genomics to perform high-throughput scRNA-seq of our *in vitro*-derived hybrid fusion cells. The analysis reaffirmed that both hybrid populations were distinct from the parental tumor and macrophage populations (Figure 13C). In addition, we observed phenotypic divergence within the hybrid cell populations: freshly fused hybrid cells (hybrid-M) expressed more immune-like characteristics and aligned more closely with the macrophage population, whereas cultured hybrids (hybrid-T) had a more tumor-like gene expression signature (Figure 2D). This shift likely reflects selection for proliferative hybrid clones during *in* vitro passaging but may also point to biologically distinct subpopulations with clinical relevance. Importantly, even within long-passaged hybrid populations, we identified clonal heterogeneity. For instance, H11 (described in Figure 12) retained strong macrophage-like gene expression and functional traits, supporting the concept of a phenotypic continuum between immune-like and tumor-like states. Gene set enrichment analysis of GO biological processes revealed that both hybrid-M and hybrid-T populations were enriched for pathways involved in cell migration, chemotaxis, antigen processing and presentation, development, and regulation of cell proliferation and differentiation (Figure 13E, F). Collectively, these findings highlight the dynamic and evolving nature of hybrid cell states, which span a phenotypic spectrum between immune and tumor identities. This flexibility may contribute to their enhanced migratory, immune evasive, and metastatic potential, positioning them as important mediators of cancer progression.

In parallel to scRNA-sequencing, we generated s3-scATAC libraries from the same cell populations to evaluate chromatin accessibility changes following cell fusion and the epigenetic reprogramming associated with the acquisition of two genomes. Consistent with our transcriptomic findings, both hybrid-M and hybrid-T populations harbored chromatin accessibility landscapes distinct from those of the parental tumor and macrophage cells, as visualized by UMAP (Figure 13G). Furthermore, similar to the scRNA-seq data, hybrid-M cells clustered more closely with

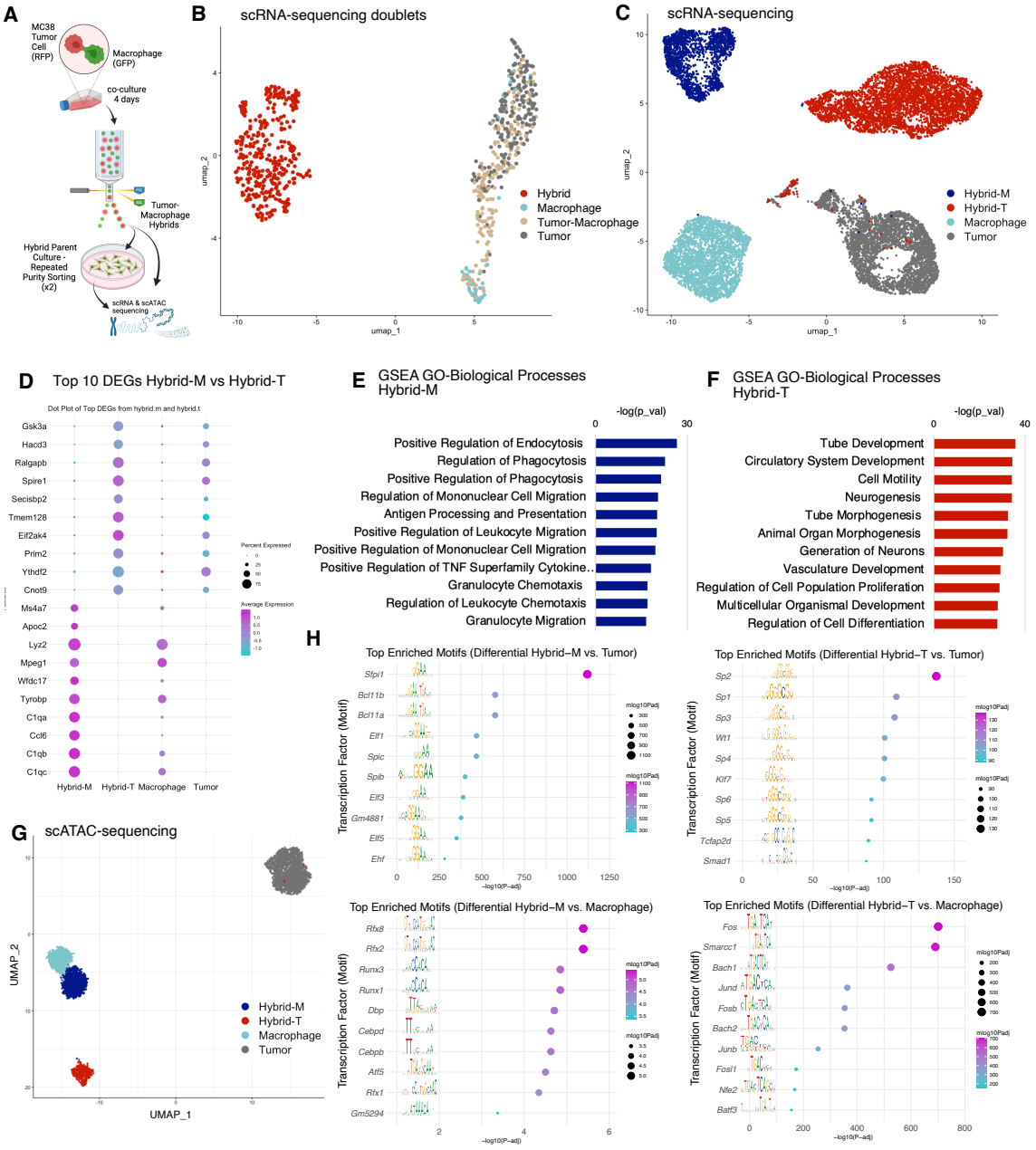


Figure 13: Multi-omic profiling reveals transcriptional and epigenetic heterogeneity of CRC hybrid cell lines.

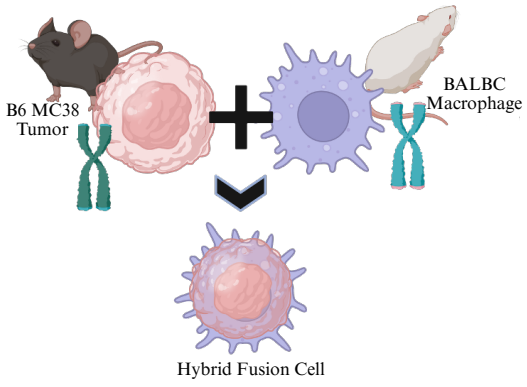
A) schematic of experimental design. B) UMAP of scRNA-sequencing using Takara iCell8 for annotation of macrophage-tumor doublets, hybrid cells and tumor and macrophage cells C) UMAP of scRNA sequencing of "freshly fused" hybrids (hybrid-M), hybrids repeatedly passaged and FACs purity sorted (hybrid-T), macrophages, and MC38 tumor cells. D) Top 10 differentially expressed genes (DEGs) in Hybrid-M and Hybrid-T cells E,F) GSEA GO-biological processes for hybrid-M and hybrid-T cells G) UMAP of scATAC-seq of "freshly fused" hybrids (hybrid-M), hybrids repeatedly passaged and FACs purity sorted (hybrid-T), macrophages, and MC38 tumor cells. H) scATAC-seq motif enrichment analysis for hybrid-M vs macrophage, hybrid-M vs tumor, hybrid-T vs macrophage, and hybrid-T vs tumor. (tumor = MC38 parental tumor cell line)

macrophages, while hybrid-T cells aligned more with tumor cells (Figure 13G), suggesting that transcriptional and epigenetic profiles evolve with prolonged *in vitro* culture. This similarity was readily apparent in analyzing track plots at a host of loci with peaks at transcription start sites, promoters, and enhancers closing trending together.

Motif enrichment analysis revealed population-specific regulatory signatures. In hybrid-M cells, we observed enrichment of ETS transcription factor family members Spi1 (Sfpi1) and Spib, as along with Runx family transcription factors Runx1 and Runx3, transcription factors critical for immune cell function and cancer progression[146,147]. In contrast, hybrid-T cells were enriched for motifs corresponding to Sp family members 1-6, as well as AP-1 transcription factor family members Fos, Fosb, Fosl1, Jund, and Junb, which are critical regulators of cancer cell function including invasion (Figure 13H)[148,149]. Track plots of peak accessibility reveal distinct chromatin accessibility patterns that, when paired with our scRNA-seq data, reinforce the functional relevance of these epigenetic changes. Importantly, the increased motif accessibility in hybrid cells indicate not only elevated transcription factor expression, but also an expanded repertoire of target sites, highlighting how cell fusion generates a distinct regulatory landscape that diverges from parental lineage.

Supplemental Contribution: scWGS in Hybrid Cells

In parallel with the transcriptomic and chromatin accessibility assays, we sought to apply the scWGS method described in Chapter 1 to definitively characterize the genomic architecture of hybrid fusion cells. The fundamental goal was to distinguish genomic contributions from the parental MC38 tumor and macrophage populations (Figure 13A). If feasible, this



Supplementary Figure 1: Visual Abstract for scWGS in Hybrid Fusions

approach would enable us to address whether the observed phenotypic heterogeneity between tumor-like and macrophage-like hybrids stems directly from differential tumor versus macrophage genomic content or reflects more complex regulation in future experiments.

To maximize our ability to discriminate between parental genomes, we designed an initial experiment leveraging strain-specific genetic variation. Primary bone marrow-derived macrophages were isolated from BALB/c mice, while the MC38-H2B-RFP colorectal carcinoma cell line was maintained in its original C57BL/6 (B6) genetic background. These strains differ by an estimated 4 million SNVs[150,151], which equates to about 1 SNV every 650 or so base pairs. This theoretically provides sufficient markers for genome attribution, with approximately half of sequencing reads expected to contain strain-specific SNVs. The subsequent bioinformatic workflow involved aligning scWGS reads from individual hybrid fusion cells to VCF files for both strains, which were publicly available from EMBL-EBI, and quantifying the proportion of strain-specific variants to determine parental genomic lineage.

Unfortunately, initial analysis revealed no discernible difference in strain attribution between hybrid fusion cells and control MC38 tumor cells sequenced in parallel. In both populations, approximately 90% of reads with identifiable strain-specific variants aligned to the B6 reference, with the remaining 10% showing ambiguous or BALB/c attribution. This unexpected result suggested either technical or biological complications in our approach.

Fundamentally, we have already discussed the coverage depth limitations of Chapter 1's scWGS method and how it is more suitable for copy number analysis than reliable SNV-based detection. Further, it's quite possible that the SNVs are not evenly distributed throughout the genome and that large swaths are identical. If with an even distribution a theoretical half of reads might be distinguishable, coupled with the stochastic fragmentation of our transposase-based approach, an uneven distribution of SNVs could represent a significant hurdle. It is also certainly possible that what we observed reflects a measure of biological reality. It is unknown whether there is a preferential retention of tumor genomic content and a selective loss of the macrophage

chromosomes following fusion. What is known, and we have further provided evidence for here, is that chromatin accessibility and transcriptomic profiles in hybrid cells do not merely reflect the tumor parentage (Figure 13C). These findings underscore the importance of the multi-omic approach employed. Future studies utilizing technological advances, particularly in long-read sequencing, may ultimately better resolve whether hybrid fusion cells represent true, stable genomic chimeras.

Increased Runx1 expression correlates with increased hybrid cell migratory and invasive phenotypes

To identify key regulators underlying the phenotypic heterogeneity and invasive capacity of hybrid cells, we performed an integrated analysis of transcriptomic profiles, pathway enrichment, and chromatin accessibility. Across all modalities, RUNX1 consistently emerged as a top candidate, supporting its role as a central regulator of macrophage-like traits and metastatic potential in hybrid cells. Notably, several RUNX1-associated pathways ranked among the top 10 upregulated gene sets in hybrid cells based on Reactome[65] analysis (Figure 14A). To further investigate this enrichment, we examined the expression of key genes within these pathways, including *Runx1, Thbs1, Spi1, Grn,* and *Pf4.* Hybrid-M cells exhibited markedly higher expression of these genes compared to hybrid-T cells, with Thbs1 showing elevated expression across both hybrid populations (Figure 14B). At the epigenetic level, chromatin accessibility at the *Runx1*, *Spi1*, *Thbs1*, and *Pf4* loci were similarly increased in hybrid-M cells, further supporting transcriptional activation of these pathways (Figure 14C–D). These findings, combined with the enrichment of migratory pathways in the hybrid-M population, support the hypothesis that the hybrid cell acquisition of Runx1 following cell fusion promotes hybrid cell migration and invasion.

To explore this, we revisited our hybrid single-cell clones to assess whether elevated Runx1 correlated with increased invasiveness. Indeed, the macrophage-like clones (H11 and H4) exhibited significantly higher RUNX1 expression at both the mRNA (Figure 14E) and protein levels (Figure 14F) compared to tumor-like clones (H3 and H7) and the parental tumor cell line. To validate the

relevance of these findings, we examined hybrid cells identified in human CRC primary tumors and lymph nodes from Figure 11. Consistent with our *in vitro* model, human hybrid cells exhibited elevated expression of *RUNX1,* and associated pathway members *THBS1, SPI1*, and *GRN* (Figure 14G-H). Together, these data indicate that RUNX1 may play a central role in promoting the invasive and immune-like features of hybrid cells, prompting further investigation into whether direct modulation of Runx1 alters their migratory and invasive behavior.

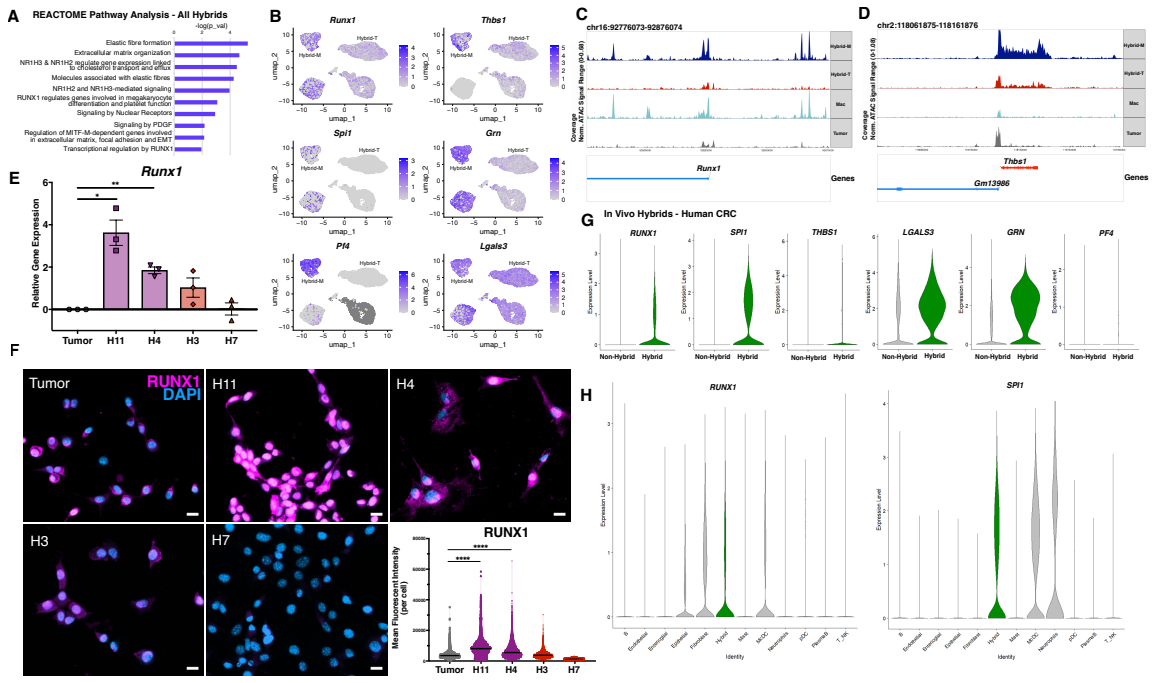


Figure 14: RUNX1 expression is elevated in invasive hybrid cell clones and associates with migratory and proteolytic programs.
A) Reactome pathway analysis using DEGs for all hybrid cells (hybrid-T and hybrid-M) vs tumor cells and macrophages. B) Expression of key Runx1 pathway associated genes contributing to Reactome analysis (*Runx1, Thbs1, Spi1, Grn, Pf4, Lgals3)*, overlaid onto the UMAP. C,D) scATAC-seq track plots for *Runx1* and *Thbs1* in hybrid-M, hybrid-T, tumor and macrophage cells. E) qRT-PCR *Runx1* gene expression in hybrid single cell clones H11, H4, H3, H7 relative to MC38 tumor cells. F) Immunofluorescent staining of Runx1 in hybrid single cell clones H11, H4, H3, H7 and MC38 tumor cells (scale bar = 50uM) and quantified as the mean fluorescent intensity per cell. G) Gene expression of *RUNX1, THBS1, SPI1, GRN, PF4,* and *LGALS3* in human hybrid cells compared to all other cells from scRNA sequencing studies presented in Figure 1. H) Gene expression of *RUNX1* and *SPI1* in hybrid cells compared to all other cells (split by cell type) from scRNA sequencing studies presented in Figure 1. All other Runx1 pathway gene plots included in supplemental figure 3.

Depletion of Runx1 impairs hybrid cell migration and invasion

To directly assess the functional role of RUNX1 in hybrid cells, we generated knockdown cell lines for H11, H7, and the MC38 tumor parent using two independent lentiviral shRNA constructs targeting *Runx1*. Knockdown efficiency was confirmed at both the transcript (Figure 15A) and protein levels (Figure 15B–E). Subsequent phenotypic assays were performed using the shRNA construct (shRNA-1) that produced more consistent knockdown across all three cell lines. In trans-well chemotaxis assays, *Runx1* depletion significantly impaired the migratory capacity of the highly invasive H11 hybrid cells but had no detectable effect on the less invasive H7 hybrid or the MC38 tumor parent (Figure 15F). Notably, *Runx1* depletion did not alter proliferation in any of the cell lines, indicating that its role in H11 cells is specific to migration rather than general cell growth (Figure 15G). Consistent results were observed in ECM invasion assays, where *Runx1* knockdown in H11 led to a marked reduction in both the number of invading cells and distance of invading cells (Figure 15H–I). This diminished invasive phenotype correlated with significant downregulation of invasion-associated proteases associated with matrix degradation (Figure 15J). Together, these findings identify *Runx1* as a critical regulator of the invasive phenotype of macrophage-like hybrid cells and underscore its role in hybrid cell driven tumor progression.

RUNX1+ hybrid cells in matched CRC primary tumor and peripheral blood correlate with disease progression

Having established RUNX1 as a key driver of hybrid cell invasiveness *in vitro*, we next assessed whether RUNX1+ hybrid cells could be detected in human CRC primary tumor and peripheral blood samples, and whether RUNX1, or its pathway components (SPI1 or THBS1) were associated with distinct hybrid phenotypic states such as EMT, stem-like, proliferative, dormant, and immune-evasive states. To do this, we employed highly multiplexed cyclic immunofluorescence (cyCIF) of to evaluate the expression of key Runx1 pathway markers, including RUNX1, THBS1, SPI1, and LGALS3, in the context of hybrid cell identity from 16 patients with CRC.

