GENOMIC EVOLUTION OF PANCREATIC CANCER AT SINGLE-CELL RESOLUTION

By

Haochen Zhang

A Dissertation

Presented to the Faculty of the Louis V. Gerstner, Jr.

Graduate School of Biomedical Sciences,

Memorial Sloan Kettering Cancer Center

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

New York, NY, USA

June 2024

Christine Iacobuzio-Donahue, MD, PhD Dissertation Mentor	Date

To all cancer patients

Abstract

Pancreatic ductal adenocarcinoma (PDAC) is one of the most lethal cancer types with a 5-year survival of less than 12%. Such a rate has increased minimally over the past decades despite improvements in surgical and medical management. Nonetheless, in recent years, collaborative research efforts across the globe have advanced our understanding of the disease's molecular underpinning and revealed several opportunities for precision medicine, while calling for more precise measurement of it molecular features to inform clinical management.

Numerous next-generation sequencing (NGS) studies have revealed PDAC clonal evolution as a hybrid model of stepwise and punctuated events. Point mutations in multiple driver genes (i.e. *KRAS, CDKN2A, TP53*) followed by allelic loss of the wild type alleles are gained through progressive waves of clonal selection, while pervasive copy number variations or larger scale genomic rearrangements that further grant growth advantages are acquired rapidly following polyploidization. Elucidation of these genomic events has largely been accomplished by sequencing patient tissues in bulk; in some cases resolution has been increased by use of microdissected samples and/or development of computational tools to deconvolve the complexity of subclonal mixtures. Regardless, given clonal evolution happens at the single-cell level one must rely on a set of strong assumptions which may overlook important events such colocalization of alterations within a singular genome or mutual exclusivity indicating distinct populations.

Thus, we developed a scalable workflow to apply high-throughput *single-cell* DNA sequencing method to primary PDAC samples and generated the highest resolution view of its genomic evolution to date. Specifically, we defined at single cell resolution the genomic bottlenecks and intercellular heterogeneity present over the course of PDAC evolution spanning early invasion to dissemination to a variety of secondary sites. From a technical perspective the combined wet/dry-lab workflow is the first to apply such single-cell DNA sequencing to solid tumor, and we believe it carries significance for clinical application especially when tumor purity is limited or when high sensitivity for low-prevalence mutations is sought after. From a biomedical perspective, we believe this work established the ground truth of PDAC genomic evolution profile in its native environment. It formed the baseline for comparison with more PDAC precision medicine studies to come.

Biographical sketch

Haochen was born and raised in Beijing, China. He graduated from the High School Affiliated to Renmin University of China in 2015. He then first moved to Houston, TX, USA and then Chicago, IL, USA for his undergraduate studies at Rice University and The University of Chicago, respectively. During the summer after his third year in college, he participated in the Summer Undergraduate Research Program (SURP) of Gerstner Sloan Kettering (GSK) Graduate School at Memorial Sloan Kettering Cancer Center (MSKCC), and was hosted by the lab of Joan Massagué, PhD for 10 weeks, where he conducted research in cancer metastasis. After earning a Bachelor of Science in Biological Sciences and a Bachelor of Arts in Economics from The University of Chicago in 2019, Haochen moved to New York to pursue his PhD at the GSK graduate school. He joined the lab of Christine Iacobuzio Donahue, MD, PhD in the summer of 2020 to conduct thesis research in pancreatic cancer genomics.

Table of Contents

List of Tables

Relevant tables are not included in this dissertation as they are too long. Please refer to the official publication.

List of figures

List of abbreviations

ADO	allelic dropout
AF	allele frequency
CCF	cancer cell fraction
CDKN2A	cyclin-dependent kinase inhibitor 2A; other aliases: *MTS-1* (multiple tumor suppressor-1); p16, where then number is derived from its molecular weight
CN	copy number
CNV	copy number variation
CONDOR	Constrained Dollo Reconstruction
DNA	Deoxyribonucleic acid
HPC	high-performance computing
ICGC	International Cancer Genome Consortium
KRAS	Kristen rat sarcoma viral oncogene homolog
LN	lymph node
LOH	loss of heterozygosity
MRCA	most recent common ancestor
NGS	next-generation sequencing
PCR	polymerase chain reaction
PDAC	pancreatic ductal adenocarcinoma
PON	panel of normal
QC	quality control
RNA	ribonucleic acid
ROS	reactive oxygen species
SMAD	*Caenorhabditis elegans* SMA ("small" worm phenotype) mothers against decapentaplegic
SnDNA-seq	single-nucleus DNA sequencing
SNP	single-nucleotide polymorphism
SNV	single nucleotide variant
SV	structural variant
TCGA	The Cancer Genome Atlas
TGF-β	transforming growth factor beta
TME	tumor microenvironment
TP53	tumor protein p53, where the number is derived from its molecular weight
VAF	variant allele frequency
WES	whole-exome sequencing
WGD	whole-genome duplication
WT	wildtype

Section 1: Introduction

1. The problem

1.1 Pancreatic cancer

Pancreatic ductal adenocarcinoma (PDAC) has surpassed breast cancer to become the third leading cause of cancer-related death in the United States, with a five-year survival rate currently standing at 12% [1]. A more alarming fact is that it is projected to overtake colorectal cancer before 2040, moving only behind lung cancer as a leading cause of cancer-related mortality [2].

Aside from a number of factors that are driving an increase in cancer incidence general, including the obesity epidemic and greater life expectancy[3], many factors unique to PDAC contribute to its lethality:

1. Outside of the minority (10%–15%) of cases ascribed to germline mutations or known risk factors such as mucinous cystic lesions and chronic pancreatitis, there is no single attributable risk factor for most patients[4], hindering preventive efforts in the general population.
2. The only curative treatment would be surgical removal of the tumor[5], but only 15-20% of patients are eligible for surgery at diagnosis [6]. Compared with the surgery's technical complexity and lasting impact on the patient's quality of life, the

difficulty of detecting the disease at a stage early enough for it to have significant benefit is perhaps the main limiting factor.

3. Because (i) localized disease is largely asymptomatic or accompanied by ill-defined symptoms (ii) there lacks diagnostic biomarkers for earlier stage tumors, (iii) the anatomical location of the pancreas is hard to access, thereby preventing routine office-based screening, the disease is usually not detected until locally advanced or metastatic, and present with highly aggressive biology [7].
4. Other tumor types have benefited from the second- and third-generation of cancer therapeutics (targeted therapy and immunotherapy, respectively) and have since seen great improvement in patient outcome[8]. There has not been a broadly-applicable targeted therapy for PDAC until very recently, KRAS inhibitors were developed [9]. Immunotherapy has not been shown to work for the majority of patients, possibly due to the poor immune cell infiltration at the pancreas as well as low immunogenicity due to PDAC biology, although more promising result have also been demonstrated very recently [10].

1.2 Basic research improves clinical management of pancreatic cancer

Although the above facts are dismal at first glance, many reasons for optimism exist, such as the development of second- and third-gen cancer therapeutics mentioned above. The progress attributes to growing understanding of the biomolecular basis of PDAC, while coincides with a newfound appreciation for the complexity of the disease, which occurs not only across patients but can also vary dramatically within individual tumors. Studies of these variances in genomics (DNA), transcriptomics (RNA), proteomics, tumor

microenvironment (TME) and treatment response are clarifying the aspects of the problem that need to be solved and paving the way for better clinical care [6]. I briefly review some of the key advances in basic research and how they have translated into the clinic.

Transcriptional subtypes of pancreatic cancer

Approaches to comprehensively interrogate the transcriptional landscape has identified several PDAC subtypes across patients which demonstrate differential biology and therefore clinical progression pattern. Overall, as methodologies to acquire tumor sample and sequence RNA evolved, consensus has best coalesced around two major pancreatic cancer subtypes, classical and basal-like, with some populations showing mixed features [11].

The recent COMPASS trial provided proof of principle validation of subtype-based treatment of patients. Patients with classical subtype of PDAC had superior responses to first line 5-flurouracil-based regimens than those with basal-like cancers [12]. This initial observation is now being further validated with a multi-center trial that is also examining whether contemporaneously established ex vivo organoid cultures from patient biopsies can provide phenotypic data on a broader range of drug sensitivities.

Tumor microenvironment

PDAC exhibits a famously substantial and intensely studied TME. Immunohistochemical staining on primary tissues and functional studies in mice have unveiled this intricate ecosystem of immune cells, endothelial cells, and cancer-associated fibroblasts (CAFs),

which provides specialized niches for cancer cells and modulates their growth and invasive behavior [13].

Aside from inspiring many ongoing trials that directly target the PDAC TME, such studies have also identified stromal subtypes that predict response to chemotherapy [14], [15]; immune milieu that is prognostic [16], [17] and predictive of response to immunotherapy [18].

Due to the limit of space, I will not be able to cover all aspects PDAC biology and treatment. For the rest of the Introduction Section I will focus on the genomic aspect of the disease which is most relevant to this dissertation.

2. Genomic study of pancreatic cancer

"The revolution in cancer research can be summed up in a single sentence: cancer is, in essence, a genetic disease." -- Vogelstein and Kinzler, Nature Medicine (2004) [19].

2.1 The significance of studying the cancer genome

In my opinion this significance is two-fold:

For one, our understanding of cancer started from the discovery that alterations in the genome drive oncogenesis [20], [21]. As we found more cancer-associated genetic mutations and pathways/phenotypes they control, we came to understand more details of

cancer's behavior - not only early oncogenesis but also acquisition of abilities to invade, metastasize, evade immune surveilance etc. [19], [22]- and developed targeted therapies, which marked the advent of second-generation cancer therapeutics [23].

The above point focuses on a static view of the cancer genome; but the cancer genome is highly dynamic [24], [25]. Oncogenesis is not just a matter of acquiring a set of mutations, but an evolutionary process where continuous acquisition of mutations in individual cells coupled with natural selection acting on the resultant phenotypic diversity sculpt the genotype that we observe. Thus, the cancer genome presents a comprehensive record of the cancer's evolutionary history which has many implications for cancer biology and clinical management:

- The underlying mechanisms that drive the acquisition of mutations (i.e. mutational signatures) can provide insights into the etiology of the disease and potentially identify new risk factors [26], [27].
- The order in which driver mutations are acquired could enable us infer the events that took place during different stages of cancer development, better understanding genes functions by putting them in context [28], [29].
- The clonal relationship between different subpopulations of cancer cells could inform us about the cancer's pattern of growth. For instance, a sample from the blood could potentially detect circulating tumor cells (CTCs) or cell-free DNA (cfDNA) that shed from the primary tumor or metastases, and inform us about the tumor's metastatic potential [30]. Samples from different spatial locations of the primary lesion could inform us how the tumor or early progenitors colonize the tissue [31], [32]. Samples

from matched primary and metastatic sites could inform us the path that the tumor took to spread.