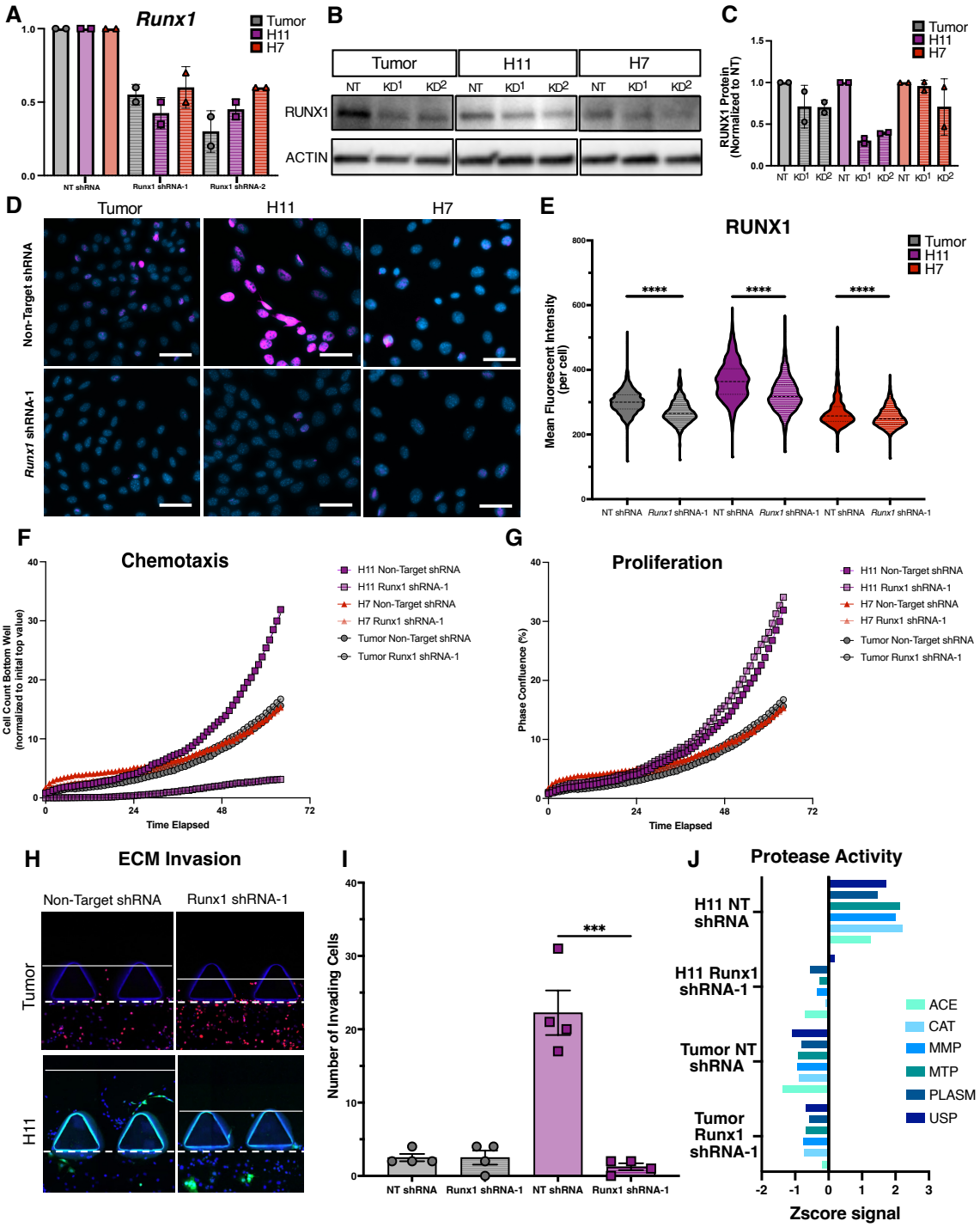

A
Runx1
Tumor
H11
H7
NT shRNA
Runx1 shRNA-1
Runx1 shRNA-2
B
Tumor
H11
H7
NT
KD1
KD2
RUNX1
ACTIN
C
RUNX1 Protein (Normalized to NT)
D
Non-Target shRNA
Runx1 shRNA-1
E
RUNX1
Mean Fluorescent Intensity (per cell)
****
F
Chemotaxis
Cell Count Bottom Well (normalized to inital top value)
Time Elapsed
H11 Non-Target shRNA
H11 Runx1 shRNA-1
H7 Non-Target shRNA
H7 Runx1 shRNA-1
Tumor Non-Target shRNA
Tumor Runx1 shRNA-1
G
Proliferation
Phase Confluence (%)
H
ECM Invasion
Non-Target shRNA
Runx1 shRNA-1
I
Number of Invading Cells
***
J
Protease Activity
H11 NT shRNA
H11 Runx1 shRNA-1
Tumor NT shRNA
Tumor Runx1 shRNA-1
ACE
CAT
MMP
MTP
PLASM
USP
Zscore signal

Figure 15: RUNX1 depletion impairs chemotaxis, invasion, and protease expression in colorectal cancer hybrid cells in vitro.
A) qRT-PCR analysis of *Runx1* expression in MC38 tumor cells and hybrid cell lines H11 and H7 using two independent shRNA constructs and non-target (NT) control. B) Western blot of RUNX1 protein levels in knockdown (KD) samples across tumor, H11, and H7 cell lines, with Actin used as a loading control. C) Quantification of Runx1 protein levels from western blot (B), each cell line normalized to each NT control after normalizing to actin loading control. D) Representative immunofluorescence images of RUNX1 staining in MC38 and H11 NT-shRNA and *Runx1* shRNA-1 cells (scale bar = *** um). E) Quantification of mean fluorescent intensity of Runx1 per cell. *Runx1* shRNA-1 groups compared to non-target controls. F) Trans-well chemotaxis assay of MC38 tumor, H11 and H7 NT-shRNA and *Runx1* shRNA-1 cells. G) Proliferation assay of MC38 tumor, H11 and H7 NT-shRNA and *Runx1* shRNA-1 cells. H) Representative images from extracellular matrix (ECM) invasion assay for MC38 tumor and H11 NT-shRNA and *Runx1* shRNA-1 cells (scale bar = 100um) I) Quantification of ECM-invading cells from (H) J) Z-score normalized protease activity in MC38 tumor and H11 NT-shRNA and *Runx1* shRNA-1 cells. (p-values: **** <0.0001, *** <0.001, **<0.01, *<0.05

Across all patient peripheral blood samples analyzed, CHCs consistently outnumbered CTCs, consistent with prior studies[132,152] (Figure 16A). In addition, CHC abundance also trended upward with advancing disease stage, though this trend did not reach statistical significance. Phenotypic analysis of CHCs and CTCs revealed extensive heterogeneity within the CHC population, identifying multiple distinct UMAP clusters based on biomarker expression profiles (Figure 16B-C, E). Notably, five of these phenotypic clusters (clusters 0, 3, 5, 6, and 14) were characterized by high expression of RUNX1 and/or downstream effectors SPI1 and THBS1 (Figure 16E), suggesting a recurring hybrid cell state defined by RUNX1 activation. Furthermore, three of the five RUNX1+ clusters (clusters 0, 3, 6) co-expressed vimentin (VIM), a marker associated with EMT. Notably, these clusters were comprised of CHCs, and not CTCs (which were more prevalent in cluster 5). This suggests that $RUNX1^{+}$ hybrid cells in human hybrids may possess a more invasive EMT phenotype as was similarly observed in our murine *in vitro*-derived hybrids. Moreover, RUNX1 expression within CHCs increased with disease stage, further supporting its association with tumor progression (Figure 16F).

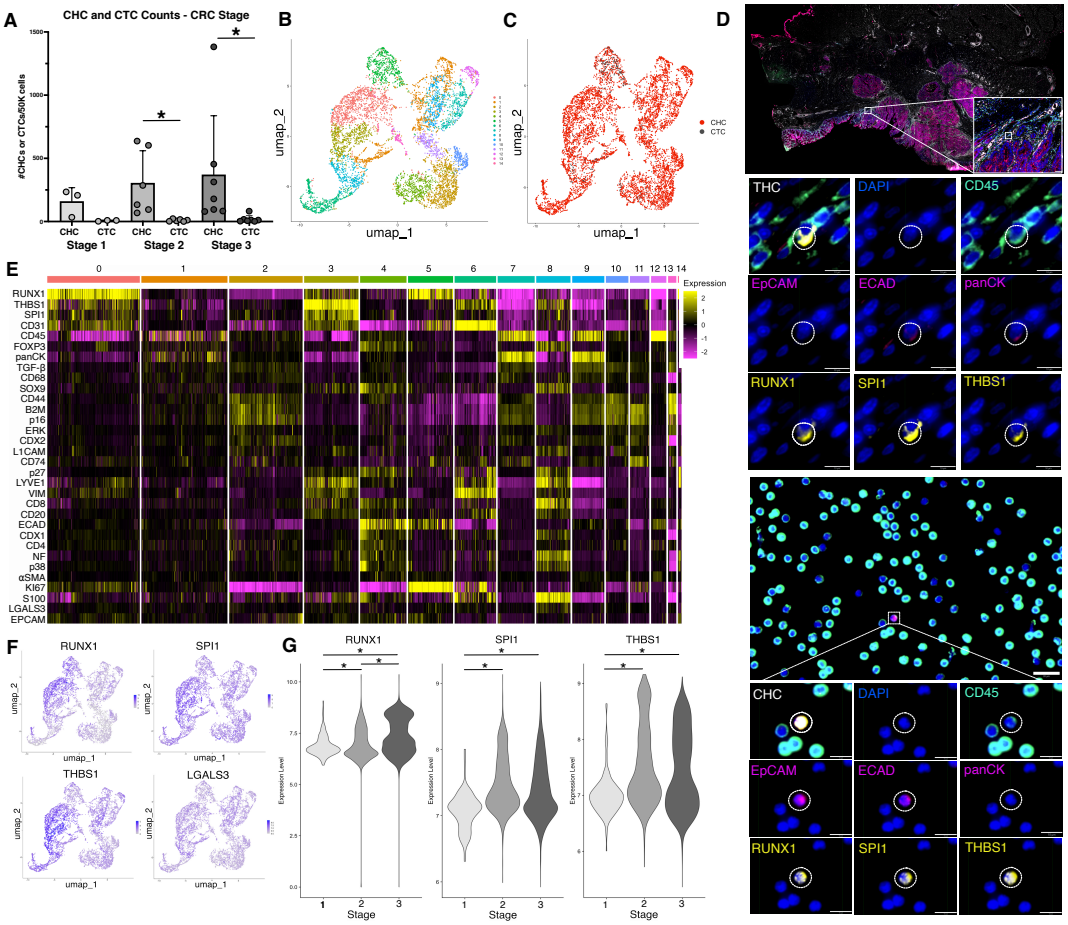


Figure 16: Cyclic immunofluorescence reveals increased RUNX1 expression in hybrid cells from colorectal cancer tumors and peripheral blood.

A) Quantification of circulating hybrid cells (CHCs) and circulating tumor cells (CTCs) per 50,000 nucleated cells in peripheral blood across CRC stages 1, 2, and 3. B) UMAP plot of cyclic immunofluorescence (cyCIF) data colored by phenotypically defined clusters. C) UMAP plot from (B) labeled by CHC versus CTC identity.

D) Representative cycIF image of a whole-slide CRC primary tumor showing RUNX1+ tumor hybrid cells (THC) and Runx1+ CHC (hybrid cells identified as DAPI+, CD45+, EpCAM+ or ECAD+ or panCK+) including Runx1 downstream pathway members SPI1 and THBS1. (scale bars; tumor inset = 50uM, THC individual images = 10uM, peripheral blood = 20uM, CHC individual images = 10uM). E) Heatmap of 34 phenotypic markers measured by cyCIF across individual CHCs and CTCs, clustered by expression patterns. F) UMAP feature plots showing normalized average fluorescent intensity levels of RUNX1, SPI1, THBS1, and LGALS3 across all single cells. G) Violin plots of RUNX1, SPI1, and THBS1 expression in CHCs grouped by CRC disease stage (p-val = * < 0.05).

These findings highlight RUNX1$^+$ hybrids cells as a transcriptionally and phenotypically distinct subpopulation enriched in the blood of patients with CRC, particularly at later disease stages. Given their abundance, EMT-like features and associated with key metastatic pathways, RUNX1$^+$ hybrid cells may serve as a clinically relevant biomarker with promising utility for both disease monitoring and therapeutic intervention in CRC.

Single-cell RNA sequencing of matched tumor and peripheral blood hybrids reveals patient-derived hybrid cell subpopulations resembling in vitro phenotypes

To gain deeper insight into the transcriptional heterogeneity of both tumor-associated hybrid cells and CHCs within an individual patient, we performed scRNA-seq on isolated hybrid cells (ECAD$^+$/EPCAM$^+$/CD45$^+$) from matched primary tumor (THCs) and peripheral blood samples (CHCs) in CRC. Given the rarity of hybrid cells obtainable from clinical specimens, we employed SMART-Seq (Takara Bio), a platform optimized for generating full-length transcriptomes from picogram-scale RNA inputs. For each patient, control tumor cells and PBMCs were also isolated based upon ECAD$^+$/EPCAM$^+$ tumor cells and CD45$^+$ immune cells (n = 16 cells/wells per patient).

Unsupervised clustering revealed multiple distinct CHC transcriptional clusters (CHC1-3), with representation from all three patients analyzed (Figure 17A-C). Both THCs and CHCs exhibited transcriptional similarity to CD45 immune cell populations, clustering closely with immune cell populations across several clusters. In contrast, very few CHCs cluster in the EpCAM$^+$ tumor control cluster suggesting that the CHCs acquire immune-like transcriptional features during dissemination or that hybrid cells require certain immune-like characteristics in order to disseminate. Differential gene expression and pathway enrichment analysis between CHCs and THCs revealed upregulation of chemotaxis, developmental and ECM reorganization pathways in CHCs (Figure 19) highlighting shared and divergent signatures.

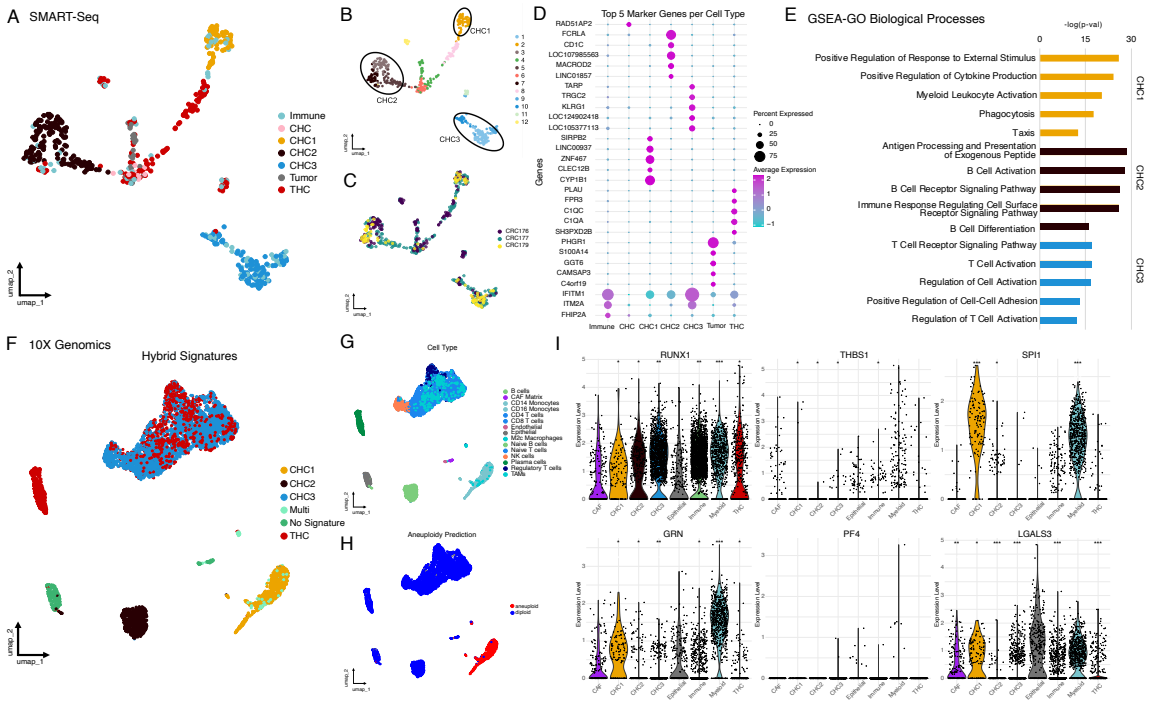


Figure 17: Single-cell RNA sequencing confirms RUNX1 pathway activation in hybrid cells from patient-matched colorectal tumors and circulation

A) UMAP projection of SMART-Seq data showing clustering of tumor hybrid cells (THCs, EpCAM$^+$/ECAD$^+$/CD45$^+$), circulating hybrid cells (CHCs, EpCAM$^+$/ECAD$^+$/CD45$^+$), tumor cells (EpCAM$^+$/ECAD$^+$), and immune cells (CD45$^+$) from patient-matched colorectal tumor and peripheral blood samples (n = 3 patients; n=3 PBMC samples, n=2 tumor samples). B) UMAP colored by CHC subclusters (CHC1–3) with distinct transcriptional states. C) UMAP colored by patient ID. D) Dot plot showing the top five differentially expressed genes defining each major cell type. E) Gene set enrichment analysis (GSEA) of GO biological processes highlighting functional differences among CHC1–3 clusters. F) UMAP of 10x Genomics data enriched for tumor and circulating hybrid cells, colored by hybrid signature scores derived from SMART-Seq CHC cluster modules. G) UMAP of the same 10X dataset annotated by major cell types, showing distribution of hybrid-enriched populations. H) Aneuploidy prediction in 10X dataset distinguishing diploid versus aneuploid populations. I) Violin plots of representative genes (RUNX1, THBS1, SPI1, PF4, LGALS3) demonstrating enrichment of RUNX1-associated transcriptional programs in both CHCs and THCs.

Differential gene expression analysis distinguished each hybrid subpopulation from tumor and immune controls (Figure 17D). Pathway analysis revealed that CHC1 cells were enriched for phagocytic and cytokine response programs similar to macrophages, CHC2 cells for B-cell activation pathways, and CHC3 cells for T-cell activation and adhesion processes (Figure 17E).

These findings suggest that discrete hybrid states in circulation reflect functional immune programs.

In a subset of patient samples where sufficient hybrid cell yields were obtained, we performed 10x Genomics-based scRNA-seq on pooled non-hybrid cells from either the primary tumor or peripheral blood spiked with an enrichment of hybrid cells ($ECAD^+/EPCAM^+/CD45^+$) to capture a broader cellular landscape (Figure 17F-I). Although these samples were enriched for THCs or CHCs, we expected them to also contain a mixture of immune cells, epithelial, stromal, and neoplastic CRC cells, particularly in tumor-derived biopsies. Because these cells were not sorted into uniquely barcoded wells, as in the SMART-Seq, surface marker-based annotation was not feasible.

To overcome this limitation, we leveraged the SMART-Seq dataset to generate a hybrid gene module for each CHC cluster 1-3, which we then applied to the 10x dataset to score and identify hybrid cells (Figure 17F). Additional cell populations were annotated using previously established gene modules corresponding to epithelial, immune, stromal, and macrophage cell types (Figure 17G). Within the 10x dataset, multiple transcriptionally distinct hybrid-enriched clusters emerged based on hybrid module scoring (Figure 17F). One cluster showed strong alignment with macrophage-like gene expression patterns, while the others aligned with annotated B-cell and T-cell populations. It is conceivable that these discrete hybrid cells arose from fusions with B and T-cell lymphocytes. B lymphocytes are commonly fused with immortalized cell lines to generate monoclonal antibody producing hybridomas[153].

Aneuploidy prediction confirmed that the macrophage-like hybrid cells are aneuploid, while the rates of aneuploidy in the B-cell and T-cell hybrids were much lower (Figure 17H). Additionally, we generated a tumor gene module from the SMART-Seq dataset and applied it to the 10x data set (Figure 19E). The macrophage-like hybrids scored high for the tumor module, much like the epithelial cells and tumor hybrid cells. Notably the T-cell hybrids in CHC1 scored distinctly low, and the B-cell like circulating hybrids contained subpopulations with both high and

low classification (Figure 19E-F). We saw this as confirmation that the hybrid sub-groups were not just misassigned macrophage or B-cells.

Across both platforms, RUNX1 and its downstream effectors (SPI1, THBS1, PF4, LGALS3) were significantly enriched in CHCs and THCs relative to parental tumor controls (Figure 17I and Figure 19C). These RUNX1+ associated transcriptional programs overlapped with pathways previously identified in vitro, including migration, chemotaxis, and immune activation. Furthermore, RUNX1 and downstream associated genes were highest in CHC1 cluster, the macrophage-like hybrid cell population with high aneuploidy prediction. Together, these results demonstrate transcriptional convergence between patient-derived and experimentally generated hybrids and implicate RUNX1-driven programs as central regulators of invasive hybrid phenotypes in CRC.

Discussion

Metastasis remains a leading cause of cancer-related death, underscoring our incomplete understanding of how tumor cells acquire and sustain metastatic capabilities while evading current therapies. To address this gap, we investigated the role of neoplastic-immune hybrid cells, formed through spontaneous fusion between tumor and immune cells, in driving CRC metastasis. Using an integrated approach, including scRNA-seq, scATAC-seq, highly multiplexed cyclic immunofluorescence (cyCIF), and functional assays, we identified hybrid cells as a transcriptionally and phenotypically diverse and distinct cell population enriched for migratory, immune-evasive and stem-like programs.

A key finding of this study is the identification of RUNX1, a transcription factor known for its role in hematopoiesis and leukemogenesis, as a key regulator of hybrid cell invasiveness and a potential driver of metastatic progression in CRC. RUNX1 was highly expressed in hybrid cells with macrophage-like features and was functionally required for invasion and ECM degradation. Depletion of *Runx1* in these hybrids (hybrid-M) significantly impaired migration, invasion, and

protease expression, indicating that RUNX1 is functionally required for hybrid cell invasion. These findings are consistent with previous reports linking RUNX1 to CRC progression, EMT, angiogenesis, and immune modulation[154–159]. For example, RUNX1 promotes metastasis and EMT by activating Wnt/β-catenin signaling, enhances PTGS2 (COX2) expression, promotes vessel co-option, and drives tumor-stromal crosstalk[160,161] via TGF-β and THBS1 signaling[162]. This effect has been validated in both *in vitro* and *in vivo* CRC models[163].

In both *in vitro* hybrid cell fusion models and patient-derived specimens, hybrid cells displayed a phenotypic continuum between macrophage- and tumor-like states, marked by transcriptional plasticity. RUNX1, along with SPI1 and THBS1, were enriched in macrophage-like hybrids, which demonstrated the greatest invasive potential across transcriptomic chromatin accessibility and proteomic analyses. These findings, combined with our hybrid-specific data, highlight RUNX1 as a multifaceted driver of metastasis that promotes EMT, migration, angiogenesis, and neoplastic-immune hybrid cell dissemination. Importantly, RUNX1+ hybrid cells were consistently present in primary tumors and circulation, with increasing prevalence in advanced-stage CRC.

Evaluation of hybrid single-cell phenotyping using both scRNA-sequencing and cyCIF, we identified RUNX1-positive hybrid cells as a transcriptionally distinct subpopulation that are present in both primary tumors and circulation. These RUNX1+ hybrid cells are enriched for genes and proteins associated with migration, immune modulation, and stemness, reinforcing their significant role in CRC metastasis.

Transcriptomic analysis of patient-derived hybrid cells revealed striking similarities to the hybrid-M and hybrid-T phenotypes observed in our *in vitro* model. This convergence supports the biological relevance of our model system and suggests that RUNX1-associated programs are conserved features of invasive hybrid cell states in human CRC. Multiple distinct hybrid cell clusters were identified across patients, with several enriched for RUNX1 or its downstream effectors including SPI1 and THBS1. These findings further implicate RUNX1 as a central

regulator of hybrid cell identity and metastatic potential. Our results advance the understanding of tumor-immune hybrid cells as dynamic contributors to metastasis and position RUNX1 as a key factor driving their invasive behavior. The consistent expression of RUNX1 across *in vitro* and patient-derived hybrid cells suggests that it may serve as a useful biomarker for identifying metastatic prone cell populations. Its functional role in regulating invasion also raises the possibility that RUNX1 could be targeted therapeutically to limit hybrid cell dissemination and improve clinical outcomes in colorectal cancer, as the RUNX1 inhibitor Ro5-3335 has shown promise in leukemia, PDAC, and glioblastoma[164–166].

While this study provides important insights into hybrid cell biology and metastatic progression in colorectal cancer using a multi-model approach, several limitations should be acknowledged. Throughout our studies, hybrid cells are identified by their co-expression of epithelial tumor and immune cell markers however this approach may not capture the full spectrum of fusion derived hybrids, particularly those that undergo phenotype switching or lose parental markers over time[108]. More sophisticated lineage tracing (i.e. MADM-CloneSeq[167,168]) or fusion reporter systems (Cre-Lox fate mapping[169]) *in vivo* would improve the specificity of hybrid cell identification. While our *in vitro* model recapitulates key features of patient-derived hybrid cells and exhibits heterogeneity that is also seen in patient hybrid cells, further validation using *in vivo* models is needed to evaluate the role of RUNX1 in hybrid cell dissemination and metastatic seeding. Additionally, although RUNX1 is necessary for hybrid cell invasion, it remains unclear whether it is sufficient to drive a hybrid-macrophage like phenotype on its own. Future work should explore how RUNX1 influences other hybrid cell traits such as immune evasion or stemness, and how it may be modulated by the tumor microenvironment or in circulation. Although this study includes hybrid cells from patient-matched tumor and peripheral blood samples, the number of patient samples and hybrid cells captured for deep transcriptomic profiling is relatively limited. This constrains the ability to assess inter-patient variability and evaluate clinical correlations. Hybrid cell evaluation in larger patient cohorts is needed to determine whether RUNX1 expression

reliably predicts metastatic potential or correlates with clinical outcomes, in particular for CHCs which can be used as a non-invasive biomarker.

In summary, this study identifies RUNX1 as a pivotal driver of hybrid cell invasion in CRC. To our knowledge, this is the first work to perform scRNA-seq of hybrid cells from patient matched tumor and peripheral blood samples, providing a valuable resource for the research community and enabling exploration of hybrid cell biology across the metastatic cascade. Furthermore, this study is the first to identify hybrid cells as distinct cellular identities from sequenced doublets, confirming that their transcriptional profiles reflect true biological fusion events rather than technical artifacts. Beyond this technical advance, our findings provide important mechanistic insight into how tumor–immune fusion contributes to cancer progression. By characterizing hybrid cells as both transcriptionally distinct and functionally invasive, we highlight this previously underappreciated driver of metastasis and a promising target for therapeutic intervention and biomarker development in CRC.

Materials & Methods

Tumor-macrophage hybrid cell identification and analysis in publicly available scRNA-sequencing

Single-Cell Data Preprocessing

scRNA-seq data of primary and metastatic CRC patient samples were obtained from a publicly available dataset[139]. Count matrices were imported using Read10X_h5() and converted into Seurat v5 objects with default thresholds (min. 3 cells per gene, min. 200 features per cell).[84,85] Metadata files containing cell-level annotations (e.g., sample origin, patient ID, and original cell type) were matched to barcode identifiers and integrated using AddMetaData(). The epithelial and non-epithelial objects were merged using merge() to generate a combined dataset. Quality control was assessed by visualizing metrics such as gene count (nFeature_RNA), transcript count (nCount_RNA), and mitochondrial gene percentage (percent.mt) using VlnPlot(). The merged object was then normalized (NormalizeData()), variable features were identified (FindVariableFeatures()), data were scaled (ScaleData()), and dimensionality reduction was performed via principal component analysis (PCA). Nearest-neighbor graphs and clusters were generated using FindNeighbors() and FindClusters() (resolution = 2.0), and a UMAP embedding was computed with RunUMAP() (dims = 1:30).

Hybrid Cell Classification

Hybrid tumor-immune cells were defined based on co-expression of epithelial and immune markers within the RNA assay. Epithelial identity was assigned to cells expressing EPCAM, CDH1 (E-cadherin), and pan-cytokeratin (including KRT2–5, KRT7–8, KRT14–16, KRT19), consistent with AE1/AE3 immunoreactivity. Immune identity was assigned based on expression of canonical markers including PTPRC (CD45), and macrophage markers CD14, CD68, CD163, and ITGAM.

Cells expressing at least one epithelial marker and one macrophage marker were classified as hybrids using WhichCells(). Subsets of hybrids enriched for macrophage lineage markers were annotated as “tumor-macrophage hybrids.” Hybrid classifications were stored in the metadata under hybridcells and counts per cell type and patient were tabulated using table().

Doublet Detection and Filtering

To evaluate doublet scores for hybrid cell populations and remove potential doublets, cells were scored using the scds package with three methods: cxds(), bcds(), and the cxds_bcds_hybrid() model based on prior work[140,170,171]. Hybrid scores were assigned to metadata (cxds-bcds-score, cxdsscore, bcdsscore) and visualized using violin plots (VlnPlot()). Cells with cxds-bcds-score ≥ 0.5 were removed from downstream analyses. Additional doublet scores were calculated using scDblFinder based on our prior work in uveal melanoma[138]. Seurat object filtering and visualization steps ensured consistency across doublet detection methods. The updated Seurat object excluding high-scoring doublets was used for all downstream hybrid analyses.

Differential Expression and Pathway Analysis

Differentially expressed genes (DEGs) were identified using Seurat's FindMarkers() function with a minimum detection threshold of 25% (min.pct = 0.25). Comparisons were made between hybrid cells and other cell types (epithelial, macrophage-like), as well as between hybrid cells from tumor and lymph node (LN) sites. Top up- and down-regulated genes were visualized using heatmaps (DoHeatmap()) and dot plots (DotPlot()), with $\log_2$ fold changes used to rank genes. For pathway analysis, gene set enrichment analysis (GSEA) was performed using the clusterProfiler package[172] (gseGO()) with the Gene Ontology Biological Process (GO-BP) database. Gene symbols were converted to Ensembl IDs via org.Hs.eg.db::mapIds(), and redundant gene entries were collapsed by retaining those with the highest absolute log fold change. GSEA plots and enrichment maps were generated using dotplot() and emapplot() from the enrichplot package.

Hybrid Cell Subsetting and Clinical Correlation

Clinical metadata including sex, age, tumor stage, MSI/MSS status, and mutational status (e.g., KRAS, BRAF, TP53, APC) were integrated into the object and stored in new metadata columns. Hybrid cell counts were assessed for all available clinical metadata. Hybrid cell analyses were restricted to matched tumor-LN patient samples (n = 7) where applicable using subset() by patient.ID in Seurat. RUNX1 pathway activity was examined by comparing expression of key

regulators (RUNX1, SPI1, THBS1, LGALS3, PF4, GRN) in hybrid vs. epithelial or macrophage populations and across cancer stage. Group comparisons were performed using two-sample t-tests and plotted using VlnPlot().

Generation of in-vitro derived fusion hybrids and single cell clones

MC38-H2B-RFP colorectal cancer cells were co-cultured with bone marrow-derived macrophages isolated from β-actin-GFP C57BL/6 mice to generate spontaneous hybrid cells as previously described[108]. Briefly, bone marrow cells were harvested from femurs and tibias of a β-actin-GFP C57BL/6 female mouse and were cultured in high glucose DMEM supplemented with 15% FBS, 1% Pen/Strep, and 25 ng/mL recombinant M-CSF to promote macrophage differentiation over 4 days. After differentiation, cells were plated at a 2:1 ratio with MC38 cells and co-cultured for 4 days in high glucose DMEM supplemented with 10% FBS, 1% NEAA, 1% HEPES, 1% Pen/Strep, and 5 ng/mL recombinant M-CSF. Double-positive GFP+/RFP+ cells were isolated via FACS using a BD influx and processed either immediately for scRNA and scATAC sequencing as described in more detail below (freshly fused; hybrid-M) or expanded through serial passaging and purity sorting for GFP+/RFP+ cells (hybrid-T). To generate single cell clones, sorted single hybrid cells were deposited into 96-well plates (H1-H12) and wells with single cells were annotated 30 minutes after FACS sorting and cell growth was assessed twice weekly until single cell colonies reached 70% confluence and were further expanded and frozen down for future studies. All experiments were completed for cell lines between passage 8-15.

Phenotypic evaluation of hybrid single cell clones

qRT-PCR

Total RNA was extracted using the RNeasy Plus Mini Kit (Qiagen). RNA integrity was assessed with a Nanodrop. cDNA was synthesized using the High-capacity cDNA synthesis kit (Applied Biosystems). qRT-PCR was performed with SYBR Green PCR Master Mix and primer sets from Table 2 on a CFX Opus 384 (Bio-Rad). Relative gene expression was calculated using the ΔΔCt method, using GAPDH as the internal reference gene. Results are mean ± SEM from

triplicate analyses averaged from three biological replicates. Heatmaps and hierarchical clustering were generated using the pheatmap package in R.

Proliferation

To assess cell proliferation, cells were plated in 96-well plates at equal density and monitored over a 72-hour period using the IncuCyte® Live-Cell Analysis System (Sartorius). Phase contrast images were acquired every hour in each well using a 10× objective. Proliferation rates were quantified using the percentage of phase object confluence over time to calculate doubling time (graphed as "1/doubling time") using IncuCyte's integrated image analysis software. All experiments were conducted in triplicate, with biological replicates, and data were normalized to initial confluence at time zero.

Chemotaxis

Cells were seeded at equal density in the upper chambers of IncuCyte® ClearView 96-well Chemotaxis Plates (Sartorius) in serum-free media. The lower wells contained complete media supplemented with 10% fetal bovine serum (FBS) as a chemoattractant. The plate was imaged every hour for 60 hours using the IncuCyte® Live-Cell Analysis System with a 10× objective, capturing phase contrast images of cells migrating through the membrane to the underside of the upper chamber. Cell migration was quantified using IncuCyte's integrated chemotaxis analysis module, measuring cell count in the lower focal plane over time, normalized to the initial top value count. Each condition was run in at least triplicate wells per experiment, with 2 or more biological replicates for both the hybrid single cell phenotyping studies (Figure 12) and Runx1 shRNA studies (Figure 15).

ECM collagen invasion on-a-chip

To establish the invasion on-a-hip, we used an identTx3 microfluidic device (Mattek). The device contains a central channel (1.30 mm in width and 0.25 mm in height) separated by pillars from two adjacent parallel channels (0.5 mm in width and 0.25 mm in height). To improve collagen adherence, the central chamber of the devices was coated with 1 mg/mL of poly-D-lysine (PDL) (Gibco) for three hours at 37 °C, then washed with ultrapure water and allowed to dry overnight

inside the biosafety cabinet. Next, we prepared collagen hydrogel by mixing 833 µl of acid-solubilized type I collagen from rat tail (3 mg/mL, Gibco) with 100 µl of 10x PBS, 20 µl of 1 M NaOH, and 47 µl of DMEM without serum to achieve a working pH of 7.2 to 7.4. Subsequently, 10 µL of collagen was pipetted into the central channel carefully to avoid bubbles. Devices were maintained in the incubator at 37^{o}C for 45 min to allow collagen fibrillogenesis, then the lateral channels were filled with media to prevent collagen dehydration.

On the next day, cells (MC38 tumor, H11, H4, H3 or H7) in serum-free media were seeded in one of the lateral channels of the device while on the other channel, DMEM with 10% FBS was placed to act as a chemoattractant for the cells. The devices were then placed in the incubator to allow cells migration into the collagen gel. Three images spanning across each channel were acquired every 24 hours for three days using a Nikon spinning disk live-cell imaging system with z-stack to assess cell migration in x,y and z planes using a 10x objective. After 72 hours, the cells were fixed and stained with DAPI and imaged using the same microscope and imaging parameters. The number of invading cells were quantified by manual annotation in a blinded observer using Zen. Data presented are the total cell counts from 2 biological replicates and 2 technical replicates n=4 per cell line.

Vasculature invasion on-a-chip

Fabrication of the microfluidic device

A CAD program was used to design a mold composed of two reservoirs connected by a central channel and a chamber as published[173]. We then used a three-dimensional (3D) printer (CADworks3D µMicrofluidic printer, Profluidics 285D) and resin (Master Mold resin, CADworks) to print the positive molds. Printed micromolds were cleaned in methanol in three rinses of 2 min each under agitation, subsequently were cast with PDMS (Polydimethylsiloxane - Sylgard 184, Dow-Corning), and left to cure overnight at 80°C, as previously described (23, 25). Next, PDMS was removed from the resin mold, and two reservoirs were prepared using a 5-mm biopsy punch for media, 1-mm biopsy punches for the collagen loading ports and 2.5 mm punch on top of the

central chamber where collagen will be placed. Molds were cleaned with ethanol and immediately plasma bonded to a glass coverslip. Assembled devices were autoclaved, then treated with 1% (w/v) glutaraldehyde (Sigma-Aldrich) for 15 min, rinsed three times with distilled water (DIW), and left overnight in DIW to remove any trace of glutaraldehyde. To mold cylindrical channels, sterile 160-μm-diameter acupuncture needles were immersed in a 0.1% bovine serum albumin (BSA) solution for at least 30 min and then inserted into the central channel of the devices ~200 μm above the glass coverslip surface. Rat tail collagen type I (3.0 mg/ml, Gibco) was prepared according to published protocol[173]. Briefly, a stock solution of collagen type I (collagen I, rat tail Gibco, 3 mg/ml) was prepared on ice by mixing 833 μl of collagen with 100 μl of 10x PBS, 20 μl of 1 M NaOH, and 47 μl of EGM to reach a working pH of 7.2-7.4. Thirty microliters of the final solution were immediately pipetted into the middle chamber of the device for the vasculature and allowed to polymerize at 37°[173]. To prevent collagen dehydration, after 1 hour, the main reservoirs and the top of the collagen hydrogel were filled with EGM, then devices were returned for incubation overnight. On the following day, the needles were carefully removed with a pair of tweezers, and the cell medium (EGM) was replaced by fresh medium. Subsequently, the devices had their reservoirs filled with fresh EGM, and were placed in the incubator at 37°C and 5% CO2 overnight before cell seeding.

Cell culture

Human umbilical vein endothelial cells (HUVECs expressing green fluorescent protein (GFP-HUVECs) (Lonza, Basel, Switzerland) were cultured in a supplemented (EGM-2 with bullet kit, Lonza). Human bone marrow mesenchymal stem cells (hMSCs) (RoosterBio, MD) were cultured in α–minimum essential medium (Gibco) with 10% FBS) and 1% penicillin/streptomycin. Cell media were changed every other day, and cells were passaged when reaching a confluency of 80 to 90%. HUVECs at passages 4 to 6 and hMSCs at passages 2 to 4 were used for all the experiments. All cells were maintained in a humidified incubator (5% CO_2 and 37°C).

Cell seeding

For seeding, GFP-HUVECs and hMSCs were detached using TripLE, counted, and mixed at a 4:1 ratio according to previous publications[173] in a cell density of 6 million cells/ml. Subsequently, the cell medium was removed from the reservoirs, and 25 µl of the cell suspension was added into one reservoir. The devices were flipped upside down, placed in the incubator for 5 min, seeded again, and left upside down in the incubator for another 5 min. Until the entire extension of the collagen channel had cells attached, we repeated the seeding, flipping the chip as needed. Next, the devices were placed in the incubator for 30 min under static conditions. Afterward, the devices were transferred to the two-dimensional (2D) rocker (BenchRocker) inside the incubator as published[173,174].

After 24h, vasculature was formed and 4 x10^5 hybrid cells or unfused parent cells were seeded on the central reservoir of the devices, on top of the collagen so that cells were 800 um away from the vascular channel. The devices were maintained in a humidified incubator (5% CO_2 and 37°C) for 24h and fixed with 4% PFA (v/v) in PBS for 30 min.

Cell staining and imaging

Cells were permeabilized with 0.1% (w/v) Triton X-100 for 10 min and blocked with 1.5% (w/v) BSA for 1 hour under agitation. After washing with PBS, samples were incubated with one of the following primary antibodies (anti-PECAM-1, rabbit anti-human, cat. no. AB32457, Abcam, 1:100; or anti-NG-2, mouse anti-human, cat. no. 14-6504-82, Invitrogen, 1:200) overnight at 4°C. Samples were washed with PBS and incubated with secondary antibody (goat anti-mouse Alexa Fluor 555, Invitrogen, 1:250; goat anti-rabbit Alexa Fluor 647, Invitrogen, 1:250) overnight at 4°C under agitation. This was followed by rinsing in 0.1% PBS, staining of the nuclei using NucBlue [Fixed Cell ReadyProbes, 4′,6-diamidino-2-phenylindole (DAPI), Molecular Probes], and staining of actin with ActinGreen 488 (ReadyProbes, Molecular Probes) for 1 hour at 37°C under agitation.