- The clonal architecture of the tumor could enable us to infer the timing of critical events in the tumor's history, such as acquisition of invasive potential, metastasis, developing resistance to therapy, etc. in relation to genomic events [33], [34], [35], [36], [37].

In the meantime, the cancer genome at the population level also forms the blueprint for its future development and enables us to apply precision medicine:

- The clonal heterogeneity of tumors as inferred by genomic diversity has been shown to correlate strongly with patient outcome [38], [39], [40].

- The clonal status of actionable driver mutations could inform us about the potential of targeted therapy [41].

2.2 Development of sequencing technologies

It took major developments in DNA sequencing technologies to go from a static and limited snapshot of the cancer genome to the dynamic and comprehensive view of all the evolutionary processes taking place in the cancer genome that enabled the above insights. I will present a personal view of this history: as I was writing up a part of Section 4 of this dissertation, I looked up papers back in the 1980s/1990s where the critical cancer genes that we are conveniently surveying today were first discovered. They were using some of the most basic genetic techniques including mutant-enriched polymerase chain reaction (PCR)-restriction fragment length polymorphism analysis and allele-specific

oligonucleotide hybridization [42]; allelotyping using microsatellite markers [43]; representational difference analysis (RDA) for detection of homozygous deletion [44] ; and direct sequencing of candidate genes [45]. Personally, these methods appeared ingeniously devised and are as exciting to read as detective stories, much more fun than many complicated sequencing methods these days. But they also appeared so primitive that I must marvel at how much we have progressed in merely two decades or so.

Then came Sanger sequencing, which was the first method to sequence the entire human genome [46]. A global genomic analysis in human pancreatic cancer was made possible by this technique in 2008 [47], which not only validated alterations in driver genes already known by that time such as KRAS, TP53, CDKN2A (P16/MTS1), and SMAD4 (DCC/DPC4), but also identified alteration, mostly point mutations, in many new genes corresponding to a core set of 12 cellular signaling pathways and processes. This seminal work set the stage for many more molecular studies of pancreatic cancer to come.

Although Sanger sequencing for pan-genome analysis was a major breakthrough, it was slow and expensive that the study was only able to profile 24 pancreatic cancer cases. It was also targeted as primers needed to be designed. Next-generation sequencing (NGS) technologies which came about in the mid-2000s [48] was much faster and cheaper and enabled sequencing of the genome in an unbiased way in large cohorts of cases [49].

It was NGS that enabled evolutionary analysis of the cancer genome: it enabled detection of more passenger mutations and the many other alteration classes, such as copy number variations (CNVs), structural variants (SVs), mutational signatures, which were all useful for inferring evolutionary processes [50].

The next development was going from bulk (lumping signals from many cells) to single-cell sequencing, which would be discussed in detail in the next Section. But aside from that, bulk sequencing technologies with higher sensitivity were also developed, such as duplex sequencing [51], [52], which enabled detection of rare mutations and mutational signatures that were previously obscured by sequencing errors. These were essential for not only studying rare, potentially resistance-driven subclones in cancer, but also for studying background mutation rate, clonal mutations in normal tissues.

2.3 History of pancreatic cancer genomic studies

In recent years, several large-scale sequencing cohorts have been published, each with its own unique characteristics such as sample sizes, sequencing methodologies, and computational innovations [34], [47], [53], [54], [55], [56], [57]. The seminal Jones et al., Science (2008) utilized high-throughput Sanger sequencing of 20,661 protein-coding genes and confirmed the presence of the four main drivers of pancreatic ductal adenocarcinoma (PDAC): KRAS, CDKN2A, TP53, and SMAD4. This study also introduced the concept of core signaling pathways which brought attention to convergent evolution through inactivating alternative genes. Although the sensitivity was high in this project thanks to the use of xenograft, the limited sample size of 24 pancreaticobiliary cancers hindered the identification of rare driver genes.

Subsequent efforts employed whole-exome sequencing with larger sample sizes, leading to the discovery of additional signaling pathways and novel driver genes. These studies also highlighted the significance of KRAS wild-type PDAC for potential targeted therapies [55]. Whole-genome sequencing, conducted on even larger sample sizes, revealed

additional candidate driver genes and a wide range of genomic instability in PDAC [56]. The most unstable genomes were found to be highly correlated with inactivation of double-strand break repair genes and sensitivity to cisplatin.

By combining whole-genome sequencing with laser capture microdissection, researchers achieved the highest sensitivity for identifying alterations in the PDAC genome [34], [57]. This approach provided valuable insights into the roles of chromothripsis and whole-genome duplication (WGD) in cancer progression. As the sample size and depth of coverage increased, the four main driver genes identified prior to the era of next-generation sequencing remained prominent. However, these studies have provided a much greater understanding of mutational targets, genome structure, and mechanisms of somatic alteration in PDAC.

2.4 General features of the pancreatic cancer genome

Single-nucleotide variants

The most prevalent form of genetic alteration in PDAC are single-nucleotide variants (SNVs). The most reliable estimates based on whole-genome sequencing of macro-dissected PDACs indicate ~2.64 mutations per megabase [56]. The number of SNVs per megabase also vary by the region of the genome studied as non-coding regions have more SNVs per megabase than coding regions, possibly due to differential selection pressures between these genomic regions [58].

A major part of these SNVs are nonfunctional mutations, either targeting the intron, or are synonymous/silent, and are deemed passenger mutations that do not provide a fitness advantage to the cell . Missense mutations that result both in a non-conservative change in the encoded amino acid and within a highly conserved region of the peptide are more likely to change protein function [22], [59], [60]. They are also the most common type of functional mutations targeting driver genes. Aside from that, nonsense mutations and small insertions and deletions (indels) that lead to frameshifts are deleterious by introducing a premature stop codon, which leads to degradation of the mutant transcript by nonsense-mediated decay or in some instances expression of a truncated protein [61]. Splice site mutations might have several consequences, including alternative splicing or intronic retention, that could also lead to formation of a downstream premature stop codon within the aberrant transcript [62].

While they do not directly affect protein function, passenger mutations in cancer cells can still have implications for cancer biology: they could generate novel amino acid sequences that can bind to MHC molecules as neoantigens [63]. PDACs with high mutational burdens, such as those associated with *MLH1, MSH2, MSH6*, or somatic POLE mutations, could be particularly sensitive to immunotherapy [64], [65].

The non-coding genome of PDAC has fewer cancer driver events compared to the coding genome. In one of the few studies of the non-coding PDAC genome, recurrent mutations in cis-regulatory promoter regions were found to be associated with alterations in gene expression related to transcriptional regulation [66]. The genes affected - *RUNX3, ROBO1, SLIT2, and CTNNA2* - had been previously implicated in PDAC [67]. Through integration

of DNA methylation and mRNA expression analysis, 96 genes have been identified as silenced by methylation in PDAC including known genes *CDKN2A, CDKN2B, BRCA1* or *MGMT* and novel PDAC candidates *ZFP82* and *PARP6* [55]. Additionally, miRNA clusters and long non-coding RNAs (lncRNAs) have been found to be differentially expressed in PDAC [55], [68].

Mutational Signatures

Somatic mutations in PDAC can be clustered into different mutational signatures representing different etiologies. The most prevalent mutational signatures in PDAC are related to ageing, smoking, defective DNA repair, dysregulated APOBEC enzymes, and reactive oxygen species (ROS) [27], [69], [70]. Mutational signatures generally are similar between matched primary and metastatic tumors [69], [70], while differ between coding and non-coding regions [71].

Copy number and structural variations

Copy number variations (CNVs) and structural variations (SVs) are common in PDAC. WGD has been identified in 45% of microdissected PDACs that underwent whole-genome sequencing and is associated with a higher number of CNVs and TP53 somatic mutations [34]. Although copy number variation in PDAC might occur in isolation, these events are more commonly associated with other SVs in which part of one chromosome is translocated to another. One such common mechanism by which this occurs in PDAC is chromothripsis, a phenomenon in which multiple SVs occur in a single catastrophic mitotic event [72]. Highly sensitive methods of detection of chromothripsis indicate that it can be

found in as many as 65% of PDACs, often before polyploidization [34]. Alternatively, ongoing genomic instability due to DNA repair deficiency might also lead to structural rearrangements [56]. A study of structural rearrangements in PDAC indicated that most (>80%) are intra-chromosomal. The majority of these are rearrangements (58%), followed by fold-back or amplified inversions (25%) and deletions (13%). Inter-chromosomal translocations are less prevalent (<20%) and duplications of large genomic regions (including tandem duplications) are relatively uncommon, accounting for ~3% of events.

2.5 The mutational landscape of the pancreatic cancer genome

The genes somatically altered at high frequency in PDAC indicate the cellular pathways whose dysregulation is selected for a survival advantage during pancreatic carcinogenesis. I briefly summarize the genes/pathways below:

High-frequency somatic genetic alterations

1. *KRAS* mutations

KRAS activation is among the earliest genetic events known in PDAC, in which it signifies the transition from a normal centroacinar or ductal cell to an initiated cell [42], [73]. *KRAS* is a 21 kDa GTPase that activates MAPK–ERK signalling, thus controlling cellular proliferation, differentiation, migration and survival [74]. *KRAS* mutations are the most common oncogenic alteration in PDAC, occurring in ~90% of cases [47], [56], [57], [57]. Given the overwhelming evidence supporting the role of chronic pancreatitis in pancreatic carcinogenesis [75], hyperactivity of MAPK–ERK signalling seems a requisite to maintain survival of a cell within the inflamed microenvironment. *KRAS* mutations have been shown

to increase cellular fitness by protecting against inflammation-associated senescence and promoting autophagy, micropinocytosis and stress granule formation [74]. In human tissues *KRAS* mutations are found in low-grade pancreatic intraepithelial neoplasia (PanIN), and some PanINs might be associated with acinar to ductal metaplasia [32], [76], [77]

Virtually all *KRAS* mutations in PDAC are SNVs occurring in codons 12 (~91%), 13 (~2%) and 61 (~7%). Subsequent to these activating mutations, allelic imbalance in association with WGD and tumour progression might occur, further increasing the dosage of the mutant *KRAS* allele [78]. In some patients more than one mechanism of *KRAS* allelic imbalance is seen within different metastatic sites, indicating convergence on a net gain in MAPK–ERK signalling [70]. This finding has been independently observed in mouse models of PDAC [76]; furthermore, in the same study compelling evidence was shown for *KRAS* allelic imbalance occurring in the setting of *CDKN2A* homozygous deletion, whereas heterozygous loss of *CDKN2A* was associated with alternative oncogenic events affecting *KRAS* [79]. These findings indicate that PDACs undergo distinct evolutionary routes to increase *KRAS* dosage depending on the mechanism of inactivation of *CDKN2A*.