Samples were imaged using a confocal microscope (Zeiss, LSM 880, Germany) with a 10× objective (numerical aperture, 0.45; Zeiss, Plan Apochromat). The depth of imaging was 100 to 400 µm, split in at least 100 *z*-stacks. *z*-Stacks were converted into TIFF files or 3D images using

Zen or Imaris software (version 9.1, Bitplane, Oxford Instruments, Zurich, Switzerland). The number of migrating cells were counted using Imaris.

In vivo tumor models

All animal procedures were conducted in accordance with protocols approved by the Oregon Health & Science University Institutional Animal Care and Use Committee (IACUC). Mice were maintained in a specific pathogen-free facility under a 12-hour light/dark cycle with unrestricted access to standard rodent chow (5001, PMI Nutrition International, Richmond, IN) and water. Experimental cohorts included adult mice aged 8 to 12 weeks, with balanced representation of both sexes, and littermate controls were used where applicable. MC38, H11, and H7 cells were injected into C57 subcutaneously at a dose of 50,000 cells diluted 1:1 in sterile PBS/Matrigel. Tumor volume and body weight was measured twice weekly until tumor endpoint (reaching a maximum tumor size of 1200mm^3). At that time, mice were euthanized, and tumors were harvested for FFPE and OCT embedding.

Runx1 shRNA studies

MC38-RFP tumor cells and two MC38xY01-derived hybrid clones (Clone 7 and Clone 11) were cultured in high glucose DMEM medium supplemented with 10% FBS, 1% NEAA, 1% HEPES and 1% penicillin-streptomycin. Neomycin (G418, Sigma Aldrich) titration was performed prior to transduction to determine the minimum concentration required to eliminate non-transduced cells. Cells were seeded in 24-well plates and treated with a dilution series of neomycin (G418). At 1000 μg/mL, all cells were killed within 4 days. MC38 tumor cells showed slightly higher sensitivity and were also maintained under the same neomycin concentration for consistency.

Lentiviral shRNA particles targeting RUNX1 (TRCN0000084810 and TRCN0000084812) and a non-targeting control (TRCN0000072229) (Sigma Aldrich) were used at calculated multiplicities of infection (MOIs) based on viral titers and cell counts (Supplementary Table 5.4). Cells were seeded into 24-well plates at densities to reach target densities for MOI calculation the following day. Immediately prior to transduction, media was replaced with fresh growth medium

containing 8 µg/mL polybrene (hexadimethrine bromide), except for one control well that received polybrene-free medium to monitor potential toxicity. The appropriate amount of viral supernatant was added directly to each well based on calculated MOI of 20 viral particles per cell. Plates were centrifuged at 2500 RPM for 90 minutes at 30 °C (no brake) to facilitate viral entry, then returned to the incubator overnight. The next day, viral-containing media were aspirated and replaced with fresh growth medium for 24hrs before starting selection with neomycin (G418) containing media. Transduced cells were selected by growing in media containing 1000 µg/mL neomycin. Brightfield images were captured at each media change to document cell viability and morphology.

Due to insufficient knockdown efficiency observed in preliminary experiments, a second transduction was performed using the same procedure. This repeat transduction was carried out after the first round of selection and expansion. Cells were re-seeded at appropriate densities, and transduction was repeated as described above. Selection with 1000 µg/mL neomycin was re-applied and cells were maintained under selection for 72hours before downstream analyses including qRT-PCR, western blot, chemotaxis, proliferation, ECM invasion, and protease activity as previously described.

Western blots

Cells were harvested for western blot analysis using trypLE (Gibco, FisherSci) to detach adherent cells. After pelleting cells were washed once with 1X PBS, pelleted and were then either flash-frozen in liquid nitrogen or immediately lysed. Cell lysis was performed using RIPA buffer (ThermoFisher Scientific, 89900) supplemented with 1 mM PMSF and one protease inhibitor mini-tablet (per 7 mL buffer) and phosphatase inhibitor cocktail 2 (Sigma-Aldrich). Lysis buffer was added at 60µL per 1×10^6 cells. Samples were incubated on ice for 20 minutes to allow solubilization. Samples were sheared to reduce viscosity via vigorous vortex. Lysates were centrifuged at 15,000 × g for 10 minutes at 4 °C, and supernatants were transferred to new tubes. Protein concentration was normalized using a BCA assay. Lysates were mixed with 3× GS loading buffer (6% SDS, 150 mM Tris pH 8.0, 30% glycerol, 0.1% bromophenol blue) containing 1× final

concentration of β-mercaptoethanol (BME) and heated at 95 °C for 5 minutes. Samples were loaded into Criterion TGX gels (Bio-Rad) and electrophoresed at 130–150 V for ~1 hour in 1× running buffer. Proteins were transferred to PVDF membranes (activated in ethanol) using the Bio-Rad Trans-Blot Turbo system according to manufacturer-specified settings based on protein molecular weight. Membranes were blocked for 5 minutes at room temperature in EveryBlot blocking buffer (Bio-Rad, 12010020). Blots were incubated overnight with primary antibodies at 4 °C (Runx1 (19555-1-AP, ProteinTech) 1:1000; Actin (2066, Sigma) 1:100) in TBST, followed by 3 x 5-minute TBST washes. Secondary antibody (goat anti rabbit HRP 1:2000) was applied for 1–2 hours at room temperature, followed by additional TBST washes (1-2 hours). Signal detection was performed using SuperSignal West Pico PLUS or Femto chemiluminescent substrates (ThermoFisher Scientific), and membranes were imaged using an iBright FL digital imaging system (ThermoFisher Scientific). Bands were quantified using iBright analysis software (ThermoFisher Scientific).

IF staining and imaging

Cells were grown on ibidi chamber slides and fixed using 4% PFA for 15mins at RT once desired confluency was reached (~70% confluent). After fixation, cells were washed 3x5mins with PBS and incubated in blocking buffer (2.5M $CaCl_2$, 1% TritionX-100, 1%BSA in PBS) for 30mins at RT. Primary antibodies (Supplemental Table 5.2) were applied diluted in blocking buffer for either 1hr at RT or overnight at 4C (Runx1 PA5-19638, 1:800).

cyCIF CRC patient matched tumor and peripheral blood

Antibody Generation

Antibodies were prepared using barcoding technologies as previously described[175–178]. In brief, each antibody was site-specifically conjugated to a 28-nucleotide (nt) single-stranded DNA docking strand (DS) using the SiteClick™ Antibody Azido Modification Kit (ThermoFisher Scientific), which targets the Fc region. For marker detection, 26 nt complementary imaging strands (IS) labeled with fluorophores and photocleavable linkers (PCLs) at both the 5′ and 3′ termini were hybridized to the docking strands. All oligonucleotides were synthesized by Integrated DNA

Technologies (IDT, Coralville, IA). A complete list of antibodies, oligo sequences, and labeling details is provided in Table S1.

Staining Protocol & Signal Removal

cyCIF was performed on formalin-fixed paraffin-embedded (FFPE) tumor tissue sections and corresponding peripheral blood mononuclear cells (PBMCs) from patient-matched specimens, as previously described[131,175–178]. FFPE tumor sections (5 μm) were first deparaffinized with xylene and rehydrated through a graded ethanol series. Antigen retrieval was conducted by incubation in citrate buffer (pH 6.0) for 30 minutes at 100 °C, followed by rinsing in heated deionized water and a 10-minute incubation in Tris-HCl buffer (pH 8.0) at 100 °C. Slides were then cooled to room temperature and washed in PBS.

PBMCs were isolated via Ficoll density centrifugation, adhered to poly-D-lysine–coated glass slides, and fixed in 4% paraformaldehyde (PFA) prior to cyCIF staining as described above. Cells were rehydrated in 3x5 min PBS washes. Tumor and PBMC slides were blocked for 30 minutes at room temperature in PBS containing 2% bovine serum albumin (BSA), 0.5% dextran sulfate, and 0.5 mg/mL sheared salmon sperm DNA. Ab-oligo conjugates targeting specific biomarkers (Supplemental Table 5.2) were diluted in blocking buffer and applied to the tissue and blood samples in a single-step staining protocol (separated into two rounds of primary antibody incubation (after R0 and after R5) as the volume of 34 antibodies left very little blocking buffer added to dilute to intended concentrations). Following incubation, unbound Ab-oligos were removed by washing, and samples were post-fixed with 2% PFA for 10 minutes. Samples were then incubated with imaging strands (IS) specific to the DNA docking sequences on the antibodies. Three to four ISs were applied per imaging round, each labeled with a distinct fluorophore, DAPI (Zeiss 96 HE), Alexa Fluor 488 (Zeiss 38 HE), AF555 (Zeiss 43 HE), AF647 (Zeiss 50), and AF750 (Chroma 49007 ET Cy7).

Image Acquisition

Whole slides were imaged using an AxioScan.Z1 digital slide scanner (Zeiss) equipped with a Colibri 7 light engine. Exposure times were adjusted per fluorophore and antibody to optimize signal detection and avoid saturation and were determined using negative staining controls for each round. Imaging was performed using a 20× Plan-Apochromat 0.8 NA objective and stitched using Zen Blue software (Zeiss).

Fluorescent Signal Removal

After each round of imaging, fluorescence signals were removed by UV treatment for 15 minutes to enable additional rounds of staining. The cycle of IS hybridization, imaging, and UV-based fluorophore cleavage was repeated until all desired markers were imaged. Round 0 (R0) images were acquired following DAPI staining and prior to antibody application to assess background autofluorescence using the 488 channel. Two tumor sections from different tumor regions and one peripheral blood specimen (~1.5 mL) were analyzed per patient.

Analysis

For Ab-oligo validation, Zen blue software (Carl Zeiss) was used for image visualization. For cyclic immunofluorescence, images from cyclic staining rounds were registered using ASHLAR feature-based image registration[179]. Cells were segmented using QiTissue default segmentation parameters for PBMC slides and using MESMER segmentation for tumor tissues[180]. The edges of the tissue sections and wells were excluded from analysis as well as areas with imaging artifacts (e.g., out of focus, bubbles). Using QiTissue, cells with high autofluorescence were removed from analysis based on the whole cell average fluorescence from the round 0 background channel. Hybrids were defined as co-expressing CD45 and at least one of the epithelial markers: panCK, ECAD, or EpCAM. Positive expression was defined in QiTissue as a whole cell intensity average value above the positive staining threshold that was determined by normalizing to an unstained control for each patient. Average cell intensity values were extracted for each antigen in the phenotyping panel for each identified hybrid cell and circulating tumor cell.

Extracted feature data of average cell intensity for all CHCs and CTCs were combined into a single data matrix for all patients and analyzed using Seurat[181]. The Seurat workflow was employed using SCTransform normalization (with regression on patient ID to correct batch effects), followed by PCA, UMAP, and Louvain clustering. Using the Seurat function FindMarkers(), differential biomarker expression was assessed across clusters and visualized using heatmaps. Marker expression levels for genes of interest (e.g., RUNX1, SPI1, THBS1) were visualized using FeaturePlot and VlnPlot functions, stratified by tumor stage and cell type. Subsets of the Seurat object were created to assess CHCs and CTCs separately. Expression differences by stage were evaluated using two-tailed t-tests comparing expression levels between pairs of stages (1 vs. 2, 2 vs. 3, and 1 vs. 3) for each biomarker.

Single cell omics in-vitro derived hybrid cells & CRC patient hybrid scRNA sequencing

Flow cytometry

All FACS used in these studies was performed on a BD Influx. For *in vitro* derived hybrid cells GFP and RFP fluorescence was used to isolate double positive hybrid cells. Cells were also stained for viability using DAPI. Tumor cell controls (RFP+) and macrophages (GFP+) were also sorted at the same time using the same methods and processed for scRNA and scATAC studies described below.

For patient samples, PBMCs were isolated using Ficoll-Paque density gradient centrifugation as previously described[106,131]. Briefly, Peripheral blood samples were collected from patients using heparinized vacutainer tubes (BD Biosciences, Franklin Lakes, NJ, USA) and diluted at a 1:2 ratio with phosphate-buffered saline (PBS; 1.37 M NaCl, 27 mM KCl, 0.1 M Na_2HPO_4, 18 mM KH_2PO_4, pH 7.4). 12 mL of Ficoll was underlaid beneath the diluted blood and centrifuged at 800 × g for 20 minutes at room temperature with no brake. The PBMC layer was collected, washed, and resuspended in PBS. Cells were then seeded onto poly-D-lysine–coated slides (1 mg/mL; Millipore, Burlington, MA, USA; Fisher Scientific, Waltham, MA, USA) and incubated at 37 °C for 15 minutes to promote adherence. Adherent cells were fixed with 4%

paraformaldehyde (PFA) for 5 minutes, permeabilized with 0.5% Triton X-100 (Fisher Scientific, BP151-100) for 10 minutes, and post-fixed with 4% PFA for an additional 10 minutes.

Patient tumor tissue obtained immediately after surgical removal was digested into single cell suspension by finely mincing, followed by enzymatic digestion with Liberase-TH (5 mg/mL in sterile ultrapure H2O) for 30-60mins on a stir plate at 300 RPM and 37C, then filtered using 100uM cell strainer, pelleted and washed. PBMCs and tumor cells were stained with antibodies targeting EpCAM, ECAD, CD45, and DAPI as previously described[108]. Briefly, cells were washed and resuspended in FACS buffer [phosphate-buffered saline (PBS), 1.0 mM EDTA, and 5% fetal bovine serum (FBS)]. Cells were incubated in FACS buffer containing Fc Receptor Binding Inhibitor (5ul per $1x10^6$ cells; eBioscience) for 20mins on ice. Cells were then incubated in FACS buffer for 30 min on ice with CD45-AF488 (1:100; Thermo Fisher Scientific), ECAD-AF647 (1:100), EpCAM-AF555 (1:100). Dapi was used to assess cell viability. Single color controls used for gating included EpCAM/Ecad expressing epithelial cell line, and donor PBMCs stained with only CD45-AF488 (1:100). Cells were defined as tumor hybrid cells or circulating hybrid cells by (EpCAM+ and/or ECAD+ and CD45+). Tumor cells (EpCAM+ and/or ECAD+ and CD45−) and immune cells (EpCAM- and ECAD- and CD45+) were also sorted into 16 wells of the 96 well plate containing cell lysis buffer (0.2% TritonX diluted in sterile, ultrapure H2O and RNase inhibitor (Takara Bio). Immediately after single cell sorting, plates were sealed, spun down and rapidly frozen and stored at -80C until further processing for Smart-seq described below. In vitro studies data reflect analyses from n = 2 hybrid fusion, sorting and downstream 10X sequencing experiments. For patient samples (Supplemental Table 5.1), n=3 patients with CRC (2 tumor samples, 3 PBMC samples) were processed for Smart-Seq plate based single cell sequencing, and if substantial hybrid cells obtained, also pooled with tumor cells or PBMCs for 10x genomics sequencing studies described in detail below.

Nuclei Isolation for 10x and s3 Libraries

After flow sorting, we spun the nuclei down (5 minutes, 500xg, 4C) to remove the media supernatant and resuspended it in 1 mL of NIB-Hepes buffer. We then incubated the cell suspension for 5 minutes on ice; and subsequently, we spun down the sample (5 minutes, 500xg, 4°C). We then resuspended the pellet in 1 mL NIB-Hepes, spun again, and resuspended in 1 mL once more before quantification.

10x Genomics scRNA

For the scRNA libraries we took the isolated nuclei and diluted them in 10x Genomics 20x Nuclei Buffer to the desirable targeted cell recovery. We then proceeded with the 10x Genomics Single Cell 3' Gene Expression Kit according to their published protocol.

SmartSeq scRNA

For these patient libraries, the nuclei were frozen down in 96 well plate in lysis buffer as previously described in FACS section above. We then followed the published Takara's SMART-Seq mRNA Single Cell LP User Manual starting from Step V, First-Strand cDNA Synthesis. The libraries were indexed using the Unique Dual Index Kits, pooled, cleaned, and quantified as described below.

ScaleBio Tagmentation for scATAC

To generate these scATAC libraries we leveraged the s3-ATAC protocol on the ICell8[84]. We put the ICell8 instrument through standard instrument start up including tip cleaning, stream checks, wash priming, and prechilled the block. Next, we assigned the source plate set up to (Cells – PCR Mix1 – Index 1 (i7 TruSeq) – PCR Mix2 – Index 2 (i5 Nextera) and read in the barcode of the nanochip.

The 100 μM ICell8 TruSeq Primer source plates was prepared. After removing the seal from nanochip and vacuum sealing it into position, the ICell8 first dispensed the i7 and then i5 primers into the chip. Following each dispense the chip was blotted, sealed, and centrifuged (3 minutes, maximum RFC, 4°C) before the next step. During the centrifugation the ICell8 tips were cleaned twice.

We prepared our quantified, isolated nuclei getting them to the appropriate concentration for tagmentation. We made a master mix of the following (2.67 µL NIB-H, 3 µL 4.3478x TAPs-TD (1M TAPs, 5M KOAc, 1M MgOAc, 1M D-glucosamine, 6.1% DMF, in H_2O), 0.1 µL pluronic). This was dispensed into a 96-well plate with a multichannel, and to this we stamped in 5 µL of 500 µM Scale SBS12/A14 TN5 on the Bravo. We next manually pipetted in 6,000 nuclei in 4.23 µL NIB-H to the respective wells. We proceeded to tagment at 55°C for 15 minutes. After tagmentation we allowed the samples to cool on ice, before they were pooled, spun down (5 minutes, 500xg, 4°C, with 180° turn) and resuspended in 1 mL TMG Buffer (1.25 mL 4x TAPS Premix, 1.5 mL 50% glycerol, 2.2 mL H_2O, 50 µL 10% pluronic F-127) to cushion the nuclei and aid in recovery. These TMG washes were repeated three more times, with the final resuspension being in 120 µL ICell8 Loading Buffer (44 µL 4.3478x TAPs-TD, 50 µL TN5 Dialysis Buffer (10 mM Hepes 7.2, 20 mM NaCl, 0.02% TritonX, 2.5% glycerol, 0.2mM DTT), 0.1% 10% pluronic f-127, 401 µL H_2O). Nuclei were quantified, with a typical recovery at approximately 50% post-tagmentation, and diluted to get 358 nuclei/µL.

Now ready to load our tagmented nuclei to the ICell8 chip. Approximately 35 µL of nuclei were dispensed into each well of the source plate, with 100 µL of cell loading buffer in the control wells. After using the ICell8 to dispense the cells into the chip, it was spun down (10 minutes, maximum RFC, 4°C) while 3 additional tip cleans were run on the ICell8. The chip was incubated at 53.7°C for 15 minutes. After adding 100 µL of PCR master mix (188.4 µL 5x hi GC Buffer, 18.84 µL dNTPs, 14.58 µL Kapa HiFi Polymerase (non-hot start), 316.5 µL H_2O) to the source plate wells, 150 nL was dispensed into each well of the ICell8 chip. Again, the chip was spun down (10 minutes, maximum RFC, 4°C) while 3 tip cleans were run concurrently. Following this, 12 cycles of sciATAC-P PCR were run with the samples on the ICell8 chip. The library was collected utilizing the ICell8 Collection Kit (Cat. No. 640212) and clean with a standard double sided-SPRI bead protocol prior to quantification.

Artificial Doublet Experiment

To generate and assess artificial doublet cell transcriptomes for comparison with hybrid fusions, we utilized Takara's SMART-Seq Pro application on the ICell8 single-cell system (Takara Cat. No. 640257). Our experimental design leveraged the previously established distinct fluorescent markers: macrophages expressed GFP, tumors expressed RFP, and hybrid cells co-expressed both fluorophores. We followed the SMART-Seq Pro manual closely, with several key modifications. The ICell8 system dispenses cells according to a Poisson distribution, which can result in imperfect single-cell loading. To increase throughput, the standard protocol involves two consecutive cell dispenses, each followed by imaging to identify wells containing single cells. These wells are then selected for downstream reagent addition. In our modified approach, we used two separate 384-well source plates. For source plate preparation, in vitro-derived hybrid cells, cultured MC38 tumor cells, and first-passage macrophages were isolated via flow cytometry as previously described. Cells were stained with DAPI (10 μg/mL) in 2 mL of media on ice and quantified. Two 384-well source plates were prepared: one containing macrophages and hybrids, and the second containing MC38 cells (loaded into wells A1 to D2). During the first dispense, we loaded macrophages into ¾ of the chip and hybrid cells into the remaining ¼. After imaging, we manually edited the "Emptywells" filter file in the Experiments Folder, changing all 0s to 1s for wells that received macrophages. Further, we set all hybrid cell wells to 1 to prevent further dispensing into those wells. Next, we swapped in the second source plate containing MC38 tumor cells for the second dispense. These cells were only dispensed into the ¾ of the chip that had previously received macrophages, creating artificial doublets. A second round of imaging followed, and the chip was frozen afterward. To identify cell types in each well, we used CELLSELECT's training feature to classify wells based on DAPI, GFP, and RFP signals. While high-confidence calls (confidence score ≥ 0.80) were accepted, wells with lower scores were manually reviewed and annotated using predefined criteria due to variability in fluorescence intensity. All subsequent steps—including first-strand cDNA synthesis, cDNA amplification, tagmentation, indexing, and

library purification—were performed according to the Takara SMART-Seq Pro manual. Libraries were quantified and sequenced as described in subsequent sections.

Table 2: Artificial Doublet Categorization

Scan 1		Scan 2		Annotation
DAPI 0	GFP 0	DAPI 0	RFP 0	Remove
DAPI 1	GFP 0	DAPI 0	RFP 0	Macrophage
DAPI 0	GFP 1	DAPI 0	RFP 0	Macrophage
DAPI 1	GFP 1	DAPI 0	RFP 0	Macrophage
DAPI 0	GFP 0	DAPI 1	RFP 1	Tumor
DAPI 0	GFP 0	DAPI 0	RFP 1	Tumor
DAPI 0	GFP 0	DAPI 1	RFP 0	Tumor
DAPI 0	GFP 0	DAPI 2	RFP 0	Tumor Doublet
DAPI 0	GFP 0	DAPI 2	RFP 1	Tumor Doublet
DAPI 0	GFP 0	DAPI 2	RFP 2	Tumor Doublet
DAPI 1	GFP 0	DAPI 1	RFP 1	Mac-Tum Doublet
DAPI 1	GFP 0	DAPI 0	RFP 1	Mac-Tum Doublet
DAPI 1	GFP 0	DAPI 1	RFP 0	Macrophage
DAPI 1	GFP 0	DAPI 2	RFP 0	Mac-Tum Doublet
DAPI 1	GFP 0	DAPI 2	RFP 1	Mac-Tum Doublet
DAPI 1	GFP 0	DAPI 2	RFP 2	REMOVED
DAPI 0	GFP 1	DAPI 1	RFP 1	Mac-Tum Doublet
DAPI 0	GFP 1	DAPI 0	RFP 1	REMOVED
DAPI 0	GFP 1	DAPI 1	RFP 0	Mac-Tum Doublet
DAPI 0	GFP 1	DAPI 2	RFP 0	REMOVED
DAPI 0	GFP 1	DAPI 2	RFP 1	REMOVED
DAPI 0	GFP 1	DAPI 2	RFP 2	REMOVED
DAPI 1	GFP 1	DAPI 1	RFP 1	REMOVED
DAPI 1	GFP 1	DAPI 0	RFP 1	REMOVED
DAPI 1	GFP 1	DAPI 1	RFP 0	Macrophage
DAPI 1	GFP 1	DAPI 2	RFP 0	Mac-Tum Doublet
DAPI 1	GFP 1	DAPI 2	RFP 1	Mac-Tum Doublet
DAPI 1	GFP 1	DAPI 2	RFP 2	Mac-Tum Doublet

Library Quantification

We first benchmarked all libraries using the Qubit dsDNA High Sensitivity assay (Thermo Fisher Q32851). For more concentrated libraries we diluted down to approximately 2ng/µL. Subsequently, the molarity of the DNA was confirmed via the Agilent Tapestation 4150 D500 tape (Agilent 5067-5592).

Library Sequencing Parameters

We sequenced all library preparations using standard chemistry and following standard protocols on the Illumina NextSeq2000 for 650pm standard flow cells and 488pm for X-LEAP flow cells.

The in vitro 10x scRNA library was sequenced was sequenced as paired-end with 28 cycles for read 1, 90 cycles for read 2, 10 cycles for index 1, and 10 cycles for index 2 utilizing a P2-200 X-LEAP flow cell (Illumina Inc. 20100986).

The Artificial Doublet library was sequenced as paired-end with 75 cycles for read 1, 75 cycles for read 2, 8 cycles for index 1, and 8 cycles for index 2 utilizing P2-200 (Illumina Inc. 20046812).

The 10x scRNA patient sample libraries were sequenced as paired-end with 28 cycles for read 1, 90 cycles for read 2, 10 cycles for index 1, and 10 cycles for index 2 utilizing P2-100 ((Illumina Inc. 20046811).

The SmartSeq scRNA patient libraries were sequenced as paired-end with 100 cycles for read 1, 100 cycles for read 2, 10 cycles for index 1, and 10 cycles for index 2 utilizing P2-200 and P3-200 X-LEAP flow cells (Illumina Inc. 20100989).

ScaleBio tagmented scATAC libraries were sequenced as paired-end with 85 cycles for read 1, 125 cycles for read 2, 10 cycles for index 1, and 10 cycles for index 2.

Computational Analysis

Primary Sequence Data Processing for 10x Genomics scRNA Libraries

Raw sequence files were processed using CellRanger[182] count and aligned to either GRCh38 or MM10 to generate matrix, features, and barcodes files for downstream processing in Seurat.

Primary Sequence Data Processing for SmartSeq scRNA Libraries

Sequence reads were demultiplexed using unidex (https://github.com/adeylab/unidex) which matches index barcodes to a whitelist. STAR (https://github.com/alexdobin/STAR.git) was utilized to align demultiplexed fastq files to GRCh38. Scitools rmdup (https://github.com/adeylab/scitools) was utilized to remove duplicate reads. The subread (https://github.com/ShiLab-Bioinformatics/subread.git) module was utilized to perform annotations against GRCh38, and the subsequent feature counts were converted to a count matrix using a custom perl script (git@github.com:adeylab/ICell8_SmartSeq_forKQ.git). This matrix was suitable for downstream analysis in Seurat.

Analysis of scRNA Data

Matrices were read into Seurat and used to generate Seurat objects[181,183]. The data was normalized, variable features were determined and scaled, and dimensions were reduced via PCR using default parameters. Native functions for finding nearest neighbors, clustering and UMAP projection were employed using standard parameters.

After metadata annotation, like experiments were merged into combined Seurat objects, and re-run through prior functions. Elbow plots were utilized to determine reasonable cut-offs for PCA dimensions, typically 10. Expression of various genes of interest were visualized across clusters using feature plot functions and GO term enrichment analysis was conducted using clusterprofiler package.[88]

Primary Sequence Data Processing for scATAC Library

Unidex mode ICell8_scale24 was employed for demultiplexing based on the two 10bp indices, matched to respective index files. Additionally, the first 8bp of read 2 served as the tagmentation index for the ScaleBio indexed Tn5. These 8 bp of Tn5 index were trimmed together with the next 20 bp of mosaic end sequence. The demultiplexed reads were then aligned to the

human reference genome hg38 using bwa-mem (v0.7.15-r1140), processed with scitools rmdup, and then filtered down to cell barcodes reaching a minimum unique read count.

Analysis of scATAC Data

Aligned, duplicate-filtered bam files were imported into ArchR to generate an arrow file, with sample names added via a separate text file, and compiled into an ArchR project[79]. Annotations were subsequently imported from a CSV and added as a meta data column in the ArchR project. Iterative LSI, UMAP projections, TSNE projections were performed using default parameters. Track plots were generated for genes of interest and genes in associated pathways. MACS2 was employed for peak calling and identifying marker peaks, which were used for peak enrichment analysis in a pair-wise fashion.

Statistical Analysis

All statistical analyses were performed using GraphPad Prism v9 or R (version [insert version number if needed]). Statistical tests were selected based on data distribution, sample size, and experimental design as follows: One-way ANOVA with Dunnett's post hoc test was used for multiple group comparisons with a shared control (tumor vs. each hybrid cell line) in normally distributed datasets (Fig. 2D, Fig. 2G, Fig. 4E). For non-normally distributed data, a Kruskal-Wallis test followed by pairwise Wilcoxon rank-sum tests was used for group comparisons (tumor vs. each hybrid cell line), with Benjamini-Hochberg false discovery rate (FDR) correction applied for multiple testing (Fig. 4F, Fig. 4H, Fig. 5E, Fig. 6G). Unpaired two-tailed Student's t-tests were used for comparisons between two groups in normally distributed data (Fig. 2H, Fig. 5I). Pairwise Wilcoxon rank-sum tests were performed for selected nonparametric comparisons, with Cliff's delta used to quantify effect sizes (Fig. 4G, Fig. 7H). A generalized linear mixed model (GLMM) with a beta distribution was employed for proportional data (Fig. 6A). Results are reported as mean ± standard deviation (SD) unless otherwise indicated. All violin plots, UMAPs, and dot plots were generated using the ggplot2 and Seurat packages in R. Statistical significance was defined as $p < 0.05$.

Additional Figures

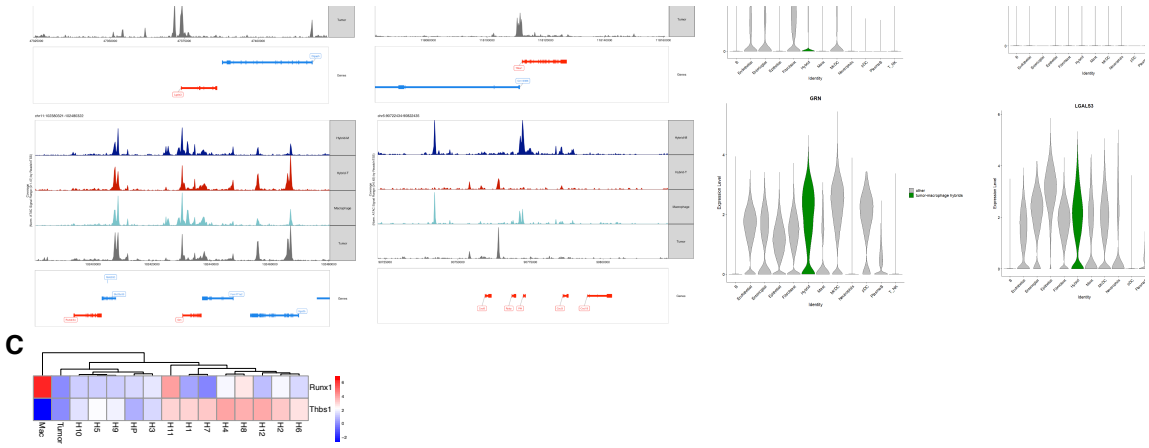


Figure 18: Additional RUNX1 pathway analyses.
A) ATAC-seq track plots, B) gene expression in CRC human hybrids, and C) *Runx1* and *Thbs1* expression in hybrid single cell clones.

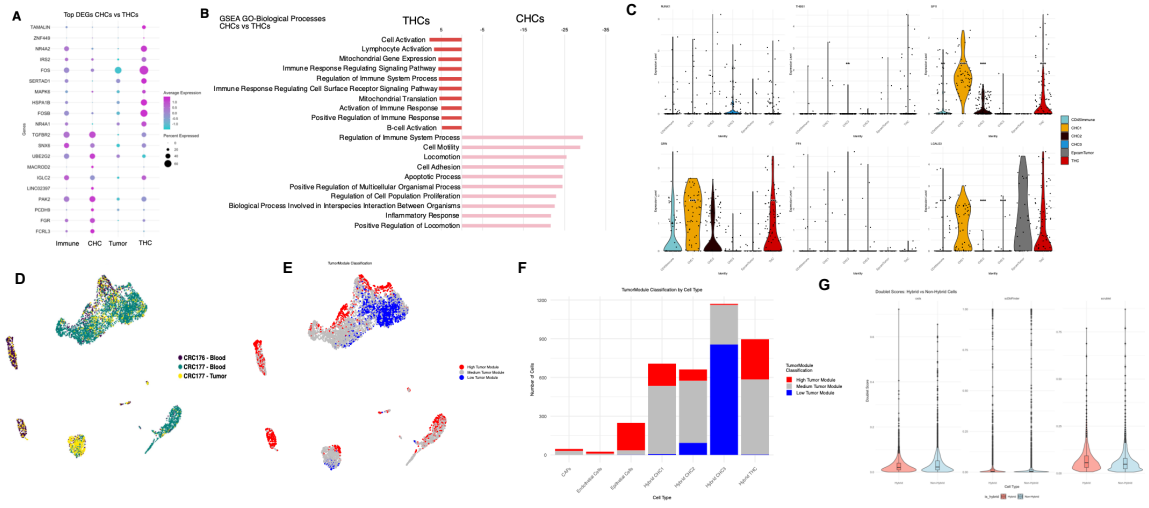


Figure 19: Additional information from patient matched CHC and THC scRNA-sequencing studies.
A) Dot plot showing top differentially expressed genes when comparing aggregated CHCs and THCs. B) GSEA GO-biological processes comparing aggregated CHCs and THCs. C) Violin plots illustrating expression of RUNX1 and downstream transcription factors in CHC vs. THC. D) 10x UMAP colored by patient ID and biopsy site. E) 10x UMAP colored by Tumor Module Classification. F) Bar Plot for disaggregated CHCs, THCs, etc. colored by Tumor Module. G) Violin plot of doublet scores in hybrid versus non-hybrid cells

Addendum 1: Clonal Tracing of Follicular Lymphoma in Conjunction with scWGS/sciMET Analysis

Authors collaborating in this work and affiliations

Konstantin Queitsch[1], Mattew Stern[3], Ruth V. Nichols[1], Brendan L. O'Connell[1,2,3], Tanya Shree[3,4], Andrew C. Adey[1,2,3,5] *

1. Department of Molecular & Medical Genetics, Oregon Health & Science University, Portland, OR, USA
2. Cancer Early Detection Advanced Research Center, Oregon Health & Science University, Portland, OR, USA
3. Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA
4. Division of Hematology/Medical Oncology, Oregon Health & Science University, Portland, OR, USA
5. Knight Cardiovascular Institute, Oregon Health & Science University, Portland, OR, USA

* Lead Contact for correspondence: adey@ohsu.edu

Author contributions

Conceptualization: A.C.A, T.S., Methodology: K.Q., M.S., R.V.N., T.S., A.C.A., Formal analysis: K.Q., M.S., Data curation: K.Q., M.S.

Abstract

Follicular lymphoma (FL) exhibits significant heterogeneity, which is thought to contribute to variable treatment responses and disease progression. Traditional bulk sequencing approaches fail to capture the subclonal heterogeneity and evolutionary dynamics within individual FL tumors, while existing single-cell methods cannot simultaneously track both genomic alterations and clonal lineage relationships. Work is ongoing on a novel coassay that combines single-cell whole-genome sequencing (scWGS) with V(D)J region capture to enable comprehensive genomic profiling of FL B-cell populations while maintaining clonal identity through unique, endogenous BCR sequences. The approach utilizes in situ reverse transcription to capture V(D)J regions prior to nucleosome disruption and genome-wide tagmentation. Combinatorial indexing then allows both datasets to be attributed to individual cells. This coassay will provide a powerful tool for dissecting FL heterogeneity, tracking clonal evolution, and potentially identifying subclonal populations with differential therapeutic responsiveness in longitudinal studies.

Main Text

The development of a single-cell coassay capable of simultaneously capturing genome-wide copy number alterations and clonal V(D)J sequences required accounting for distinct technical requirements for each molecular component. Adapting the established scWGS method from Chapter 1 for B-cell lymphomas proved straightforward. Integrating V(D)J capture necessitated extensive optimization to overcome issues with primer specificity, amplification artifacts, and compatibility with the tagmentation-based genomic workflow. This section details the iterative development process, from initial proof-of-concept experiments through the resolution of key technical obstacles, culminating in a novel approach that enables paired genomic and clonal lineage profiling at single-cell resolution.

Method development utilized lymphoblast cell lines (JVM2 and DoHH2), healthy donor B-cells, and CLL lines. The intended application targets B-cells isolated from cryopreserved blood draws from FL patients. Currently, six patient samples are banked, with additional samples anticipated by the Shree lab.

Proof of concept experiments on CLL samples indicated that there would be no issues adapting Chapter 1's scWGS method – previously run on PDAC-derived cell lines, murine cell lines, and murine and human brain tissue – for B-cell lymphomas. Standard conditions were employed for nuclei isolation, fixation, nucleosome depletion, tagmentation, and cell recovery[184]. The library complexity metrics met thresholds of prior assays (~2.5 – 3e5+ unique molecules per cell) and the TSSe (~0.8) indicated that the nucleosome disruption was uniformly successful. Experiments demonstrated that the resulting scWGS library was suitable material for the Twist Biosciences Exome Capture Kit. Although it only achieved 15 to 20-fold enrichment rather than the 30-fold enrichment promised, exome capture opened the possibility for addressing the coverage limitations that might have precluded the detection of novel SNVs in the scWGS data. This initial library served as a proof-of-concept vehicle for establishing a computational pipeline including GATK[60], CellSNP-lite[61], and VCFR[62], as described in Part III of the Introduction. While the scWGS

component showed promising results, developing the V(D)J capture component presented several technical challenges.

The development of the V(D)J component of the coassay presented complications. The initial concept was to integrate published gDNA V(D)J primers[185,186], and spike them into 10x droplets alongside standard 10x chemistry to enrich for the region of interest. All fragments, tagmented and primer amplified, within the droplet would share 10x GEM barcodes for cell identification. The published gDNA V(D)J primers proved unreliable. Exhaustive testing pointed to a lack of specific amplification, a propensity to self-amplification, and the generation of primer-dimer artifacts. Amplifying the V(D)J genomic region is complex due to somatic hypermutation affecting the potential primer binding site specificity and the region containing highly homologous stretches. Given these limitations with genomic DNA primers, an alternative approach was pursued.

The coassay was reimagined to target nuclear RNA for the V(D)J with primers for the cDNA template[185]. Using RNA starting material enables priming off the L1 leader region and constant region which are less affected by somatic hypermutation and are downstream of V(D)J rearrangement. The requirement for reverse transcription complicated the additive droplet-based chemistry. Ultimately, the approach shifted away from the 10x scWGS Chromium platform to combinatorial indexing in plate-based s3WGS[33]. Rethinking the approach also opened the door to integrating sciMET in place of scWGS in the future; as both the tagmentation, ligation, and indexing steps are compatible between the two assays with minor adjustments.

The shift to cDNA primers (Table 3) showed immediate promise, yielding products of the appropriate size when assessed on the Agilent TapeStation. The cDNA products were validated via amplicon sequencing. With this validation complete, the next phase focused on adapting these primers for integration into the coassay workflow. The cDNA primer sequences served as the foundation for designing modified IGHC primers (Fig. 2), which incorporated several functional elements: a

Table 3: Langlois et. al[185] cDNA Primer Design	
IGHL1_1	CTCACCATGGACTGSAYYTGGAG
IGHL1_2	ATGGACAYACTTTGYTMCACRCTCC
IGHL1_3	ATGGARTTKGGGCTKWGCTGGGTTT
IGHL1_4	CTGTGGTTCTTYCTBCTSCTGGTGG
IGHL1_5	CCTCCTCCTRGCTRTTCTCCAAG
IGHL1_6	CTGTCTCCTTCCTCATCTTCCTGCC
IGHC_mu	GGTTGGGGCGGATGCACT
IGHC_gamma	CGATGGGCCCTTGGTGGA

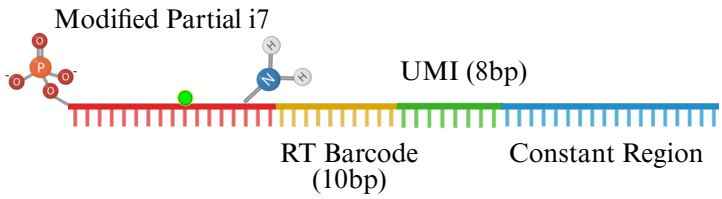


Figure 20: Graphical Depiction of Modified RT Primers for V(D)J Seq Coassay.
The amine modification on the partial i7 sequence allows for crosslinking of the targeted RNA RT product. The 5' phosphate group is necessary for the splint ligation chemistry. The biotin modification is required for the streptavidin bead pulldown to separate the coassays.

primary barcode for cell identification, a UMI for molecule counting, and a modified partial i7 adapter sequence. These modifications served specific purposes—the partial adapter enabled additional barcoding through subsequent ligation chemistry, the amine modification allowed for crosslinking during fixation, and the biotin modification facilitated streptavidin-based separation of the V(D)J and genomic libraries. The IGHL primers were modified to include an i5 adapter to enable subsequent indexing with TruSeq chemistry and the addition of Illumina flow cell primers on the 3' end. Following iterative optimization of RNA reverse transcription, temperature conditions, and cycling durations consistent generation of cDNA with the modified IGHC primers was achieved.