About 10% of PDACs do not have an activating mutation in *KRAS* [55]. These cases are notable for mutations or copy number alterations in alternative drivers such as activating mutations or amplifications of *BRAF, FGFR1 or ERBB2*, inactivating mutations in *NF1, DUSP6 or SPRED1* [54] or fusions involving *NRG1* and *NTRK1* [80], [81], [82], further underscoring convergent evolution towards the MAPK–ERK hyperactivity phenotype. Wild-type *KRAS* PDACs also seem to be enriched in patients with germline mutations in known familial risk genes [55].

2. *CDKN2A* inactivation

CDKN2A is a tumour suppressor gene whose protein product controls the G1/S checkpoint. *CDKN2A* encodes two proteins: the INK4 family member p16 (or p16INK4a) and p14ARF [83]. Whereas p16 arrests the cell cycle in the G1 phase by inhibiting binding of CDK4 or CDK6 with cyclin D1, p14ARF initiates p53-dependent cell cycle arrest. The mutational profile of *CDKN2A* or *CDKN2B* indicates that p16 is the most likely target, as a substantial proportion of mutations do not affect the p14ARF coding region [84]. Inactivation of *CDKN2A* is found in 90% of PDACs by multiple mechanisms, each occurring in approximately equal proportions: homozygous deletion, mutation coupled with loss of the wild-type allele, or hypermethylation [55], [56], [77]. In PDACs in which *CDKN2A* is not inactivated, alternative mechanisms of inhibiting the G1/S checkpoint — including RB1 inactivation by somatic mutation or hypermethylation, CDK4 amplification or CCND1 amplification — have been identified, indicating convergence towards the phenotype of loss of G1/S checkpoint [54], [85].

3. *TP53* inactivation

Unlike *KRAS* and *CDKN2A*, which are almost always targeted during carcinogenesis and before invasion into the pancreatic parenchyma, *TP53* and *SMAD4* inactivation are relatively late events that, when identified, signify the presence of a neoplasm with invasive potential [31], [86]. *TP53* is a tumor suppressor gene whose protein product serves as a major guardian of genome integrity by modulating transcription, DNA repair, genomic

stability, cell cycle control and apoptosis [87]. Alterations of *TP53* in cancer occur in 80% of PDACs, the majority of which are missense mutations in association with allelic loss that confer gain of function via altered DNA binding and interactions with other transcription factors. Consequences of these gain-of-function mutations include cell cycle activation, loss of apoptosis regulation and metabolic changes; these phenotypes are most likely selected for to maintain survival in association with an increasingly unstable genome during clonal expansion within the pancreatic ducts [31]. A subset of PDACs exhibit TP53 loss-of-function mutations via truncating mutations or homozygous deletion [88]me. Although these mutations also confer an aggressive phenotype, the specific mechanisms have been relatively less explored [89], [90]. TP53 alterations in PDAC are highly correlated with polyploidization, and, although they do not seem to specifically cause WGD, might maintain cell survival in the presence of it [91]Shu.

4. *SMAD4* inactivation

SMAD4, also a tumour suppressor, is a mediator of the canonical TGFβ signalling pathway that controls tissue homeostasis throughout human body [92]. Inactivation of *SMAD4* occurs in just over 50% of resected PDACs by homozygous deletion or somatic alteration with loss of the wild-type allele [43]. About 10% of PDACs have inactivating *TGFBR2* mutations or deletions that are mutually exclusive with *SMAD4* loss. Inactivating mutations in other members of the TGFβ family of receptors are also observed [93]. Loss of *SMAD4* promotes growth by loss of tumor-intrinsic responsiveness to canonical TGFβ pathway signalling, abrogating growth control while resulting to increased migratory behaviour, immune evasion and autocrine activation [92]. *SMAD4* inactivation has been shown to be

a subclonal event (only present in a subset of tumor cells) and roughly coincide with the moment of invasion, indicating that it is selected for to maintain cell survival and fitness upon exposure of the neoplastic cells to the stromal microenvironment [94], [95]. Considering that the presence or absence of *SMAD4* expression also correlates with the mode of invasion — that is collective or mesenchymal, respectively [94]— it is likely that specific features of the stromal microenvironment that are yet uncharacterized act as a potent selective pressure that favours neoplastic cells with *SMAD4* loss.

Low-frequency somatic genetic alterations

Many genes are recurrently altered in PDAC but at frequencies lower than 10% of analysed cases in large-scale studies. Emerging data suggest some might serve as biomarkers of PDAC subtypes with differing therapeutic susceptibilities [96], [97].

SWI/SNF and COMPASS complexes

Two chromatin remodelling complexes are affected by inactivating mutations in up to 10% PDAC - the Switch/Sucrose-Nonfermentable (SWI/SNF) complex and the COMPASS complex [98], [99] - indicating selection towards epigenetic remodeling.

Genes targeted in the SWI/SNF complex include *ARID1A* and *ARID1B*, both of which are components of the BAF subunit, or *ARID2* and *PBRM1*, both of which are components of the PBAF subunit. *SMARCA2* and *SMARCA4*, present in both the *BAF* and *PBAF* subunits, might also be affected in a minority of PDACs [55]. *ARID1A* is genetically inactivated at

a slightly higher rate (~6% of PDACs) than all other genes in the SWI/SNF complex, indicating a more general importance for disruption of the BAF subunit. Further, mutations in any one SWI/SNF gene are mutually exclusive of each other, indicating convergence [100]. In normal cells the SWI/SNF complex contains DNA-stimulated ATPase activity and utilizes the energy provided by ATP hydrolysis to remodel chromatin through nucleosome sliding and removal [101]. These distinct SWI/SNF family complexes bind to a multitude of genomic loci, including distal enhancers, promoters and CCCTC-binding factor binding sites at which they facilitate and maintain DNA accessibility to regulate gene transcription [102]. Importantly, they are significantly intertwined with the TGFβ signalling pathway which is frequently altered in PDAC as described above.

Genes targeted in the COMPASS complex include the histone H3 lysine 4 (*H3K4*) methyltransferases *KMT2C* and *KMT2D* and the complex-related gene *KDM6A*, which acts as a H3 lysine 27 (*H3K27*) lysine demethylase [103]. As for genes of the SWI/SNF complex, mutations in any one of these three genes are mutually exclusive of each other [56]. These proteins serve as major regulators of monomethylation at enhancer regions, indicating convergence towards a phenotype of enhancer attenuation [58].

Genetic inactivation of genes in the SWI/SNF or COMPASS complexes is correlated with the development of basal-like transcriptional features, particularly if this event occurs early in the evolutionary life history of the neoplasm [104]. It demonstrates that attenuation of enhancer activity is an important factor in the development of basal-like transcriptional features and worse outcome [105], [106].

GATA6 and *MYC*

GATA6 and *MYC* are recurrently amplified in PDAC and usually associate with whole-genome duplication or chromothripsis [107]. *GATA6* is a transcription factor that contributes to the normal development of mesodermal and endodermal tissues including the pancreas . In human PDAC cells and tissues *GATA6* amplification, or its transcriptional upregulation, occurs late during carcinogenesis and activates canonical WNT signaling [108]. *GATA6* amplification or overexpression correlates with the classic transcriptional phenotype as well as with improved overall survival [96], [97]. *GATA6* amplification is also correlated with *SMAD4* deletion, in part because they are in close proximity to each other on chromosome 18 and are likely altered simultaneously [107]. On the contrary, in some PDACs *GATA6* might instead undergo hypermethylation-mediated silencing, which is correlated with basal-like features of PDAC and with poor outcome [107], [109].

MYC is part of a family of transcription factors that form a network to regulate metabolism and cell proliferation and induce expression of genes required for these processes [110]. Unlike *GATA6*, which is subject to either gain or loss of expression, *MYC* is usually only upregulated by gene amplification[47], [55], [56], [57]. It promotes cancer progression though altered metabolic pathways, survival signals in the setting of hypoxia, and promotion of cell competition [111]; it is inversely correlated with basal-like features of PDAC and poor outcome [96], [97].

Germline alterations

The best-characterized germline variants linked to PDAC are BRCA1, BRCA2 and PALB2, the Fanconi anaemia genes FANCC and FANCG, and ATM. Mutations in BRCA1 and BRCA2 are the most commonly identified germline variant, accounting for almost 50% of

all germline mutations found [112], [113]. These proteins are all components of the DNA double-strand break repair machinery and mutations in them cause increase genomic instability by faulty homologous recombination at stalled replication forks and increase the rate of somatic mutations, as seen by the signature SBS3 and ID6 [27].

2.6. Pancreatic cancer genomic evolution through space and time

In the general cancer field, more than half of all single nucleotide variants (SNVs) found in cancerous epithelial cells have been observed to occur prior to the initiating event [114], mostly as a result of random DNA errors that accumulate over time [26]. Oncogenic mutations in the *KRAS* gene, the most common initiating genetic alteration in pancreatic cancer, can also occur randomly and commonly in phenotypically normal pancreatic cells or in samples of chronic pancreatitis from otherwise healthy individuals [32]. This suggests that *KRAS* mutations alone are not sufficient for the development of invasive cancer.

The presence of both acute and chronic pancreatitis has been found to significantly increase the risk of developing pancreatic ductal adenocarcinoma (PDAC). In humans, many of the known risk factors for PDAC are associated with chronic inflammation [115]. It has been shown that activated *KRAS* creates a feedback loop that that promotes cell survival by inducing a pro-tumor microenvironment while downregulating other cellular pathways that limit formation of acinar to ductal metaplasia (ADM) [116], [117], [118], [119]. Therefore, activate MAPK-ERK signaling increases the fitness of cells to withstand the inflamed environment.

Then, as genetic damage accumulates due to one or more mutational processes (e.g. ROS due to inflammation, ageing, smoking), additional alterations, such as in the *CDKN2A* gene, can occur, further increasing the fitness of the mutant clone. The corresponding phenotype of loss of the G1/S checkpoint perhaps increases genomic instability by causing catastrophic events such as chromothripsis [79], which might together select for additional alterations that allow the cells to sustain such genomic instability, such as *TP53* inactivation which limits apoptosis in such context [87]. *TP53* alterations are highly correlated with cells with high-grade cytological features (PanIN3), increased proliferation and the capacity to migrate through the pancreatic ductal system [31], at which point the cells are very close to invasive cancer,

The pancreatic stromal response to invasive carcinoma is a crucial factor in pancreatic carcinogenesis, and is an expansive topic that I will not go into much detail here. I found it worth highlighting that the stroma response to epithelial injury in the normal pancreas include fibroblast activation, immune suppression, remodeling of the extracellular matrix and trophic signals to promote re-epithelialization, all features coordinated in large part by TGFβ [120], [121]. This drastically changing microenvironment forms an additional round of selection force that my research lab postulate is the primary reason for the enrichment of TGFβ inactivation, signified by loss of *SMAD4* (occurs in >50%) which is a central mediator of cell-intrinsic TGFβ signaling [43], [57]. Notably, it has been shown that TGFβ inactivation is a subclonal event, indicating that it is not selected for during early oncogenesis but rather during the invasion process [94], [95].