The primers were ready to integrate into the targeted RNA prong of the coassay workflow (Fig. 21). An initial promising pilot experiment in intact nuclei from CLL, JVM2, and DoHH2 cell lines produced clean TapeStation readouts with a sharp peak at ~550bp, the expected product

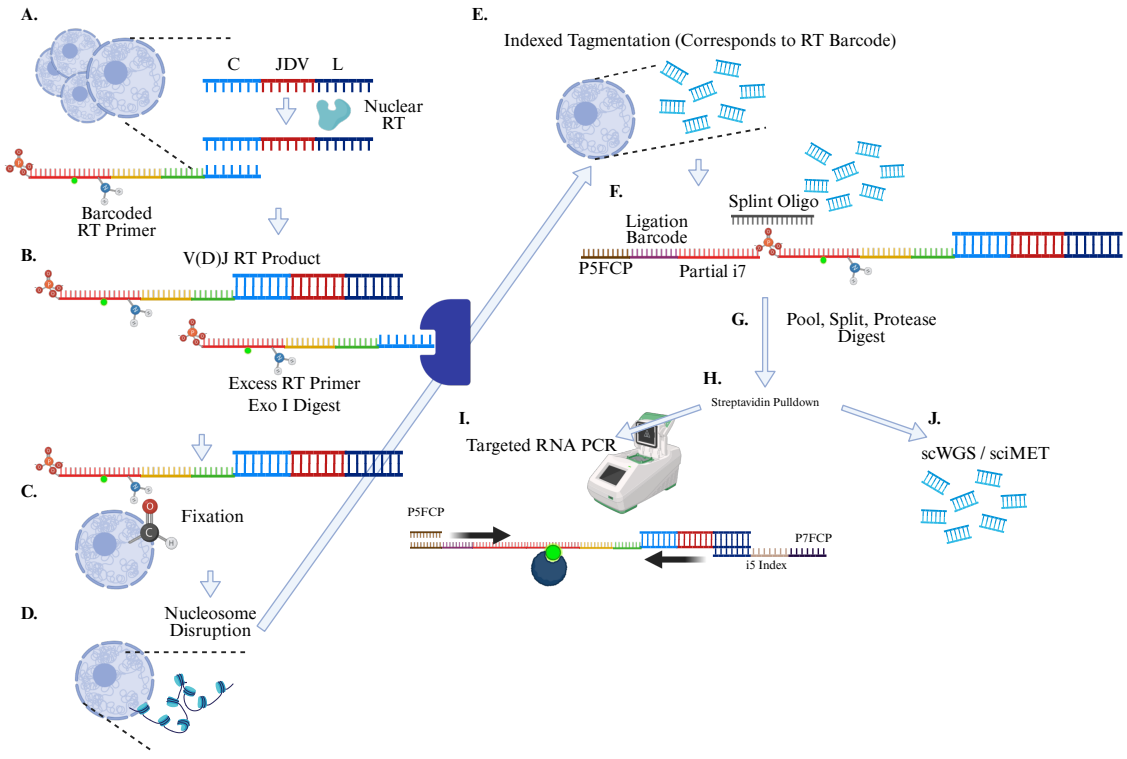


Figure 21: Graphical Abstract of V(D)J Seq Coassay.

size. However, sequencing revealed an abundance of concatemer artifacts. Further experiments on isolated RNA from JVM2 and DoHH2, while reducing the cycles, suggested the crux of the issue was remnant IGHC RT primer (Fig. 21A) priming during the subsequent targeted RNA PCR (Fig. 21I). The issue was ameliorated with the introduction of an exonuclease I digest prior to the fixation (Fig. 21C, Fig22A-B). Exonuclease I preferentially cleaves ssDNA from the 3' end, leaving the double-stranded RT product protected while chewing up the single-stranded primer. A series of experiments demonstrated a significant reduction in the amount of observed artifact (90 – 200bp) with the addition of the exonuclease, in both isolated RNA and intact nuclei. The most recent round of the targeted RNA arm of the V(D)J coassay in nuclei indicated that the addition of the exonuclease vastly improves the issue. The immediate next steps will center on optimization of the post-PCR clean up conditions, as the desired versus undesired product is still not where it needs to be for sequencing (Fig. 22D). Should that prove successful both arms of the V(D)J-seq coassay would be ready for implementation.

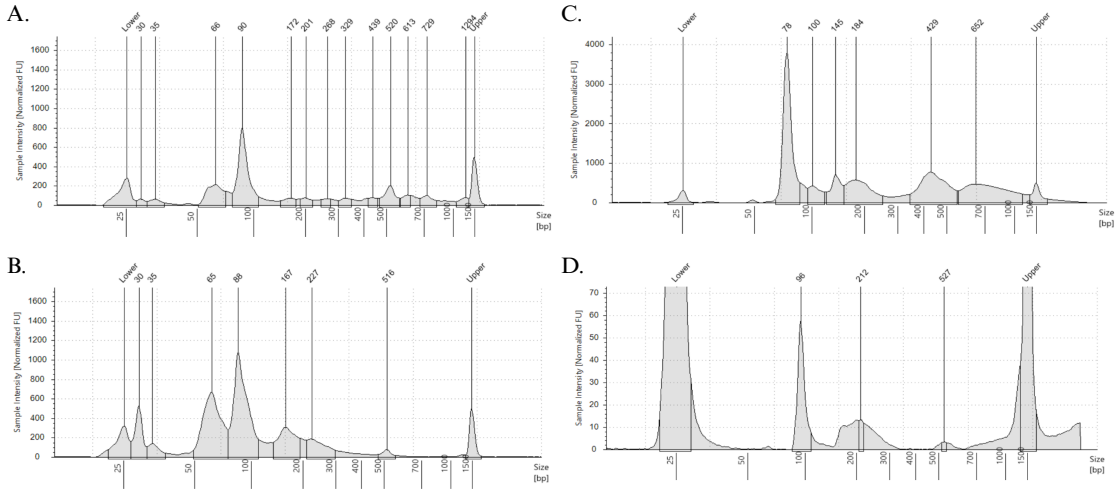


Figure 22: Preliminary QC Data from Aligent TapeStation.
A. Starting with isolated RNA from JVM2 cells, introduction of exo I treatment resulted in desired peak at ~520bp post-targeted PCR. **B.** Without exo I digestion, additional artifact is observed at ~160-220bp and a reduced peak at ~520bp relative to **A. C.** Pre-SPRI-bead clean up a (shifted) TapeStation trace shows an estimated desired peak, but considerable artifact at ~430bp and ~160bp. **D.** Starting with intact nuclei and in situ RT, the desired peak at ~520bp is observed; however post clean up considerable ~220bp fragments remain.

In theory, following the digestion, crosslinking, and nucleosome disruption steps the nuclei are then prepared for the uniform tagmentation required for either scWGS and sciMET assays, with an indexed s3 Tn5 (Fig. 21E-J). Each Tn5 index corresponds to a specific barcoded RT primer so the reads may be associated with the corresponding V(D)J sequence. Following tagmentation, an i7 ligation is utilized to ligate an additional layer of indexing on the 5' end of the fragments as well as add an i7 flow cell adapter required for Illumina sequencing. Nuclei are pooled and protease digested. Targeted RT product, which is biotinylated, and scWGS tagmented fragments are separated via a streptavidin bead pulldown. The targeted RT product is then amplified on the streptavidin beads, using the i7 flow cell primer and indexed primers complementary to the flanking leader sequence on the 3' end of the V(D)J fragments. In parallel, the scWGS fragments in the supernatant undergo linear extension and PCR amplification with corresponding indexed i5 primers. Assuming 8 initial paired indexes (RT/Tn5), 96 ligation indexes, and 96 i5 indexes, a 96-well coassay allows for the processing of an approximately 9,000 nuclei. While optimization of the coassay remains ongoing, significant progress has been made in addressing key technical challenges. The successful adaptation of scWGS for B-cell lymphomas, combined with the resolution of primer-dimer artifacts through exonuclease I treatment, demonstrates the feasibility of this approach. Future work will focus on finalizing post-PCR cleanup conditions and validating the complete workflow on primary FL samples.

To our knowledge, this represents the first attempt to simultaneously capture both genome-wide copy number alterations and clonal V(D)J sequences at single-cell resolution in lymphoma samples. Once fully optimized, this coassay will enable unprecedented insights into FL heterogeneity by linking genomic alterations directly to clonal lineages defined by unique BCR sequences. This capability will be particularly valuable for longitudinal studies, where tracking the expansion or contraction of specific subclones in response to therapy could reveal mechanisms of treatment resistance and relapse. Moreover, by maintaining clonal identity while profiling genome-wide alterations, this approach may identify previously hidden subclonal populations with distinct

genomic features that confer differential therapeutic responsiveness, ultimately informing more personalized treatment strategies for FL patients.

Table 4: Addendum 1 Materials		
Reagent / Resource	**Source**	**Identifier**
Chemicals, Peptides, and Recombinant Proteins		
Hepes, pH 7.5	Sigma-Aldrich	H4034
$MgCl_2$	Sigma-Aldrich	M8226
NaCl	Fisher Scientific	M-11624
IGEPAL	Sigma-Aldrich	I8896
Tween-20	Sigma-Aldrich	P7949
Formaldehyde	Fisher Scientific	PI28906
Glycine	Sigma-Aldrich	G8898-500G
UltraPure Sodium Dodecyl Sulfate	Invitrogen	15525-017
D-(+)-Glucosamine hydrochloride	Sigma-Aldrich	G1414-100G
SEQURNA	SEQURNA	Cat. SQ00203
Maxima H-minus Reverse Transcriptase	Thermo Fisher	Cat. EP0752
T4 Polynucleotide Kinase (PNK)	New England Biolabs	Cat. M0201L
T7 DNA Ligase	New England Biolabs	Cat. M0318L
RNaseOUT™ Recombinant Ribonuclease Inhibitor	Thermo Fisher	Cat. 10777019
Qiagen Protease	Qiagen	
KAPA HiFi HotStart Ready Mix	Roche	
VersaSeq 2.0 Polymerase		
Watchmaker U 2x Master Mix		
Critical Commercial Assays		
Chromium Next GEM Single Cell ATAC Kit v2	10x Genomics	PN-1000390
Next GEM Chip H Single Cell Kit	10x Genomics	PN-1000162
Single Index Kit N, Set A, 96	10x Genomics	PN-1000212
Single Cell ATAC Gel Beads v2	10x Genomics	PN-2000210
scATAC Pre-Indexing Kit	ScaleBio	N/A
Qubit 1x dsDNA HS Assay Kit	Invitrogen	Q33231
High Sensitivity D1000	Agilent	5067-5584

ScreenTape		
High Sensitivity D1000 Sample Buffer	Agilent	5067-5603
NextSeq2000 Kits	Illumina	20046811, 20046812
Experimental Models: Cell Lines		
CLL	Shree Lab	N/A
JVM2	Shree Lab	N/A
DoHH2	Shree Lab	N/A
Patient-Isolated B-Cells	Shree Lab	N/A
Donor-Isolated B-Cells	Shree Lab	N/A
Software		
unidex	Adey Lab	https://github.com/adeylab/unidex
bwa-mem (v0.7.15-r1140)	Li H, Durbin R[82]	https://github.com/lh3/bwa
scitools	Adey Lab	https://github.com/adeylab/scitools
Change-O	Gurkan Lab	https://github.com/immcantation/changeo
GATK	Granja JM, *Corces MR* et al.[79]	https://gatk.broadinstitute.org/hc/en-us

V(D)J-seq Coassay

Nuclei Isolation

To isolate the nuclei, sorted cells were resuspended in 2 mL 1 mL of NIB-Hepes buffer (10 mM Hepes, pH 7.5, 3 mM $MgCl_2$, 10 mM NaCl, 0.1% IGEPAL (v/v), 0.1% Tween-20 (v/v)). We then incubated the cell suspension for 5 minutes on ice; and subsequently, we spun down the sample (5 minutes, 500xg, 4C). We then resuspended the pellet in 1 mL NIB-Hepes, spun again, and resuspended in 1 mL once more before quantification. Nuclei were quantified.

Reverse Transcription of Targeted V(D)J RNA

Approximately 280,000 nuclei were dispensed into each well of an 8-strip together with 2.8 µL 1M glucosamine, 2.8 µL 100 µM modified IGHC RT primer, adding NIB-H for a total volume of 70 µL. The 8-strip was incubated for 5 minutes at 55C, while mixing at 400 rpm. Immediately afterwards it was placed on ice. During the incubation, the RT conditions were set on the thermocycler as the following (4C Hold, 4C 2 min, 10C 2min, 20C 2min, 30C 2min, 40C 2min, 50C 2min, 55C 2min, 4C Hold). One ice, RT mix was added to each well (56 µL 5x RT Buffer, 1.4 µL 1M DTT, 14 µL 10 mM dNTPs, 1.4 µL RNaseOUT RNase Inhibitor, 14 µL Maxima H-minus reverse transcriptase, 44.8 µL dH_2O). The reaction was mixed via pipetting on ice, and the

200 µL reactions were split across 4 wells each for more even heating and then moved onto the pre-set thermocycler.

Exonuclease I Digestion

Like samples were pooled. The exonuclease digestion was straight forward, simply adding 1 µL per 10 µL of RT reaction. This was incubated for 15 minutes at 37C. Heat inactivation was skipped, and samples proceeded directly into crosslinking.

Fixation & Nucleosome Disruption

We added 780 µL NIHB-H and 46.9 µL 16% formaldehyde (final concentration is ~0.75% formaldehyde) to each reaction. After pipette mixing, we fixed the sample at room temperature over 10 minutes. We next added 46.9 µL 2.5M glycine to quench the reaction, incubated for 5 minutes on ice, added 3 µL of BSA, and then spun the suspension down (5 minutes, 500xg, 4°C). We next resuspended the sample in 970 µL NIB-H and then added 30 µL 10% SDS. We incubated the nuclei for 20 minutes at 37°C in this solution. We carefully spun down the samples (5 minutes, 500xg, 10°C), as the SDS can precipitate and taint the pellet if left cold for too long. Nuclei were then resuspended in 1 mL NIB-H buffer and quantified.

Tagmentation & Wash

With a targeted 13 µL remaining in each tube after the SDS supernatant is discard, we added 5 µL of 4x TAPs-TD and 2 µL indexed s3-Tn5 to each tube. The Tn5 index is paired with a specific RT barcode index. Tagmentation is completed with an incubation at 55C for 15 minutes. After incubation, the tubes are pooled and washed in 3 mL cold NIB-H with 15 µL BSA. This is spun down (5 minutes, 500xg, 4°C), washed again in 3 mL cold NIB-H with 15 µL BSA, and spun down again. After the supernatant is removed, the pellet is resuspended in 100 µL NIB-H and transferred to a 1.5 mL Eppendorf tube.

i7 Ligation

To this tube we added the following: 33 µL 10x PNK Buffer, 33 µL 10 mM ATP, 22 µL dH_2O, and 132 µL T4 PNK. This was incubated at 37C for 30 minutes and then put on ice. On ice we added the following: 677 µL 2x StickTogetherBuffer, 33.8 µL 100 µM splint oligo, and 169 µL

T7 DNA ligase (3000 U/μL). 11 μL of the mix was dispensed to each well of a 96-well plate. Subsequently, 2 μL of 15 μM ligation barcode was stamped into the plate. This was incubated overnight at 15C, shaking at 400 rpm.

Post-Ligation and Dilution

Nuclei in the plate were pooled into a 5 mL tube and topped off with 3 mL NIB-H and with 3 μL BSA. After a spin (5 minutes, 500xg, 4°C) they were resuspended in 100 μL NIB-H, prepared separately without protease inhibitor. This is a crucial consideration for the subsequent protease digestion. 96-well plates are prepared with 1 μL Qiagen Protease and 1 μL 90 mM Tris-HCl. Nuclei are quantified and diluted to 60 – 90 nuclei per μL, and 1 μL is added to the plate. Plates can be frozen down at this step.

C1 Streptavidin Bead Pulldown

If frozen, plates are spun down briefly. Nuclei are then digested in a 55C, 15-minute incubation step, followed by a 72C, 20-minute incubation step to denature the protease. Concurrent to the incubation 196 μL of C1 Streptavidin beads are washed twice in 500 μL 2x BW Buffer (10 mM Tris-HCl pH 7.5; 1 mM EDTA; 2 M NaCl). This is an approximate 1.5 μL of beads per well. After the final wash the beads are suspended in 300 μL 2x BW Buffer.

Once the protease is completely denatured, the plate is briefly spun down and 3 μL of washed beads are added to each well. The plate is incubated for 1 hour at 25C on a shaker set to 1000 rpm. The plate is put on a magnet to concentrate the beads to the side of the well. The supernant, approximately 3 to 4 μL, is removed and moved into a new plate while carefully maintaining the correct orientations. The supernatant faction contains the scWGS side of the coassay. The orientation matching is critical to ensure that PCR barcodes can be paired for the two sides of the co-assy during subsequent indexing steps. An additional 13 μL. 2x BW Buffer are added to the beads to wash them. The beads are concentrated on the magnet again, with the 13 μL being transferred to the same scWGS plate.

Targeted V(D)J RNA Side: Beads Only

The beads are resuspended in a PCR master mix. Each reaction requires 25 µL Equinox Amplification Master Mix, 0.5 µL SyberGreen, 2 µL 10 µM flow cell i7 primer, and 2 µL 10 µM i5 TruSeq-indexed IGHL primer, and H_2O for a total reaction volume of 50 µL. These are initial run a thermocycler for 6 cycles with the following conditions: 98C for 45sec, [cycles: 98C for 15sec, 63C for 30sec, 72C for 30sec], 72C for 1 minute. Subsequently the reaction was set on a magnet block, the supernatant was moved to a new plate, and it was continued running on a BioRad CFX thermocycler under the same PCR conditions. This was done to enable fluorescence image acquisition after each cycle, with amplification typically observed after an additional 16 cycles. The presence of the C1 beads otherwise interferes with proper interpretation of the RFUs on the machine. Following the completion of the PCR, 25 µL from a well can be pooled for sets of 8 wells. Add a volume of NGS SPRI beads equal to the reaction volume (individual reactions are 50 µL, depends on if how many you pool) is added and allowed to bind for 5 minutes. Using a magnet to remove the supernatant, we proceeded with two consecutive 200 µL 80% EtOH washes, made fresh. After the second wash, the plate is briefly spun down, residual EtOH is removed with a 10 µL pipette, and the bead pellet is air dried for approximately 5 minutes. It's careful observed that bead pellet does not begin to crack. Subsequently the pellet is resuspended in 30 µL of elution buffer and incubated for 5 minutes at room temperature off the magnet. The magnet is used to remove the library from the beads. The supernatant is moved to a new tube and ready for quantification via Qubit dsDNA High Sensitivity assay (Thermo Fisher Q32851). If required, based on the Qubit reading, libraries are dilute to an approximate 2ng/µL. The molarity is confirmed and visualized via the Agilent Tapestation 4150 D1000 tape (Agilent 5067-5592).

s3WGS Side: Supernatant Only

The Linear Extension Master Mix (Per Reaction: 0.15 µL 10 µM LNA oligo, 2 µL 5x VeraSeq 2.0 Buffer, 0.1 µL VersaSeq 2.0 DNA Pol, 0.1 µL 10 mM dNTPs, and 4.15 µL dH_2O) is added to the supernatant plate. This is incubated for 10 minutes 72C. The plate is briefly spun down and then the PCR master mix is added directly (Pre Reaction: 12.5 µL WatchMaker Ultra 2x MM,

0.25 µL SyberGreen, 1.0 µL 10 µM i7 flow cell primer, 0.25 µL dH_2O). Subsequently 1 µL 10 µM i5 TruSeq-indexed primer is added so that the indexes match the orientation of the V(D)J side of the coassay. The PCR conditions are as follows: 98C for 30sec, [cycles: 98C for 15sec, 63C for 30sec, 72C for 30sec], 72C for 1 minute. The plate should be pulled when approximately 75% of the wells appear to be exponentially amplifying. Following the PCR, the 25 µL from a well can be pooled for sets of 8 wells. Proceed with the SPRI clean as described above. Finally, run on the Qubit and TapeStation to quantify as previously described for the V(D)J side of the coassay.