Mutational processes continue as the invasive cancer develops which results in clonal heterogeneity. Whole-genome duplication, with or without chromothripsis, further adds to this heterogeneity [34]. This diversity manifests as variable gene dosages that are adjusted for a net survival advantage in the context of the immediate microenvironment and available nutrients [79]. As a result, by the time a PDAC is clinically evident it contains demonstrable intratumoural heterogeneity characterized by subclones that differ with respect to somatic alterations, allelic gains or losses and gene deletions or amplifications; in some patients with resectable PDAC these subclones contain mutations or amplifications of known driver genes that are associated with disease recurrence after adjuvant therapy [70]. By contrast, sequencing studies of treatment-naive stage IV PDAC failed to identify any genetic heterogeneity, indicating these two patient cohorts differ by at least one major clonal expansion [36], [122]. Ultimately, despite the complexity of a primary PDAC genome, focused studies of the genomic alterations associated with metastasis indicate that metastatic efficiency is largely established by the driver genes that accumulate during carcinogenesis [123], after which epigenomic and metabolic perturbations play a greater role [124].

2.7 Limitations of bulk NGS

Elucidation of genomic events/processes presented above has largely been accomplished by sequencing of bulk tissues, which means genomic signals from thousands of cells are lumped together and analyzed as a whole. While this approach has been highly successful

in identifying the major genetic drivers and general genetic features, it is subject to several shortcomings:

For one, pancreatic cancer stands out with a uniquely extensive stroma reaction. This entails that most samples taken from the primary/metastatic site consist of a significant amount of non-tumor cells, which dilutes the signal from the tumor cells. This is particularly problematic for detecting rare genetic events, such as subclonal mutations or copy number variations, which could be important for understanding the evolutionary history of the tumor. In the worst case there can be no tumor cell captured from a precious sample at all. Present solutions include multiregional sampling [125] or laser-capture tissue microdissection to enrich for tumor content [57], but they are laborious and not amenable to high-throughput.

Another point is generally, given clonal evolution happens at the single-cell level, with bulk sequencing data one must rely on a set of strong assumptions which may overlook important events such colocalization of alterations within a singular genome or mutual exclusivity indicating distinct populations.

3. My perspective and objectives of the thesis

3.1 The need for single-cell resolution

The field of single-cell genomics, since its advent about 10 years ago [126], has been striving to increase the throughput and resolution of cancer research. Single-cell DNA sequencing (scDNA-seq) offers many advantages over traditional "bulk" DNA-seq. Most importantly, it circumvents the issue of "mixed signals" [127], i.e. the admixture of normal and tumor cells, and/or of distinct tumor subclones. Solving the former allows for much higher sensitivity in calling rare genetic events, which opens opportunities to validate and discover cancer-related somatic mutations. Solving the latter allows for more confident identification of different clonal lineages within a single tumor, which could inform understanding cancer evolution as well as targeted treatment decisions.

Being able to separate tumor cells and other cells directly, single-cell sequencing carries an additional layer of technical advantage for studying pancreatic cancer which is characterized by a dense stroma reaction.

3.2 The need for spatial resolution

Early/large-scale genomic studies have usually used solely one piece of tissue to represent each tumor. However, as described above, tumor is a spatially heterogeneous population and needs to be studied across varying spatial locations. There are two levels of spatial resolution - at the micro-level, one could seek to profile single cells in sitro in their native tissue context, which would provide information on the spatial distribution of different cell

types and their interactions within tissue structures such as the pancreatic ducts or the stroma. At the macro-level, one could seek to profile different regions of the tumor, which would provide information on the spatial distribution of different clonal lineages and their interactions. The first has seen major development in the past few years in basic research, but I would argue that the second is more relevant for clinical practice and would be a major focus of my thesis. The TRACERx series of studies in lung have already made a great demonstration of the power of adding a spatial dimension to sampling: the spatial heterogeneity across different subregions of the primary tumor, between primary and metastasis, and potentially to be captured by circulating tumor DNA (ctDNA) surveillance, can be used for predicting many clinical outcomes such as disease recurrence, metastasis, and treatment response [39], [40]. My lab has established a large library of such multiregionally-sampled pancreatic cancer cases through research autopsy programs [125].

3.3 The need for temporal resolution

Similar to the spatial dimension, the temporal dimension of pancreatic cancer evolution has been understudied in the past, due to the lack of appropriate sampling. For one part this is because unlike other tumor types where patient survival is long and biopsies can be take at multiple time points (treatment-naive, neoadjuvant-treated, resected, relapsed), pancreatic cancer is usually diagnosed at a late stage and patients deteriorate so rapidly that biopsies at multiple time points are relatively more difficult [7]; another reason is simply there has not been such initiative as it requires lots of clinical coordination. As we are gradually entering the era of precision medicine for pancreatic cancer, the temporal

resolution is particularly valuable for studying cancer evolution through treatment bottlenecks and devising strategies to overcome resistance [128].

3.4 The need for a clinically-applicable workflow to apply the high-resolution technologies

Last but not least, while I acknowledge that single-cell technologies are flourishing for basic science, many have been limited to studying cell lines or model organisms [129]. There are a few reasons:

(1) The technologies are still expensive and laborious, and require a high level of technical expertise to operate. Therefore, they are hard to scale and achieve statistical confidence over a large population of clinical samples.

(2) The lack of robustness of the technologies do not allow for application on precious clinical samples.

(3) The high noise profile, high complexity of the data limit its use for hypothesis-generating for basic biology studies, less for generating clinically-relevant or actionable insights.

Given the three points, I recognize the necessity of not only increasing the resolution of sequencing methods, but also adapting them into workflows that are (1) scalable (2) robust for precious samples (3) directly relating to clinical questions.

Section 2: Development of a scalable high-throughput single-nucleus DNA sequencing (snDNA-seq) pipeline for pancreatic cancer

This work has been published at *Nature Communications* (2023) [130]. I am adapting a major part of it into this thesis, as I believe that although a lot of the workflow was preliminary and will be updated in the later sections, it shows the exploratory phase of the project thereby completing the story.

1. Optimization of the wet-lab workflow

To date, several high-throughput single-cell partitioning systems have been developed, including microfluidic platforms, nanowells and microdroplets, which have resulted in several reliable single-cell DNA library preparation technologies [129]. Tapestri [131], [132], as a microdroplet-based, targeted sequencing approach, allows for high cell throughput (up to 10,000 cells per sample) and high coverage depth (>80X) of genomic sites of interest, and is therefore suited for high-resolution studies of key genetic variants within diseases. However, its use so far has been limited to cell lines and liquid primary tumor samples, or a small batch of solid tumor tissues at a time [133], [134], [135], [136], [137], [138]. While methods to quickly and effectively dissociate clean single nuclei suspensions from solid tumor tissues have been extensively tested for single-nucleus RNA-seq (snRNA-seq) [139], they have not been applied for single-nucleus DNA-seq (snDNA-seq) usage.

1.1 Development of a single-nuclei extraction protocol

With Ronan Chaligne, Ignas Masilionis, Elias-Ramzey Karnoub, Juan Li, Neeman Mohibullah, and other MSK Integrated Genomics Operations (IGO) members

We recognized the need for a nuclei extraction workflow that has reduced hands-on operation, sample resuspension times, and total processing time, all of which hinder scalability and could potentially cause between-sample inconsistency in quality (clumping, debris) and final yields. Thus, we used an automated nuclei extraction machine [cite] for homogenizing frozen tissues into single nuclei suspensions. The nuclei were then passed through a sucrose gradient to strip away debris before microdroplet encapsulation. The entire procedure takes ~30 min per sample and requires a single step of pelleting and resuspension **[Figure 1a].** Although the resulting nuclei clumping percentage and nuclei concentration varied across samples of different starting sizes, cellularity and morphology, most primary pancreas tumor samples of volumes ≥ 8 mm^3, regardless of collection method (resection vs autopsy), resulted in final nuclei suspensions with ≤10% clumping and >2000 nuclei/µl suspended in > 35 µl buffer as input for Tapestri **[Figure 1b].** Exceptions were samples with particularly high fat/stroma content or random technical errors, which either limited input nuclei concentration or increased clumping rate as observed by microscopy.