Discussion

This dissertation presents the development and application of a novel single-cell whole genome sequencing (scWGS) method that addresses accessibility barriers in cancer genomics research, with applications to two distinct cancer biology systems.

Chapter 1 addressed the accessibility barrier in the field by developing a high throughput scWGS method that leverages widely available commercial platforms. By combining nucleosome disruption methodologies with the 10x Genomics ATAC-seq workflow, this approach democratizes scWGS technology, making it accessible to researchers without specialized equipment or custom reagents. The method successfully achieves uniform genome coverage suitable for copy number analysis while maintaining the high throughput necessary for comprehensive characterization. The double tagmentation optimization additionally offers a unique advantage: the potential to capture both genome-wide copy number and chromatin accessibility information from the same nuclei. As the field increasingly demands integrated multi-omic approaches, this dual-readout capability, despite undeniable cell recovery limitations, provides a tool for more comprehensive single-cell characterization.

The assay retains limitations common to short read, single-cell whole genome assays. The primary challenges encountered included insufficient coverage depth for reliable structural variant detection, which would be desirable when studying hybrid fusion cells and cancer biology. The Tn5 transposase tagmentation patterns are stochastic and uniform to a point; a bias towards DNA in the periphery of the nucleus relative to the center was observed. Another limitation to the scWGS approach is that it does not provide sufficient coverage for de novo single nucleotide variant calling due to the lack of a genome amplification step – something common to direct tagmentation scWGS techniques, restricting its application to copy number assessment, genotyping of known variants, or variant calling from pseudo-bulked data within copy-number-defined clusters.

Despite cost barriers to the 10x Chromium platform, several strategies make this method economically more viable. The well-documented potential for overloading[187] enables processing of 75,000-90,000 nuclei per lane, with eight lanes per chip, while maintaining acceptable collision rates. Additionally, in Chapter 1, the ability to multiplex several samples in a single chip lane is demonstrated when utilizing ScaleBio's indexed Tn5s to deconvolve samples after sequencing. This offers flexibility, and addresses cost, by allowing many samples to be run in multiplex in a single chip lane. Finally, it should be considered that the presented scWGS methodology also works on the Takara ICELL8 system, by similar preprocessing of nuclei to modify published sciATAC methods[84], for high-throughput cells at a lower reagent cost.

Application of this method to both hybrid fusion cells and V(D)J region capture revealed additional technical boundaries that inform future development priorities and research strategies. Long-read sequencing adaptation presents a compelling future direction. Oxford Nanopore's recent advances in duplex sequencing and improved basecalling algorithms now enable reliable capture of 1–5 kbp fragments with high fidelity. This capacity would revolutionize several applications. The entire V(D)J region (~550-650 bp) could be captured intact, complex structural variants could be fully resolved, and chromatin domain boundaries could be preserved. While adapter exchange from Illumina to Nanopore is straightforward, the primary challenge remains sequencing through crosslinked DNA fragments, where adducts[188] may interfere with nanopore translocation. An additional consideration would be a titration of the Tn5 transposase enzyme to produce genomic fragments of larger size for the scWGS fraction, as the concentrations are currently optimized for Illumina read lengths. Nanopore is also inherently suitable for sequencing methylation information[189], which may be of particular interest in FL.

The field's rapid advancement in just two years, particularly in single-cell Fiber-sequencing technologies, offers transformative potential for studying complex genomic arrangements like those in hybrid fusion cells. Single-cell Fiber-seq can capture entire chromatin fibers (10–100 kb)[190–192]. At baseline, this technology offers the ability to capture complex structural

rearrangements, tandem duplications or inversions, potentially even TADs. In the context of hybrid fusion biology, the short-read scWGS approach provided no discrimination between contributing genomes and failed to assign chromosomal segments to their cancer or immune cell origins. Haplotype-resolved long-read sequencing would potentially enable these critical analyses. Moreover, by capturing full-length, intact chromatin fibers, Fiber-seq would confirm whether cancer and immune fibers exist in parallel or if they undergo recombination in the hybrid fusion cells. If genomes are in fact recombined, it could reveal whether the hybrid genomes maintain distinct nuclear territories or show reproducible preferential associations. As Fiber-seq conveys chromatin accessibility information, it could serve as a deeper-dive into which parental genome contributes actively transcribed alleles. This could be valuable in further explaining the tumor-like and macrophage-like accessibility profiles, transcriptomes, and phenotypes observed in Chapter 2.

While whole-genome interrogation of hybrid fusion cells revealed methodological limitations, Chapter 2 demonstrates the power of complementary approaches. Through integrated single-cell transcriptomics, single-cell chromatin accessibility assays, phenotypic characterization, and cyclic immunofluorescence, we identified how RUNX1 and its downstream targets drive metastasis and are associated with disease progression in human colorectal cancer patients.

A future longitudinal study could collect circulating hybrid fusion cells at multiple timepoints to determine whether tumor-like and macrophage-like hybrids represent terminal differentiation fates or continue evolving, particularly in response to therapeutic intervention.

Addendum 1's proposed V(D)J-seq coassay extends the Chapter 1 methodology to address unique challenges in B-cell malignancies, specifically follicular lymphoma. By simultaneously capturing genome-wide alterations and V(D)J clonal identifiers, this approach enables unprecedented resolution of intratumoral clonal diversity. The endogenous molecular barcoding of B-cell receptors provides a framework for tracking subclonal evolution and understanding differential therapeutic responses.

Preliminary data indicates successful generation of appropriately sized V(D)J fragments, representing a key validation milestone. Given the established s3WGS protocol, banked patient samples can be processed rapidly upon completion of the V(D)J protocol. This technology could fundamentally transform B-cell malignancy management by serving as both a molecular diagnostic tool and precision medicine foundation. The minimally invasive nature of blood draws and fine needle aspirations enables longitudinal tracking with minimal patient burden, allowing clinicians to monitor subclonal shifts throughout therapeutic interventions. Despite coverage limitations, this unbiased approach offers potential for discovering novel biomarkers and therapeutic targets.

The V(D)J-seq approach represents an initial paradigm shift. Future integration with the Adey lab's sciMET methodology would constitute a second paradigm shift, adding epigenetic methylation landscapes to genomic alterations within the same cells. The shared tagmentation chemistry ensures fundamental compatibility, with sciMET's additional downstream steps already published and validated. This dual-layer approach would reveal epigenetic states driving therapeutic resistance or disease progression in FL subclones. Such multi-dimensional single-cell resolution could further enable routine clinical monitoring of clonal evolution, enhanced detection of minimal residual disease, and early identification of resistance mechanisms—extending beyond follicular lymphoma to other B-cell malignancies and T-cell cancers with V(D)J recombination signatures.

Beyond the specific applications demonstrated here, this accessible scWGS platform addresses a critical gap in cancer research infrastructure. By eliminating the need for specialized equipment like the DLP platform or custom microfluidics, it enables smaller research groups and clinical laboratories to participate in cutting-edge genomic studies, potentially accelerating discovery across diverse cancer types and patient populations. Although technical limitations remain – particularly in structural variant detection and read length constraints – the convergence of this approach with emerging long-read technologies and multi-omic integration strategies promises to unlock previously intractable questions in tumor heterogeneity, clonal evolution, and

therapeutic resistance. Ultimately, this work establishes a technological foundation that further bridges the gap between basic cancer biology research and clinical application, offering a pathway toward more precise, mechanistically informed therapeutic strategies. This dissertation demonstrates how methodological innovation in single-cell genomics can democratize access to cutting-edge technologies while revealing fundamental insights into cancer biology.

References

1. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576 (2010).
2. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
3. Turner, B. M. Histone acetylation and an epigenetic code. *Bioassays* **9**, 836–845 (2000).
4. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature 2012 489:7414* **489**, 57–74 (2012).
5. Tolhuis, B., Palstra, R. J., Splinter, E., Grosveld, F. & De Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**, 1453–1465 (2002).
6. Spilianakis, C. G., Lalioti, M. D., Town, T., Lee, G. R. & Flavell, R. A. Interchromosomal associations between alternatively expressed loci. *Nature 2005 435:7042* **435**, 637–645 (2005).
7. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
8. Dixon, J. R. *et al.* Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions. *Nature* **485**, 376 (2012).
9. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–329 (2015).
10. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
11. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res* **22**, 1748–1759 (2012).
12. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* **48**, 1119–1130 (2016).
13. Minussi, D. C. *et al.* Breast tumours maintain a reservoir of subclonal diversity during expansion. *Nature 2021 592:7853* **592**, 302–308 (2021).
14. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature 2020 578:7793* **578**, 122–128 (2020).
15. Losic, B. *et al.* Intratumoral heterogeneity and clonal evolution in liver cancer. *Nature Communications 2020 11:1* **11**, 1–15 (2020).
16. Loeb, L. A., Bielas, J. H. & Beckman, R. A. Cancers exhibit a mutator phenotype: Clinical implications. *Cancer Res* **68**, 3551–3557 (2008).
17. Waddington, C. H. Genetic Assimilation of the Bithorax Phenotype. *Evolution (N Y)* **10**, 1–13 (1956).
18. Waddington, C. H. The epigenotype. 1942. *Int J Epidemiol* **41**, 10–13 (2012).
19. Waddington, C. H. The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology. *George Allen & Unwin* (1957).
20. Holliday R, P. JE. DNA modification mechanisms and gene activity during development. Science. *Science (1979)* **187**, 226–232 (1975).
21. Riggs, A. D. X inactivation, differentiation, and DNA methylation. *Cytogenet Cell Genet* **14**, 9–25 (1975).

22. Bird A. DNA methylation patterns and epigenetic memory. . *Genes Dev.* **16**, 6–21 (2002).
23. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature 2001 409:6822* **409**, 860–921 (2001).
24. Ohno, S. So Much "'Junk"' DNA in Our Genome. *In Evolution of Genetic Systems* 366–370 https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1834025 (1972).
25. Farh, K. K. H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
26. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature 2011 472:7341* **472**, 90–94 (2011).
27. Telenius, H. *et al.* Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718–725 (1992).
28. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci U S A* **99**, 5261–5266 (2002).
29. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, (2017).
30. Chen, C. *et al.* SINGLE-CELL GENOMICS Single-cell whole-genome analyses by Linear Amplification via Transposon Insertion (LIANTI). https://www.science.org.
31. 10x Genomics. (n.d.). Discontinuation of Linked-Reads. *https://www.10xgenomics.com/products/linked-reads*.
32. O'Connell, B. L. *et al.* Atlas-scale single-cell chromatin accessibility using nanowell-based combinatorial indexing. *Genome Res* **33**, 208–217 (2023).
33. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nature Biotechnology 2021 39:12* **39**, 1574–1580 (2021).
34. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* **14**, 302–308 (2017).
35. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
36. Garvin, T. *et al.* Interactive analysis and assessment of single-cell copy-number variations. *Nat Methods* **12**, 1058–1060 (2015).
37. Moore, T. W. & Yardımcı, G. G. Robust CNV detection using single-cell ATAC-seq. *bioRxiv* 2023.10.04.560975 (2023) doi:10.1101/2023.10.04.560975.
38. Gast, C. E. *et al.* Cell fusion potentiates tumor heterogeneity and reveals circulating hybrid cells that correlate with stage and survival. *Sci Adv* **4**, (2018).
39. Pawelek, J. M. Tumour-cell fusion as a source of myeloid traits in cancer. *Lancet Oncol* **6**, 988–993 (2005).
40. Sutton, T. L., Walker, B. S. & Wong, M. H. Circulating Hybrid Cells Join the Fray of Circulating Cellular Biomarkers. *Cell Mol Gastroenterol Hepatol* **8**, 595–607 (2019).
41. Kaseb, H., Ali, M. A., Gasalberti, D. P. & Koshy, N. V. Follicular Lymphoma. *StatPearls* https://www.ncbi.nlm.nih.gov/books/NBK538206/ (2024).
42. Merryman, R., Mehtap, Ö. & LaCasce, A. Advancements in the Management of Follicular Lymphoma: A Comprehensive Review. *Turk J Haematol* **41**, 69–82 (2024).
43. Cyster, J. G. & Allen, C. D. C. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell* **177**, 524–540 (2019).

44. Mesin, L., Ersching, J. & Victora, G. D. Germinal Center B Cell Dynamics. *Immunity* **45**, 471–482 (2016).
45. Victora, G. D. & Nussenzweig, M. C. Germinal Centers. *Annu Rev Immunol* **40**, 413–442 (2022).
46. Vodicka, P., Klener, P. & Trneny, M. Diffuse Large B-Cell Lymphoma (DLBCL): Early Patient Management and Emerging Treatment Options. *Onco Targets Ther* **15**, 1481 (2022).
47. Schatz, D. G. *et al.* The Mechanism, Regulation and Evolution of V(D)J Recombination. *Molecular Biology of B Cells, Third Edition* 13–57 (2024) doi:10.1016/B978-0-323-95895-0.00004-0.
48. Tonegawa, S. Somatic generation of antibody diversity. *Nature 1983 302:5909* **302**, 575–581 (1983).
49. Schatz, D. G. & Swanson, P. C. V(D)J recombination: mechanisms of initiation. *Annu Rev Genet* **45**, 167–202 (2011).
50. Haebe, S. *et al.* Follicular lymphoma evolves with a surmountable dependency on acquired glycosylation motifs in the B-cell receptor. *Blood* **142**, 2296–2304 (2023).
51. Merkenschlager, J. *et al.* Regulated somatic hypermutation enhances antibody affinity maturation. *Nature 2025 641:8062* **641**, 495–502 (2025).
52. Fitzgerald, D. *et al.* A single-cell multi-omic and spatial atlas of B-cell lymphomas reveals differentiation drives intratumor heterogeneity Differentiation drives variation in B-cell lymphomas. https://doi.org/10.1101/2023.11.06.565756 doi:10.1101/2023.11.06.565756.
53. Tomlinson, I., Sasieni, P. & Bodmer, W. How Many Mutations in a Cancer? *Am J Pathol* **160**, 755 (2002).
54. Werner, B. *et al.* Measuring single cell divisions in human tissues from multi-region sequencing data. *Nat Commun* **11**, 1035 (2020).
55. Haebe, S. *et al.* Single-cell analysis can define distinct evolution of tumor sites in follicular lymphoma. *Blood* **137**, 2869–2880 (2021).
56. Nichols, R. V. *et al.* Atlas-scale single-cell DNA methylation profiling with sciMETv3. *Cell Genomics* **5**, 100726 (2025).
57. Jiang, Y., Dominguez, P. M. & Melnick, A. M. The many layers of epigenetic dysfunction in B-cell lymphomas. *Curr Opin Hematol* **23**, 377–384 (2016).
58. Liu, M. K. *et al.* Methylation alterations and advance of treatment in lymphoma. *Front Biosci (Landmark Ed)* **26**, 602–613 (2021).
59. Bouska, A. *et al.* Combined copy number and mutation analysis identifies oncogenic pathways associated with transformation of follicular lymphoma. *Leukemia* **31**, 83 (2016).
60. Van der Auwera, G. A. *et al.* From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.1-11.10.33 (2013).
61. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* **37**, 4569–4571 (2021).
62. Knaus, B. J. & Grünwald, N. J. vcfr: a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour* **17**, 44–53 (2017).
63. De Bourcy, C. F. A. *et al.* Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc Natl Acad Sci U S A* **114**, 1105–1110 (2017).

64. Gupta, N. T. *et al.* Change-O: a toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* **31**, 3356–3358 (2015).
65. Baslan, T. *et al.* Genome wide copy number analysis of single cells. *Nat Protoc* **7**, 1024 (2012).
66. Kim, C. *et al.* Chemoresistance Evolution in Triple-Negative Breast Cancer Delineated by Single-Cell Sequencing. *Cell* **173**, 879-893.e13 (2018).
67. Navin, N. *et al.* Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90–95 (2011).
68. Wang, Y. *et al.* Clonal Evolution in Breast Cancer Revealed by Single Nucleus Genome Sequencing. *Nature* **512**, 155 (2014).
69. Danilenko, M. *et al.* Single-cell DNA sequencing identifies risk-associated clonal complexity and evolutionary trajectories in childhood medulloblastoma development. *Acta Neuropathol* **144**, 565–578 (2022).
70. Cai, X. *et al.* Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. *Cell Rep* **8**, 1280–1289 (2014).
71. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat Biotechnol* **39**, 1574–1580 (2021).
72. Vitak, S. A. *et al.* Sequencing thousands of single-cell genomes with combinatorial indexing. *Nat Methods* **14**, 302–308 (2017).
73. Mulqueen, R. M. *et al.* High-content single-cell combinatorial indexing. *Nat Biotechnol* https://doi.org/10.1038/s41587-021-00962-z (2021) doi:10.1038/s41587-021-00962-z.
74. Nichols, R. V. *et al.* High-throughput robust single-cell DNA methylation profiling with sciMETv2. *Nature Communications 2022 13:1* **13**, 1–10 (2022).
75. Jiang, Y. *et al.* CODEX2: full-spectrum copy number variation detection by high-throughput DNA sequencing. *Genome Biol* **19**, 202 (2018).
76. Wang, R., Lin, D. Y. & Jiang, Y. SCOPE: A Normalization and Copy-Number Estimation Method for Single-Cell DNA Sequencing. *Cell Syst* **10**, 445-452.e6 (2020).
77. 10x Genomics. 10xgenomics.com/resources/datasets.
78. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nature Biotechnology 2019 37:8* **37**, 916–924 (2019).
79. Granja, J. M. *et al.* ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics 2021 53:3* **53**, 403–411 (2021).
80. Yardımcı, G. G. *et al.* Measuring the reproducibility and quality of Hi-C data. *Genome Biol* **20**, (2019).
81. Minussi, D. C. *et al.* Breast Tumors Maintain a Reservoir of Subclonal Diversity During Expansion. *Nature* **592**, 302 (2021).
82. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
83. Liu, X. *et al.* Conditional reprogramming and long-term expansion of normal and tumor cells from human biospecimens. *Nat Protoc* **12**, 439–451 (2017).
84. O'Connell, B. L. *et al.* Atlas-scale single-cell chromatin accessibility using nanowell-based combinatorial indexing. *Genome Res* **33**, 208–217 (2023).
85. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLoS One* **7**, e30377 (2012).

86. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
87. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)* **57**, 289–300 (1995).
88. Mani, K. *et al.* Causes of death among people living with metastatic cancer. *Nat Commun* **15**, (2024).
89. Qiu, M., Hu, J., Yang, D., Cosgrove, D. P. & Xu, R. Pattern of distant metastases in colorectal cancer: a SEER based study. *Oncotarget* **6**, 38658–38666 (2015).
90. Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J Clin* **74**, 12–49 (2024).
91. Welch, D. R. & Hurst, D. R. Defining the Hallmarks of Metastasis. *Cancer Res* **79**, 3011–3027 (2019).
92. Gerstberger, S., Jiang, Q. & Ganesh, K. Metastasis. *Cell* **186**, 1564–1579 (2023).
93. Peng, X.-C. *et al.* Cell-cell fusion as an important mechanism of tumor metastasis (Review). *Oncol Rep* **46**, (2021).
94. Wang, H.-F. *et al.* Cell fusion in cancer hallmarks: Current research status and future indications (Review). *Oncol Lett* **22**, (2021).
95. Brown, K. E. & Fisher, A. G. Reprogramming lineage identity through cell–cell fusion. *Current Opinion in Genetics & Development* **70**, 15–23 (2021).
96. Hass, R., von der Ohe, J. & Dittmar, T. Cancer Cell Fusion and Post-Hybrid Selection Process (PHSP). *Cancers (Basel)* **13**, 4636 (2021).
97. Sieler, M., Weiler, J. & Dittmar, T. Cell–Cell Fusion and the Roads to Novel Properties of Tumor Hybrid Cells. *Cells* **10**, 1465 (2021).
98. Rizvi, A. Z. *et al.* Bone marrow-derived cells fuse with normal and transformed intestinal stem cells. *Proceedings of the National Academy of Sciences* **103**, 6321–6325 (2006).
99. Aguirre, L. A. *et al.* Tumor stem cells fuse with monocytes to form highly invasive tumor-hybrid cells. *Oncoimmunology* **9**, (2020).
100. Barreto, S. G., Gardi, N. & Dutt, S. Birth of a solid organ cancer—the cell fusion hypothesis presented with pancreatic cancer as a model: a narrative review. *Chin Clin Oncol* **10**, 45 (2021).
101. Berndt, B., Zanker, K. S. & Dittmar, T. Cell Fusion is a Potent Inducer of Aneuploidy and Drug Resistance in Tumor Cell/ Normal Cell Hybrids. *Crit Rev Oncog* **18**, 97–113 (2013).
102. Brito, A. *et al.* Cell fusion enhances energy metabolism of mesenchymal tumor hybrid cells to sustain their proliferation and invasion. *BMC Cancer* **21**, (2021).
103. Carloni, V., Mazzocca, A., Mello, T., Galli, A. & Capaccioli, S. Cell fusion promotes chemoresistance in metastatic colon carcinoma. *Oncogene* **32**, 2649–2660 (2012).
104. Clawson, G. A. *et al.* Macrophage-Tumor Cell Fusions from Peripheral Blood of Melanoma Patients. *PLoS One* **10**, e0134320 (2015).
105. Cozzo, A. J., Coleman, M. F. & Hursting, S. D. You complete me: tumor cell-myeloid cell nuclear fusion as a facilitator of organ-specific metastasis. *Front Oncol* **13**, (2023).
106. Parappilly, M. S. *et al.* Circulating Neoplastic-Immune Hybrid Cells Predict Metastatic Progression in Uveal Melanoma. *Cancers (Basel)* **14**, 4617 (2022).

107. Clawson, G. A. *et al.* “Stealth dissemination” of macrophage-tumor cell fusions cultured from blood of patients with pancreatic ductal adenocarcinoma. *PLoS One* **12**, e0184451 (2017).
108. Gast, C. E. *et al.* Cell fusion potentiates tumor heterogeneity and reveals circulating hybrid cells that correlate with stage and survival. *Sci. Adv* **4**, 7828–7840 (2018).
109. Zhang, L.-N., Huang, Y.-H. & Zhao, L. Fusion of macrophages promotes breast cancer cell proliferation, migration and invasion through activating epithelial-mesenchymal transition and Wnt/β-catenin signaling pathway. *Arch Biochem Biophys* **676**, 108137 (2019).
110. Dörnen, J., Myklebost, O. & Dittmar, T. Cell Fusion of Mesenchymal Stem/Stromal Cells and Breast Cancer Cells Leads to the Formation of Hybrid Cells Exhibiting Diverse and Individual (Stem Cell) Characteristics. *Int J Mol Sci* **21**, 9636 (2020).
111. Merle, C., Lagarde, P., Lartigue, L. & Chibon, F. Acquisition of cancer stem cell capacities after spontaneous cell fusion. *BMC Cancer* **21**, (2021).
112. Chen, Y.-C. *et al.* Mesenchymal Stem/Stromal Cell Engulfment Reveals Metastatic Advantage in Breast Cancer. *Cell Rep* **27**, 3916-3926.e5 (2019).
113. Cowan, C. A., Atienza, J., Melton, D. A. & Eggan, K. Nuclear Reprogramming of Somatic Cells After Fusion with Human Embryonic Stem Cells. *Science (1979)* **309**, 1369–1373 (2005).
114. Ding, J., Jin, W., Chen, C., Shao, Z. & Wu, J. Tumor Associated Macrophage × Cancer Cell Hybrids May Acquire Cancer Stem Cell Properties in Breast Cancer. *PLoS One* **7**, e41942 (2012).
115. Dittmar, T. *et al.* Characterization of hybrid cells derived from spontaneous fusion events between breast epithelial cells exhibiting stem-like characteristics and breast cancer cells. *Clinical & Experimental Metastasis* **28**, 75–90 (2010).
116. FAN, H. & LU, S. Fusion of human bone hemopoietic stem cell with esophageal carcinoma cells didn’t generate esophageal cancer stem cell. *Neoplasma* **62**, 540–545 (2014).
117. Gauck, D., Keil, S., Niggemann, B., Zänker, K. S. & Dittmar, T. Hybrid clone cells derived from human breast epithelial cells and human breast cancer cells exhibit properties of cancer stem/initiating cells. *BMC Cancer* **17**, (2017).
118. Hass, R., von der Ohe, J. & Ungefroren, H. Potential Role of MSC/Cancer Cell Fusion and EMT for Breast Cancer Stem Cell Formation. *Cancers (Basel)* **11**, 1432 (2019).
119. He, X. *et al.* Cell fusion between gastric epithelial cells and mesenchymal stem cells results in epithelial-to-mesenchymal transition and malignant transformation. *BMC Cancer* **15**, (2015).
120. Islam, M. Q., Meirelles, L. da S., Nardi, N. B., Magnusson, P. & Islam, K. Polyethylene Glycol-Mediated Fusion between Primary Mouse Mesenchymal Stem Cells and Mouse Fibroblasts Generates Hybrid Cells with Increased Proliferation and Altered Differentiation. *Stem Cells Dev* **15**, 905–919 (2006).
121. Melzer, C., von der Ohe, J. & Hass, R. Involvement of Actin Cytoskeletal Components in Breast Cancer Cell Fusion with Human Mesenchymal Stroma/Stem-Like Cells. *Int J Mol Sci* **20**, 876 (2019).
122. Ramakrishnan, M., Mathur, S. R. & Mukhopadhyay, A. Fusion-Derived Epithelial Cancer Cells Express Hematopoietic Markers and Contribute to Stem Cell and Migratory Phenotype in Ovarian Carcinoma. *Cancer Res* **73**, 5360–5370 (2013).

123. Rappa, G., Mercapide, J. & Lorico, A. Spontaneous Formation of Tumorigenic Hybrids between Breast Cancer and Multipotent Stromal Cells Is a Source of Tumor Heterogeneity. *Am J Pathol* **180**, 2504–2515 (2012).
124. Sottile, F., Aulicino, F., Theka, I. & Cosma, M. P. Mesenchymal stem cells generate distinct functional hybrids in vitro via cell fusion or entosis. *Sci Rep* **6**, (2016).
125. Wang, R. *et al.* Fusion with stem cell makes the hepatocellular carcinoma cells similar to liver tumor-initiating cells. *BMC Cancer* **16**, (2016).
126. Wang, R. *et al.* Spontaneous Cancer-Stromal Cell Fusion as a Mechanism of Prostate Cancer Androgen-Independent Progression. *PLoS One* **7**, e42653 (2012).
127. Zhang, L. *et al.* Roles of cell fusion between mesenchymal stromal/stem cells and malignant cells in tumor growth and metastasis. *FEBS J* **288**, 1447–1456 (2020).
128. Melzer, C., von der Ohe, J. & Hass, R. Enhanced metastatic capacity of breast cancer cells after interaction and hybrid formation with mesenchymal stroma/stem cells (MSC). *Cell Communication and Signaling* **16**, (2018).
129. Melzer, C., von der Ohe, J. & Hass, R. In Vivo Cell Fusion between Mesenchymal Stroma/Stem-Like Cells and Breast Cancer Cells. *Cancers (Basel)* **11**, (2019).
130. Melzer, C., von der Ohe, J. & Hass, R. In Vitro Fusion of Normal and Neoplastic Breast Epithelial Cells with Human Mesenchymal Stroma/Stem Cells Partially Involves Tumor Necrosis Factor Receptor Signaling. *Stem Cells* **36**, 977–989 (2018).
131. Dietz, M. S. *et al.* Relevance of circulating hybrid cells as a non-invasive biomarker for myriad solid tumors. *Sci Rep* **11**, (2021).
132. Manjunath, Y. *et al.* Tumor-Cell–Macrophage Fusion Cells as Liquid Biomarkers and Tumor Enhancers in Cancer. *Int J Mol Sci* **21**, 1872 (2020).
133. Henn, T. E. *et al.* Circulating hybrid cells predict presence of occult nodal metastases in oral cavity carcinoma. *Head & Neck* **43**, 2193–2201 (2021).
134. Walker, B. S. *et al.* Circulating Hybrid Cells: A Novel Liquid Biomarker of Treatment Response in Gastrointestinal Cancers. *Ann Surg Oncol* **28**, 8567–8578 (2021).
135. Sutton, T. L., Walker, B. S. & Wong, M. H. Circulating Hybrid Cells Join the Fray of Circulating Cellular Biomarkers. *CMGH* **8**, 595–607 (2019).
136. Zhang, Y. *et al.* Patterns of circulating tumor cells identified by CEP8, CK and CD45 in pancreatic cancer. *Int J Cancer* **136**, 1228–1233 (2014).
137. Ganesh, K. *et al.* L1CAM defines the regenerative origin of metastasis-initiating cells in colorectal cancer. *Nat Cancer* **1**, 28–45 (2020).
138. Anderson, A. N. *et al.* Detection of neoplastic-immune hybrid cells with metastatic properties in uveal melanoma. *Biomark Res* **12**, (2024).
139. Joanito, I. *et al.* Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. *Nat Genet* **54**, 963–975 (2022).
140. Ali, A. M. & Raza, A. scRNAseq and High-Throughput Spatial Analysis of Tumor and Normal Microenvironment in Solid Tumors Reveal a Possible Origin of Circulating Tumor Hybrid Cells. *Cancers (Basel)* **16**, 1444 (2024).
141. Ye, X. *et al.* Myeloid-like tumor hybrid cells in bone marrow promote progression of prostate cancer bone metastasis. *Journal of Hematology & Oncology* **16**, (2023).

142. Zhou, C., Gao, Y., Ding, P., Wu, T. & Ji, G. The role of CXCL family members in different diseases. *Cell Death Discov* **9**, (2023).
143. Garlanda, C. & Mantovani, A. Interleukin-1 in tumor progression, therapy, and prevention. *Cancer Cell* **39**, 1023–1027 (2021).
144. Koroleva, E. P., Fu, Y.-X. & Tumanov, A. V. Lymphotoxin in physiology of lymphoid tissues – Implication for antiviral defense. *Cytokine* **101**, 39–47 (2018).
145. Bharti, R., Dey, G., Lin, F., Lathia, J. & Reizes, O. CD55 in cancer: Complementing functions in a non-canonical manner. *Cancer Lett* **551**, 215935 (2022).
146. Ito, Y., Bae, S.-C. & Chuang, L. S. H. The RUNX family: developmental regulators in cancer. *Nat Rev Cancer* **15**, 81–95 (2015).
147. Huang, J., Chen, W., Jie, Z. & Jiang, M. Comprehensive Analysis of Immune Implications and Prognostic Value of SPI1 in Gastric Cancer. *Front Oncol* **12**, (2022).
148. Safe, S. Specificity Proteins (Sp) and Cancer. *Int J Mol Sci* **24**, 5164 (2023).
149. Milde-Langosch, K. The Fos family of transcription factors and their role in tumourigenesis. *Eur J Cancer* **41**, 2449–2461 (2005).
150. Meier, M. J., Beal, M. A., Schoenrock, A., Yauk, C. L. & Marchetti, F. Whole Genome Sequencing of the Mutamouse Model Reveals Strain- and Colony-Level Variation, and Genomic Features of the Transgene Integration Site. *Sci Rep* **9**, 13775 (2019).
151. Doran, A. G. *et al.* Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biol* **17**, 1–16 (2016).
152. Walker, B. S. *et al.* Circulating Hybrid Cells: A Novel Liquid Biomarker of Treatment Response in Gastrointestinal Cancers. *Ann Surg Oncol* **28**, 8567 (2021).
153. Mitra, S. & Tomar, P. C. Hybridoma technology; advancements, clinical significance, and future aspects. *Journal of Genetic Engineering & Biotechnology* **19**, 159 (2021).
154. Chen, Y., He, Y. & Liu, S. RUNX1-Regulated Signaling Pathways in Ovarian Cancer. *Biomedicines* **11**, 2357 (2023).
155. Hong, D. *et al.* RUNX1-dependent mechanisms in biological control and dysregulation in cancer. *J Cell Physiol* **234**, 8597–8609 (2018).
156. Lie-A-Ling, M. *et al.* RUNX1 positively regulates a cell adhesion and migration program in murine hemogenic endothelium prior to blood emergence. *Blood* **124**, e11–e20 (2014).
157. Mercado-Matos, J., Matthew-Onabanjo, A. N. & Shaw, L. M. RUNX1 and breast cancer. *Oncotarget* **8**, 36934–36935 (2017).
158. Planagumà, J. *et al.* A Differential Gene Expression Profile Reveals Overexpression of RUNX1/AML1 in Invasive Endometrioid Carcinoma. *Cancer Res* **64**, 8846–8853 (2004).
159. Sangpairoj, K. *et al.* RUNX1 Regulates Migration, Invasion, and Angiogenesis via p38 MAPK Pathway in Human Glioblastoma. *Cell Mol Neurobiol* **37**, 1243–1255 (2016).
160. Zheng, W. *et al.* RUNX1-induced upregulation of PTGS2 enhances cell growth, migration and invasion in colorectal cancer cells. *Sci Rep* **14**, (2024).
161. Guo, X. *et al.* RUNX1 promotes angiogenesis in colorectal cancer by regulating the crosstalk between tumor cells and tumor associated macrophages. *Biomark Res* **12**, (2024).
162. Rada, M. *et al.* Runt related transcription factor-1 plays a central role in vessel co-option of colorectal cancer liver metastases. *Commun Biol* **4**, (2021).

163. Li, Q. *et al.* RUNX1 promotes tumour metastasis by activating the Wnt/β-catenin signalling pathway and EMT in colorectal cancer. *Journal of Experimental & Clinical Cancer Research* **38**, (2019).
164. Cunningham, L. *et al.* Identification of benzodiazepine Ro5-3335 as an inhibitor of CBF leukemia through quantitative high throughput screen against RUNX1–CBFβ interaction. *Proceedings of the National Academy of Sciences* **109**, 14592–14597 (2012).
165. She, C. *et al.* Combination of RUNX1 inhibitor and gemcitabine mitigates chemo-resistance in pancreatic ductal adenocarcinoma by modulating BiP/PERK/eIF2α-axis-mediated endoplasmic reticulum stress. *Journal of Experimental & Clinical Cancer Research* **42**, (2023).
166. Santamarina-Ojeda, P. *et al.* Multi-omic integration of <scp>DNA</scp> methylation and gene expression data reveals molecular vulnerabilities in glioblastoma. *Mol Oncol* **17**, 1726–1743 (2023).
167. Cheung, G., Pauler, F. M., Koppensteiner, P. & Hippenmeyer, S. Protocol for mapping cell lineage and cell-type identity of clonally-related cells in situ using MADM-CloneSeq. *STAR Protoc* **5**, 103168 (2024).
168. Short, S., García-Tejera, R., Schumacher, L. J. & Coutu, D. L. Next generation lineage tracing and its applications to unravel development. *NPJ Syst Biol Appl* **11**, (2025).
169. Lee, S. E., Rudd, B. D. & Smith, N. L. Fate-mapping mice: new tools and technology for immune discovery. *Trends Immunol* **43**, 195–209 (2022).
170. Bais, A. S. & Kostka, D. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics* **36**, 1150–1158 (2019).
171. Xi, N. M. & Li, J. J. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Syst* **12**, 176-194.e6 (2021).
172. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS* **16**, 284–287 (2012).
173. Franca, C. M. *et al.* Perivascular cells function as key mediators of mechanical and structural changes in vascular capillaries. *Sci Adv* **11**, (2025).
174. Polacheck, W. J., Kutys, M. L., Tefft, J. B. & Chen, C. S. Microfabricated blood vessels for modeling the vascular transport barrier. *Nat Protoc* **14**, 1425–1454 (2019).
175. McMahon, N. P. *et al.* Oligonucleotide conjugated antibodies permit highly multiplexed immunofluorescence for future use in clinical histopathology. *J Biomed Opt* **25**, 1 (2020).
176. McMahon, N. *et al.* Signal removal methods for highly multiplexed immunofluorescent staining using antibody conjugated oligonucleotides. in *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues XVII* (eds. Farkas, D. L., Leary, J. F. & Tarnok, A.) 32 (SPIE, 2019). doi:10.1117/12.2510573.
177. Jones, J. A. *et al.* Oligonucleotide conjugated antibody strategies for cyclic immunostaining. *Sci Rep* **11**, (2021).
178. McMahon, N. P. *et al.* Flexible Cyclic Immunofluorescence (cyCIF) Using Oligonucleotide Barcoded Antibodies. *Cancers (Basel)* **15**, 827 (2023).
179. Muhlich, J. L. *et al.* Stitching and registering highly multiplexed whole-slide images of tissues and tumors using ASHLAR. *Bioinformatics* **38**, 4613–4621 (2022).

180. Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nat Biotechnol* **40**, 555–565 (2021).
181. Hao, Y. *et al.* Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat Biotechnol* **42**, 293–304 (2023).
182. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications 2017 8:1* **8**, 1–12 (2017).
183. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
184. Queitsch, K. *et al.* Accessible high-throughput single-cell whole-genome sequencing with paired chromatin accessibility. *Cell reports methods* **3**, (2023).
185. Langlois de Septenville, A. *et al.* Immunoglobulin Gene Mutational Status Assessment by Next Generation Sequencing in Chronic Lymphocytic Leukemia. *Methods Mol Biol* **2453**, 153–167 (2022).
186. van Dongen, J. J. M. *et al.* Design and standardization of PCR primers and protocols for detection of clonal immunoglobulin and T-cell receptor gene recombinations in suspect lymphoproliferations: report of the BIOMED-2 Concerted Action BMH4-CT98-3936. *Leukemia* **17**, 2257–2317 (2003).
187. Wu, B. *et al.* Overloading And unpacKing (OAK) - droplet-based combinatorial indexing for ultra-high throughput single-cell multiomic profiling. *Nature Communications 2024 15:1* **15**, 1–12 (2024).
188. Gavrilov, A., Razin, S. V. & Cavalli, G. In vivo formaldehyde cross-linking: it is time for black box analysis. *Brief Funct Genomics* **14**, 163–165 (2015).
189. Kong, Y. *et al.* Critical assessment of nanopore sequencing for the detection of multiple forms of DNA modifications. *bioRxiv* 2024.11.19.624260 (2024) doi:10.1101/2024.11.19.624260.
190. Swanson, E. G. *et al.* Deaminase-assisted single-molecule and single-cell chromatin fiber sequencing. *bioRxiv* 2024.11.06.622310 (2024) doi:10.1101/2024.11.06.622310.
191. Stergachis, A. B., Debo, B. M., Haugen, E., Churchman, L. S. & Stamatoyannopoulos, J. A. Single-molecule regulatory architectures captured by chromatin fiber sequencing. *Science (1979)* **368**, (2020).
192. Bohaczuk, S. C. *et al.* Resolving the chromatin impact of mosaic variants with targeted Fiber-seq. *Genome Res* **34**, 2269 (2024).