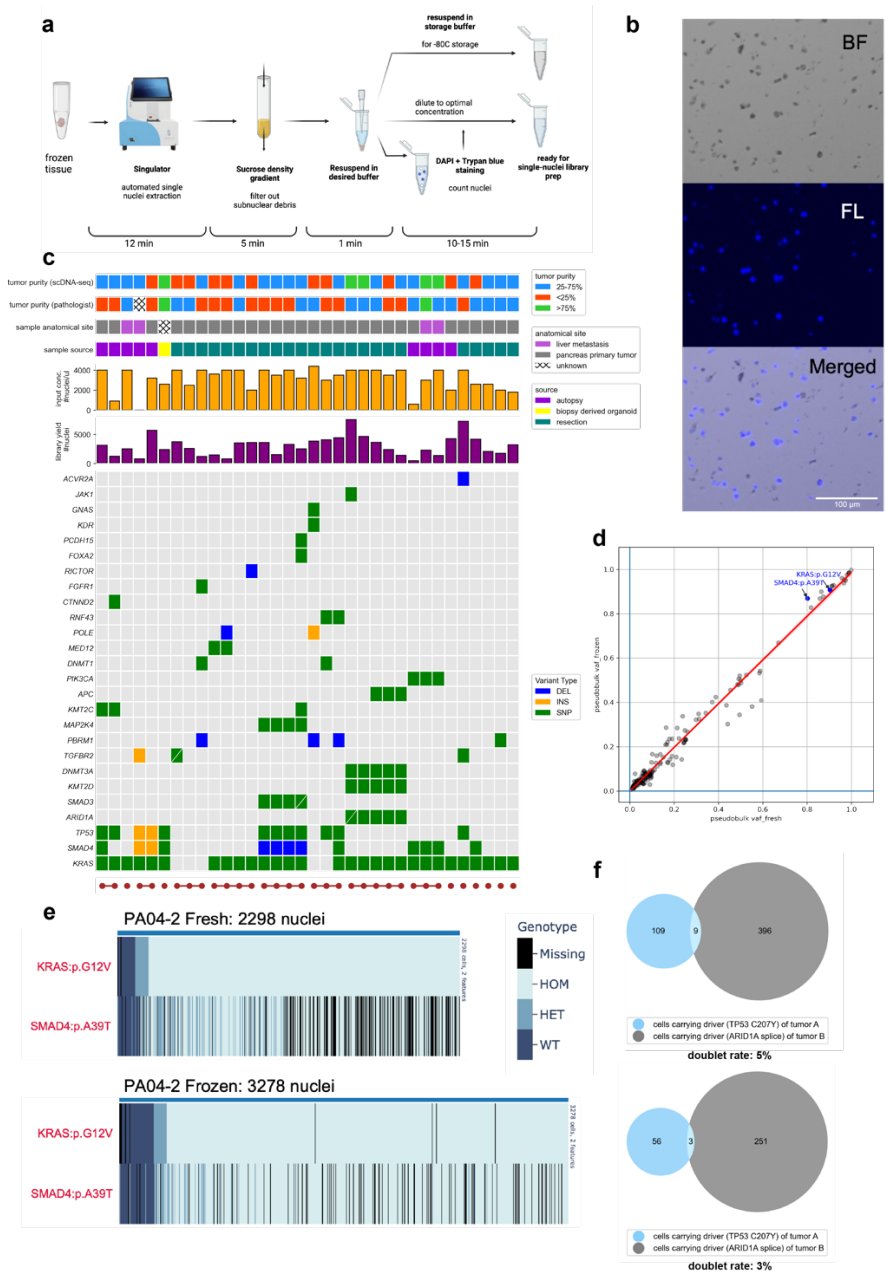

Figure 1. Frozen tissue single nuclei extraction workflow for snDNA-seq.

[a] Overview of frozen sample single nuclei extraction workflow for Tapestri snDNA-seq. [**b]** Representative microscopic images of extracted single nuclei, stained with Trypan blue (brightfield, BF), DAPI (fluorescence, FL). At least 3 representative pictures were taken per sample and yielded similar results. [**c]** Technical and genetic profile of each biologically distinct sample. A total of 38 samples were processed, 34 of which were biologically distinct samples from 16 unique patients. Genetic profiles were based on bulk sequencing. [**d]** Pseudobulk (p-bulk) variant allele frequency (VAF) comparison of all 160 shared variants between libraries prepared with fresh vs frozen (3 weeks) nuclei of sample PA04-2. Key drivers pre-identified by bulk WES in this case are

highlighted; regression line with 90% confidence interval is drawn. **[e]** Single-cell genotype (HOM homozygous mutation, HET heterozygous mutation, WT wildtype) heatmap of snDNA-seq libraries generated by fresh vs frozen nuclei of sample PR04-2. Each row represents a bulk data-validated driver variant of this case, while each column represents a single nucleus in the library. The nuclei were sorted based on *KRAS* variant's VAF in ascending order from left to right. **[f]** Venn diagrams showing colocalization of genotypes belonging to two separate tumor cell populations in one snDNA-seq library (two replicates shown); the cells carrying both genotypes were identified as doublets. Nuclei suspension extracted from two tumor samples from different patients were mixed and subject to snDNA-seq. The two distinct tumor populations were identified by genotype for their respective driver variants *TP53* p.C207Y and *ARID1A* splice. Source data are provided as a Source Data file.

With this protocol, we prepared 38 snDNA libraries from 34 biologically distinct snap-frozen PDAC samples from 16 patients. These samples were purposely selected to represent primary tumors and metastases, different tissue collection methods, and with varying tumor purities **[Figure 1c].** Each was analyzed with a custom 186-amplicon panel covering 93 frequently mutated genes in PDAC. The mean library yield of single nuclei that passed quality control was 2867 complete nuclei (standard deviation = 1672.67). For context, two previous studies [cite] using primary acute myeloid leukemia (AML) cell suspensions for Tapestri resulted in on average 5072 complete cells/sample (146 samples) and 6102 complete cells/sample (154 samples) in the final libraries.

1.2 Quantification of technical parameters and noises

To determine the optimal sequencing depth (in this case, mean reads per cell per amplicon) required to extract significant insights from the snDNA-seq data, we selected two samples that we deemed as over-sequenced (respectively, 278 million and 341 million total read pairs; 168 and 342 mean reads/cell/amplicon) and performed downsampling experiments

by subsampling the raw FASTQ files at intervals of 50 millions read pairs **[Figure 2a, b].** The elbow point could be spotted at roughly 100 mean reads per cell per amplicon for the proportion of DNA read pairs assigned to cells, total number of barcodes captured, and number of cells called **[Figure 2c–e].** However, the proportion of tumor cells, as defined by the orthogonally validated KRAS mutation for each case, did not vary much along with total depth and mean depth per cell per amplicon **[Figure 2f].** This indicates that the relative proportions of major clones and associated biological insights could be robust to varying technical parameters such as read depth and cell throughput. However, for the purpose of rare clone discovery where cell throughput would be critical, the results here suggest at least a mean depth of 100 is required for our panel specifically.

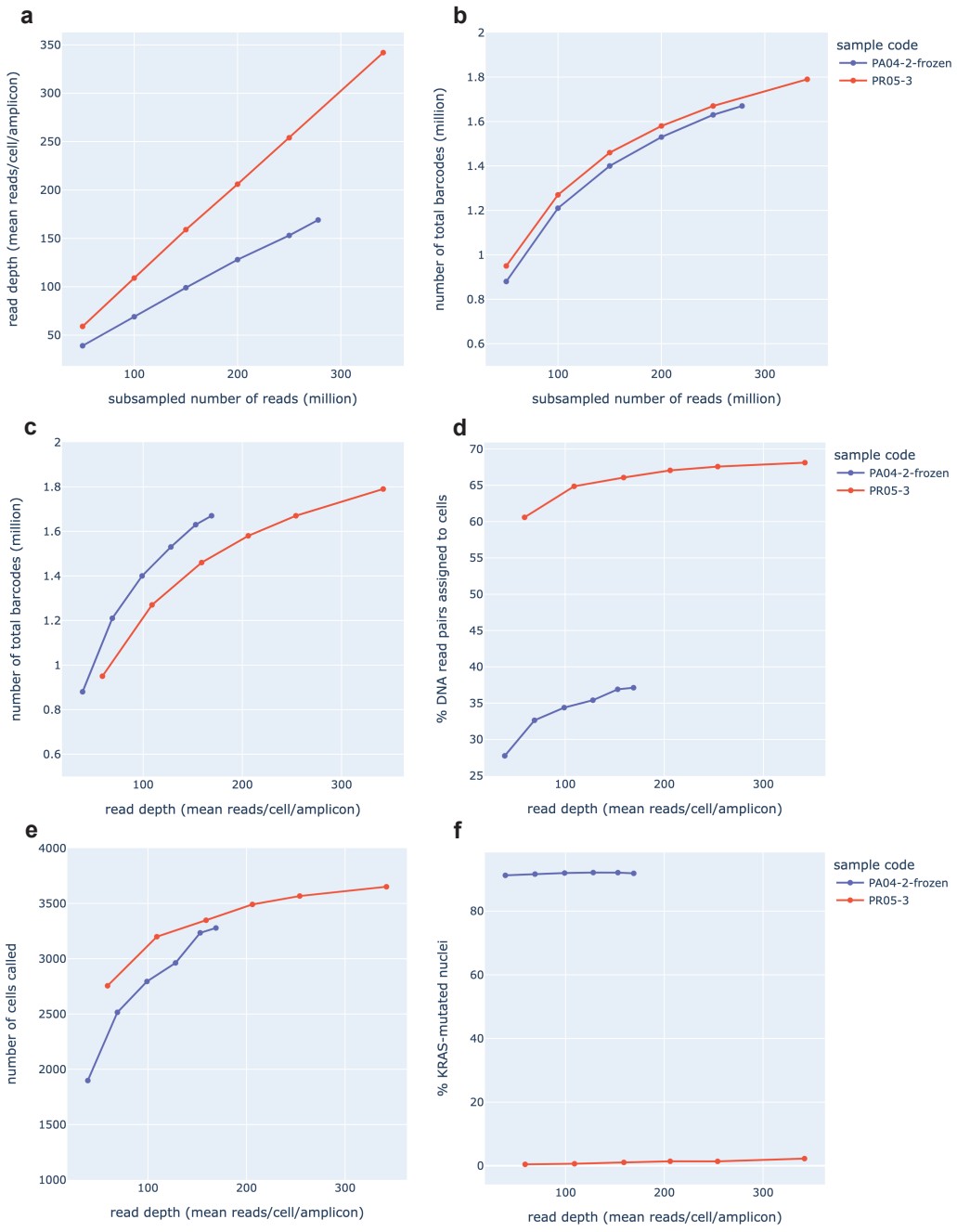

Figure 2. Subsampling experiments to optimize sequencing hyperparameters

Line plots showing correlation between total/mean read depth and several import technical parameters. **a.** total number of reads vs. total number of cell barcodes detected. **b.** total number of reads vs. mean read depth (per cell per amplicon). **c.** mean read depth vs. percentage of DNA read pairs assigned to cells. **d.** mean read depth vs. number of total barcodes. **e.** mean read depth vs. number of cells called. **f.** mean read depth vs percentage of KRAS mutated single nuclei

We next determined the extent to which extracted single nuclei could be stored in suspension without affecting the yield. For two different samples (PA04-2 and PA04-3) we cryopreserved a portion of the extracted nuclei then thawed these frozen nuclei suspensions after 3 weeks and 14 weeks respectively as input for snDNA-seq. This allowed us to compare both the nuclei morphology and the resulting snDNA-seq results between freshly extracted and cryopreserved nuclei of the same biological samples. We found that the freeze-thaw-resuspension process had >80% recovery rate and minimal change in nuclei morphology and clumping percentage Moreover, the final library yield did not decrease when prepared with frozen nuclei; on the contrary, frozen nuclei gave slightly higher yield than fresh nuclei for both samples. Frozen nuclei generated similar quality results as freshly extracted nuclei as measured by sharing the majority of high-quality variants and having largely linearly correlated pseudobulk variant allele frequency (VAF) for all shared variants **[Figure 1d].** For both samples, fresh and frozen nuclei revealed highly concordant genotypes as well **[Figure 1e].**

While the doublet rate inherent to the Tapestri scDNA library preparation method has been estimated to be 5–8% by its manufacturer [140], we sought to estimate the multiplet rate associated with our entire workflow when using snap-frozen tissue. We selected two samples from two distinct patients, each with a tumor population characterized by a distinct driver mutation (TP53 p.C207Y vs. ARID1A splice) that was orthogonally validated by bulk WES as well as independent Tapestri runs. Similar sized pieces of each tissue sample were mixed together and subjected to the entire nuclei extraction workflow followed by snDNA library generation. Next, we assigned each cell to the two originating tumors based

on its genotype for the two driver mutations. The number of barcodes that carried somatic variants from both tumors are as shown in the Venn Diagrams **[Figure 1f].** Using a mixture model we derived the doublet rate to be 3–5%.

As a PCR-based method, our snDNA library preparation is subject to technical noise introduced by uneven amplification of both alleles in a cell (allelic dropout, ADO). To quantify the ADO rate inherent to the workflow, we selected 19 of our samples which came from resection of early-stage PDAC. We identified germline single-nucleotide polymorphisms (SNPs) by comparing single-nuclei sequencing results with bulk-sequencing results of matched normal sample of the same patient, and calculated the mean ADO rate to be 19.6% (of 100 nuclei, 19.6 nuclei only have one of the two alleles). However, we deem this as overestimation as we observed obvious single-nucleus colocalization pattern of ADO of SNPs on different chromosome arms, which more likely suggests real copy number variation rather than technical noise.

1.3 Validation of snDNA-seq results against bulk sequencing results

To validate the robustness of our method for detection of single-nucleotide variants (SNVs), we selected 18 samples previously used for bulk WES sequencing. When limiting to variants present within the targeted panel, a strong concordance was noted between unfiltered variants called from frozen nuclei versus those called by WES **[Figure 3a].** When we next limited to filtered, higher-confidence variants from bulk, there was virtually complete concordance **[Figure 3b].** Moreover, despite technical differences inherent to

WES versus Tapestri (input tissue slice, DNA library prep technology and sequencing depths) the bulk WES VAFs and snDNA-seq pseudobulk VAFs for all shared variants, except for those with ≤2 alternative reads in bulk WES results, were linearly correlated for the majority of samples.

We next sought to determine if our snDNA-seq approach can detect subclonal copy number variations (CNVs). For this analysis we used sample PA02-1 with known homozygous deletions of SMAD4 and CDKN2A [cite Jones et al., Science 2018]. By calculating the single-cell per-amplicon (~200 bp) ploidy based on read counts , we were able to identify both homozygous deletions (ploidy~0) in the neoplastic population despite it comprising 28.5% of all cells **[Figure 3c, d].**

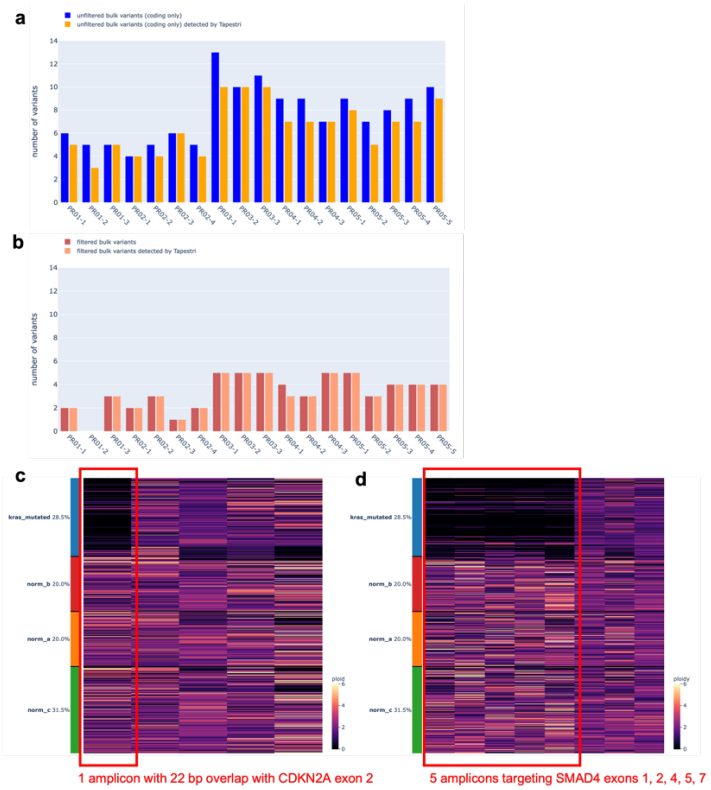

Figure 3. SnDNA-seq can capture SNV and CNV pre-identified by bulk-sequencing with high sensitivity.

a. For 18 samples of the same bulk WES sequencing cohort, histogram showing the number of unfiltered ("Methods"), coding variants called by bulk (blue) and among them, the number detected by snDNA-seq (orange). **b.** For the same samples, histogram showing the number of filtered ("Methods") variants called by bulk (dark red) and among them, the number detected by snDNA-seq (light red). Single-cell per-amplicon ploidy heatmap of select amplicons covering chromosomes 9 **(c)** and 18 **(d)** of sample PA02-1. Each row represents one cell while each column is one amplicon. Cells are divided into *KRAS*-mutated group (blue) and 3 normal groups (red, yellow, green) and hierarchically clustered within group. The amplicons spanning *CDKN2A* and *SMAD4* genes are labeled.

---

2. Feasibility on low-cellularity, low-tumor content samples

Low cellularity and tumor content in certain clinical settings, such as with fine needle aspirations, core needle biopsies or for PDAC in general [141], represent major hurdles for bulk-sequencing and hinder the quality of resulting genomic information. We therefore tested our workflow's performance in such settings.

We identified one sample PA04-1 of a primary pancreas tumor with both low-tumor cellularity (neoplastic cells occupying <50% area of tissue section) and low overall cellularity due to high fat content **[Figure 4a].** Despite use of a tissue sample of similar size (~8 mm^3) as others studied, we extracted a total of 30,000 nuclei, much lower than the optimal number of 200,000 as input for Tapestri. Ultimately, we captured only 479 nuclei in the final library, approximately 6-fold lower than average. Nonetheless, both driver gene variants (KRAS p.G12V, SMAD4 p.A39T) identified by bulk WES of this same sample

were identified with high-quality read data and indicated 113 likely tumor cells (23.6% of all) captured for this sample based on the presence of the clonal KRAS variant **[Figure 4b, c].**

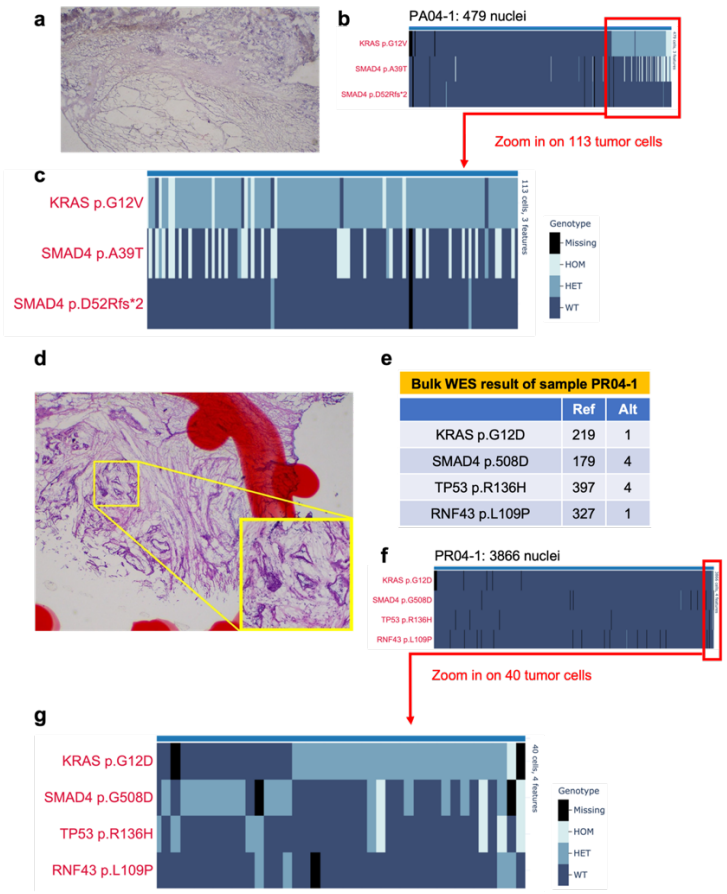

Figure 4. SnDNA-seq's performance in limiting settings.

a. Representative hematoxylin and eosin (H&E) stained histology image of sample PA04-1, a sample of particularly low cellularity due to high fat content. At least 3 representative pictures were taken per sample and yielded similar results. **b.** Single-cell genotype (HOM: homozygous mutation; HET: heterozygous mutation; WT: wildtype) heatmap of sample PA04-1. A total of 479 captured single nuclei were sorted based on *KRAS* VAF in ascending order from left to right. **c.** Single-cell genotype (HOM: homozygous mutation; HET heterozygous mutation, WT wildtype) heatmap of sample PA04-1, zoomed in on 113 putative tumor cells, defined by the presence of *KRAS* p.G12V/*SMAD4* p.39T mutation. **d.** Representative H&E histology image of sample PR04-1. At least 3 representative pictures were taken per sample and yielded similar results. **e.** Bulk WES result of sample PR04-1. **f.** Single-cell genotype heatmap of

PR04-1. A total of 3866 captured single nuclei are sorted based on *KRAS* VAF in ascending order from left to right. Tumor cells, defined by the presence of *KRAS* p.G12D/*SMAD4* p.G508D/*TP53* p.R136H mutations, cluster to the right and are barely visible. **g.** Single-cell genotype heatmap of PR04-1, zoomed in on 40 putative tumor cells.

With a second sample (PR04-1) of particularly low-tumor purity as revealed by pathology review **[Figure 4d]** and bulk sequencing **[Figure 4e],** Tapestri data identified 40 out of 3866 nuclei (1%) of this sample carrying at least one of KRAS p.G12V, TP53 p.R186H, or SMAD4 p.508D **[Figure 4f, g].** Again, the driver variants were genotyped with high-quality read data.

3. Preliminary biological insights to be gained

For a PDAC surgical resection case PR02, based on both MSK-IMPACT sequencing (high-depth targeted sequencing) and bulk WES, we identified that the major tumor clone carried the hotspot KRAS p.G12D mutation; hints of a minor KRAS p.G12V clone existed but were on the borderline of the technologies' detection sensitivity **[Figure 5a].** Single-nucleus genotype heatmaps and Venn diagrams **[Figure 5b–e]** of multiregional samples PR02-3 and PR02-4 suggested colocalization of the major KRAS p.G12D with another likely driver TP53 p.C203Y, which signified the major tumor clone in this sample; the minor KRAS p.G12V -bearing clone was mutually exclusive with the above two drivers and did not colocalize with any known driver gene mutation at similar clonal frequency. In PR02-4, while the major clone consisted of 221 cells (6.12% of all cells), the minor clone was only 12 cells (0.33% of all cells, **Figure 5e**), further buttressing the technology's

sensitivity. The minor KRAS p.G12V clone in both samples was substantiated with high-quality read data; digital droplet PCR (ddPCR) also proved the presence of the KRAS p.G12V [130]. Pathology review did not identify any apparent secondary neoplastic or metaplastic structure **[Figure 5f].** This observation aligns with several other studies [55], [142] in suggesting that multiple KRAS genetic variants may coexist in one patient's PDAC precursor/tumor.

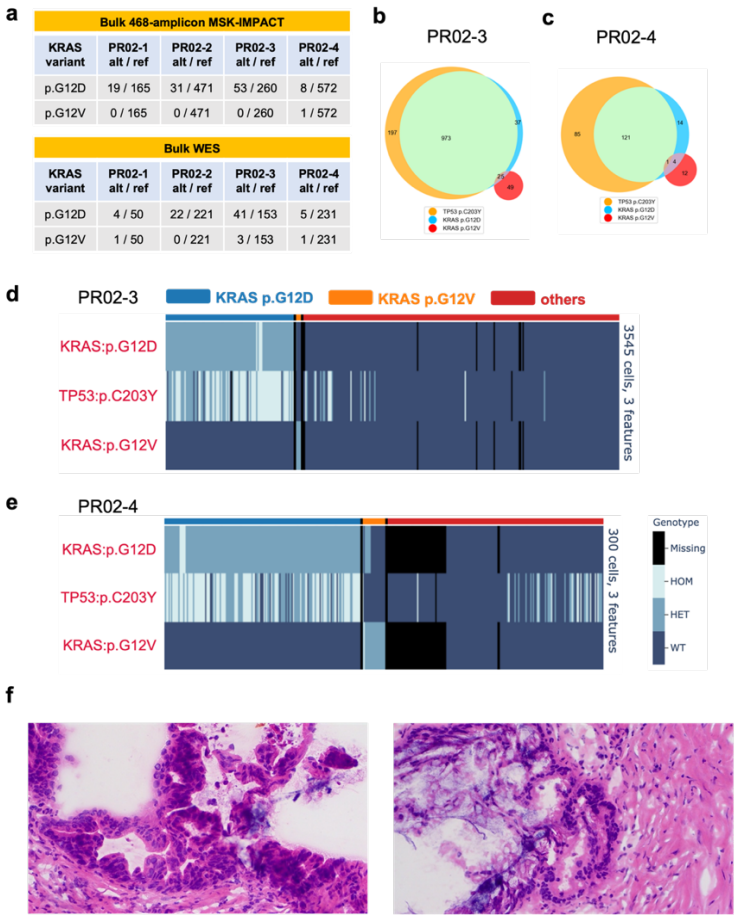

Figure 5. SnDNA-seq identified two mutually exclusive clones bearing two different KRAS mutations in one pancreatic cancer patient.

a. Bulk-sequencing calls of *KRAS* variants of patient PR02's 4 multiregional primary tumor samples. Venn diagram showing colocalization pattern of genetic variants *TP53* p.C203Y, *KRAS* p.G12D, *KRAS* p.G12V in single nucleus of samples PR02-3 **(b),** PR02-4 **(c).** Single-cell genotype (HOM homozygous mutation, HET

heterozygous mutation, WT wildtype) heatmap of samples PR02-3 **(d),** PR02-4 **(e, zoomed in on 300 cells where tumor cells cluster).** The *KRAS* p.G12D & p.G12V clone identities (above heatmaps) are identified as cells having "HET" or "HOM" genotype of each variant and are colored as labeled. Cells are hierarchically clustered based on the two KRAS variants' single-cell AF. *KRAS* p.G12V clone size was 57 cells in PR02-3, 17 cells in PR02-4. **f**. Representative H&E histology images of sample PR02-3. At least 3 representative pictures were taken per sample and yielded similar results.

---

4. Discussion

In this preliminary study, we used two commercially available products to assemble a highly automated workflow to generate Tapestri snDNA-seq libraries from snap-frozen patient tissues. The workflow is fast, efficient and can be applicable for high-throughput clinical and translational research. Additionally, we recognized the value of storing excess single-nuclei suspensions for later use and verified a corresponding workflow that was free from issues such as nuclei quantity loss, nuclear envelope damage, or nuclei clumping. This workflow further illustrates the ability to maintain information pertaining to relative VAFs compared to matched WES data. Furthermore, while our custom panel was not designed for identification of CNVs, specifically homozygous deletions, we demonstrate the proof of principle that these events could be identified with confidence. These encouraging results, combined with the added information gleaned pertaining to co-occurring and mutually exclusive genetic events, suggests this technology and workflow is ideally suited to settings in which samples sizes are small or limited.

A caveat of this snDNA-seq technology is that albeit its high cell throughput, it is a targeted approach that focuses on a pre-designed panel of genes which might be insufficient for

certain research questions that require unbiased study of the cancer exome/genome [143], [144]. This also implies a general challenge of single-cell research: because the total library size is limited by sequencing cost and data processing capacities, there is always a trade-off between the total cell throughput and the amount of information one can extract from each single cell.

Sequencing depth, usually defined as the mean number of times any nucleotide is read, is a straightforward and critical metric for bulk sequencing- it can be easily controlled by adjusting the total sequencing depth of a sample and decides the accuracy of estimating relative clonal composition of SNVs based on variant allele frequencies (VAFs) as well as the power of discovering rare SNV clones. Similarly, for Tapestri we calculated depth as the mean number of reads per amplicon per cell and determined the optimal depth. However, because reads need to be first demultiplexed and assigned to individual cell barcodes, and then undergo barcode quality control (QC or "cell-calling", "Methods"), the depth is affected by not only the total read depth, but also the number of single cells that pass QC and the proportion of reads assigned to these cells, which could be panel- and sample-specific. For example, at the same total read depth, a sample with more incomplete cells/nuclei (e.g. autopsy samples with high necrosis rate) would have a lower percentage of reads assigned to quality cells and therefore lower depth. Therefore, it is difficult to precisely control depth through altering the total depth; it might be advisable to sequence to a total read depth that is economically feasible first, inspect the data, and sequence more if needed.

Another caveat, in this instance related to our computational analysis used for this published project, is that the current variant calling pipeline for Tapestri applied GATK HaplotypeCaller and enabled default read downsampling which might be suboptimal for such high-depth single-cell data in a cancer setting. We did not apply matched normal/panel of normal (PON)-aided filtering in our pipeline; a benefit of this approach is that it enabled us to see many germline variants that can be used for quality control or phylogeny modeling, although this benefit is balanced by outputs containing many artifacts. To enable novel somatic variant discovery, a more robust variant calling pipeline is needed to adapt to the noise profiles particular to this PCR-based high-throughput single-cell DNA sequencing technology. This will be addressed in **Section 3.**

The field of single-cell transcriptomics began earlier than single cell genomics, and accompanying analysis methods have been flourishing in the past 5 years [145], [146]. Single-cell transcriptome-oriented methods, such as clustering or gene-set enrichment analysis, are generally not optimal for targeted single-cell genomic data, where the cell-cell difference is much smaller, often only on a small set of genomic regions. The best method would be to rely on the ancestor-descendent relationship between every pair of cells' genomic sequence and build single-cell phylogenies. Several models for the evolution of SNVs in cancer have been developed to date [147], [148], [149], [150], [151], [152], [153], but only two [152], [153] have been constructed to also account for the frequent, complex aneuploidy in cancer, one of which is directly applicable to the dataset presented in this paper [153] but its accuracy is yet to be tested. Clustering single cells based on both SNV and CNV in PA04's multiregional samples, we were able to resolve

stepwise evolution patterns in metastatic PDAC, showing single-cell-level convergent evolution to inactivating the TGF-β pathway, although at a limited resolution [130]. Therefore, better methods need to be developed and tested to analyze this new dataset further. This was another significant motivation for our work to be described in **Section 3.**

Section 3: Development of a pipeline to resolve genomic evolution from snDNA-seq data

Single-cell/nucleus RNA sequencing, popular in general biological sciences and having wide-ranging applications nowadays, surveys the transcriptomics of single cells which is much more abundant in quantity than the genome, and asks questions about sets of genes in groups of single cells. But for targeted single-nucleus DNA sequencing the genome needs to undergo amplification before sequencing, which results in a different error profile. Given the data's high density (for each gene, there's significant read information in virtually all single cells), we also seek to investigate events at the single-cell level and utilize them to infer evolutionary dynamics. As mentioned in the previous section, due to the novelty of this single-cell DNA library preparation method, there has not been a lot of development on its corresponding bioinformatics and downstream analysis. Therefore, we saw the need to develop a set of bioinformatics/computational analysis tools ourselves to maximize the insights to be gained from this dataset.

1. The bioinformatics pipeline for mutation calling

1.1 background

Given the snDNA-seq's high sensitivity, I was not only expecting it to (1) output single-cell level information on small mutations (single-nucleotide variants, SNVs; insertions and deletions, indels) already detected by bulk sequencing on the matched tissue, but also (2)

discover low-prevalence mutations *de novo*. The technology's inventors Benjamin Demaree and Cyrille Delley developed a pipeline [154], which was later adapted by the startup company Mission Bio into its Tapestri pipeline. But upon testing, I found that both versions could only achieve (1) but not (2), as (2) requires a more stringent quality control process to remove false positives which are of high abundance in the data.

From a technical perspective, neither of the above two pipelines were written in a way that optimizes efficiency while being run on a high-performance computing (HPC) system, where thousands of cores could be leveraged in parallel and greatly speed up the whole process. To give an example, the Tapestri pipeline's variant calling step has to run as one entity over 36 cores and usually take overnight (6hr+); but the pipeline that I developed and will be described in detail below can scatter over thousands of cores of the MSK HPC system and finish in 1 hour or so.

1.2 an efficient pipeline for mutation calling and preliminary variant filtering

As a first pass, with aligned single-nucleus DNA libraries, I used Mutect2 (GATK 4.2.5.0) for mutation calling. Different from the original Tapestri pipeline, I applied technical filters at the single-cell level right after mutation calling. Filters included: base_qual/low_allele_frac/weak_evidence/slippage/multiallelic/clustered_events.
Required minimum depth was 4 reads and variant allele frequency (VAF) was 0.2. This showed effectiveness in removing many alignment-related artifacts.

Then, I concatenated all single-cell mutation calls into the sample's consensus mutation list and do a joint genotyping. But in the midst, I applied a prevalence filter to remove

mutations present in < 3 as these are very likely PCR-caused false positives, and are of a number high enough to be computationally expensive for the rest of the pipeline.

Importantly, I used **Snakemake** [155], a workflow management system that was designed to be seamlessly scaled to server, cluster, grid and cloud environments, to parallelize each step that only required looking at each individual cell (i.e. single-cell-level variant calling/genotyping). Additionally, it enables high-fidelity execution of the pipeline due to its output-dependent execution model that ensures that all desired output files are generated and up-to-date, which is crucial for such high-throughput snDNA-seq data with thousands of intermediate files per step.

1.3 a series of quality control steps to remove false positive variants for *de novo* mutation calling

With an average throughput of 2000+ single nuclei per sample, this snDNA-seq approach has a theoretical sensitivity of calling short mutations, including single nucleotide variants (SNVs) and insertions and deletions (indels), present at 0.15% single-cell prevalence, assuming events present in only 3 cell are considered. But many rare mutations could be from PCR /sequencing errors. To enable de novo mutation calling, we took several steps of quality control, inspired by bulk DNA sequencing bioinformatics.

Filtering by relative prevalence and mutational signature

In samples taken from different organ sites, we noticed a significantly uniform mutational signature enriched in mutations <0.5% single-cell prevalence detected by this snDNA-seq technology [to-be-published data]. As different types of human tissues have been shown to

display distinct mutational signatures and the uniform signature did not seem to match any of them [27], [156], we decided to use 0.5% single-cell prevalence as the threshold to call mutations. Rare T>C variants, enriched in this likely artifact signature, were also filtered out from the final output.

Detecting rare variants leveraging multiregional sampling

Sampling multiple regions of a tumor enables mutations that have a large clonal fraction at one site but lower than detection threshold fraction at another to be captured. Therefore, mutations from all samples of the same tumor were pooled and genotyped in all single nuclei with bcftools.

Assembly of a panel of normal

A panel of normal (PON), proven effective in eliminating library artifacts in bulk sequencing [cite], was assembled for snDNA-seq from three sources:

8 unrelated normal pancreas samples were subject to the same snDNA-seq sequencing and mutational calling process (up to filtering by prevalence). Variants called in >= 4 unmatched normal samples were discarded.

SnDNA-seq SNV lists of all PDACs studied here were concatenated, and variants called in >= 3 patients, except for those in known hotspot (KRAS codon 12), were discarded. Manual curation proved this effective in removing low-prevalence SNVs likely caused by library error, and confirmed that no driver mutations identified by matched bulk sequencing were discarded.

Variants present in a bulk WES PON (gs://gatk-best-practices/somatic-b37/Mutect2-exome-panel.vcf) were discarded.

To retain real SNPs/SNVs with high occurrence frequency in the population, variants with a frequency > 0.01 in the 1000 Genome Project were whitelisted.

Additional filters for phylogenetic analysis

As our phylogenetic analysis relies on high-density single-cell VAF signal of mutations, those with >0.3 biallelic dropout (ADO) rate in any sample were discarded.

A diagram describing the above pipeline is shown in Supplementary Figure X, and the code repository is available at https://github.com/haochenz96/TapVarCallSmk.

2. Single-cell copy number clones clustering

With Palash Sashittal

2.1 background

Multiple evolutionary models have been proposed to construct tumor phylogenies from scDNA-seq data. Early works used SNVs as evolutionary markers, and relied on the *infinite-sites model* which states that an SNV can be gained only once and never be subsequently lost in the phylogeny [157]. While the same SNV occurring independently more than once is rare [158], loss of SNVs due to copy-number deletions is common in cancer [147]. To account for these losses, other works [159], [160], [161], [161] use some

variant of the *k-Dollo model* [162], in which a mutation can be gained at most once but may be lost at most k times during the course of the evolution, where k is a user-defined integer.

A major limitation of the methods above is that they do not utilize any information about CNVs, which can often also be derived from scDNA-seq data. This limitation is addressed in methods such as SCARLET [152], BiTSC2 [163], and COMPASS [153] which incorporate copy-number information during phylogeny inference. SCARLET introduced a novel loss-supported Dollo model that requires the copy-number profile of each cell and the copy-number phylogeny as input. BiTSC2 and COMPASS, on the other hand, construct a joint phylogeny with both SNV and CNV events. However, these methods rely heavily on accurate and simultaneous identification of SNVs and CNVs on the same set of cells, which is challenging with the current scDNA-seq technologies [164].

Current scDNA-seq technologies fall into one of two classes which focus on *either* SNV or CNV. First, *whole genome* scDNA-seq technologies yield data with roughly uniform coverage of the *whole genome* but with low depth at any particular locus, making it suitable for detection of CNV in large genomic region but too sparse for detecting point mutations in single-cells [143], [144]. By contrast, the Mission Bio Tapestri platform used in our work [130] performs high-coverage targeted sequencing (>80× coverage) of hundreds of amplicons from thousands of cells, allowing accurate identification of SNVs. While this entailed the infeasibility of detecting large-scale CNVs, based on our initial testing it still allowed for clustering cells into clones with distinct copy number profiles [165], and crude absolute copy number calling at genes of interest in each clone.

Herein we have introduced a new evolution model for scDNA-seq - the constrained k-Dollo model - which takes into both SNV and CNV signals to compute phylogenetic trees. The majority of the project's preliminary technical details and validation have been written and published with my collaborator Palash as the first author; I have mostly contributed to adapting this model for our particular dataset and building the updated tool named fast Constrained Dollo Reconstruction (fast-CONDOR) to derive new biological insights, which I will describe briefly below. A full publication is being prepared as this thesis is being written.

As the first step, according to the model, cells must first be partitioned into copy number (CN) clones. Several methods were attempted, and a negative binomial mixture model stood out because (1) the CN clones called roughly conformed to those from other callers [166] (2) more importantly, LOH events were easily identified from mutation VAF signals between clones, which orthogonally validated the partitioning of cells, while also provided ideal inputs for fast-CONDOR which models mutation losses.

2.2 Clustering for copy number clones from snDNA-seq data

Suppose we measure the number of reads aligned to m genes, each covered by d amplicons, from n cells. We represent this measurement by a read count tensor R, where $r_{i,j,l}$ is the number of reads aligned to amplicon l of gene j in cell i. Suppose the cancer tumor is comprised of p copy number clones. We assume that all loci of a gene have the same copy number, except in regions of focal homozygous deletions, where the copy number is 0. The copy number clones thus be represented by two matrices, (1) copy number matrix C, where

$c_{k,j}$ is the copy number of gene j in clone k and (2) homozygous deletion matrix H where $h_{k,j,l}$ is 0 if amplicon l of gene j lies in a region that has been homozygously deleted in clone k . We encode the membership of cell i to one of the p copy number clones by random variable Z where $z_{i,k} = 1$ if and only if cell i belongs to copy number clones k. Note that since each cell belongs to exactly one copy number clone, we have $\sum_{k=1}^{p} z_{i,k} = 1$.

We model the number of reads aligned to an amplicon by a negative binomial model, where the mean of the distribution is proportional to the copy number of the genomic region spanned by the amplicon. Specifically, we model the read count $r_{i,j,l}$ for cell i belonging to copy number clone k as follows,

$$r_{i,j,l} \mid z_{i,k} = 1 \sim \boldsymbol{NB}\left(\mu = h_{k,j,l} c_{k,j} f_{j,l} T_i, \phi = 1/\beta_{j,l}\right)$$

where T_i is the total number of reads in cell i , $f_{j,l}$ is the amplicon-specific factor amplification factor of amplicon l in gene j and $\beta_{j,l}$ is the heterogeneity (dispersion) parameter of amplicon l in gene j. Importantly, we would empirically measure $f_{i,l}, \beta_{j,l}$ by orthogonally running the same snDNA-seq experiment on diploid nuclei from normal samples and fit the model to the observed read count matrix with C set to a one-dimensional vector of 2's (we assume every cell belongs to the diploid cluter) using maximum likelihood.

Then in tumor samples where there are mixtures of tumor and normal cells, we assume that the cells are sequenced independently. Let $\vec{\alpha} = (\alpha_1, \ldots, \alpha_p)$ be the mixing proportions of

the p copy number clones such that $\sum_{k=1}^{p} \alpha_k = 1$. Importantly, p is a hyperparameter we need to define based on observation of model performance, which will be detailed below.

The likelihood of observing read count matrix R for a given copy number matrix C , homozygous deletion matrix H and clone proportions $\vec{\alpha}$ is as follows,

$$\begin{aligned}
\mathbf{Pr}(R \mid C, H, \vec{\alpha}) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \prod_{\mathrm{l}=1}^{d} \mathbf{Pr}(r_{i,j,\mathrm{l}}) \\
&= \prod_{i=1}^{n} \sum_{k=1}^{p} \prod_{j=1}^{m} \prod_{\mathrm{l}=1}^{d} P(r_{i,j,\mathrm{l}} \mid z_{i,k} = 1)\, \mathbf{Pr}(z_{i,k} = 1) \\
&= \prod_{i=1}^{n} \sum_{k=1}^{p} \alpha_k \prod_{j=1}^{m} \prod_{\mathrm{l}=1}^{d} \mathbf{Pr}(r_{i,j,\mathrm{l}} \mid z_{i,k} = 1)
\end{aligned}$$

We define the copy number calling problem as follows.

Copy number calling

For a given read count tensor R for n cells and m genes with d amplicons, find the copy number matrix C , homozygous deletion matrix H and copy number clone proportions $\vec{\alpha} = (\alpha_1, \ldots, \alpha_p)$ where p is a hyperparameter, that maximize the likelihood $\Pr(R \mid C, H, \vec{\alpha})$.

As we ran the model, we tried different p hyperparameters, usually from 2 to 10 because we observed that as p increased, the number of unique clones generally plateaued around 10. In the end, we would pick the one that resulted in the highest inter-cell heterogeneity, so as to provide the highest granularity to phylogeny inference later.

While this initial clustering proved to be useful for directly calling homozygous deletions and high-level amplifications, it usually overfit by calling too many clones due to the overdispersion of PCR-amplified read counts. Therefore, a refinement step *after phylogeny inference* would merge clones that have no SNV gain/SNV LOH/SNP LOH event between them, ensuring that the final copy number clones were supported by both CNV and SNV signals.

3. Phylogenetic inference from snDNA-seq data

With Palash Sashittal

3.1 background

Again, the backbone of the phylogenetic inference algorithm has been well described in our previously published paper [165]. Herein I will highlight major updates we made to the model since then.

3.2 a simple statistical test for clonal relationship between two mutations

Before we jump into phylogenetic inference with the many complex SNV and CNV signals, a simple question we would run into with snDNA-seq is: given two SNVs' distribution of reads across single cells, how can we measure their clonal relationship – whether they are colocalizing in the same set of cells or mutually exclusive? Or expanding this to more general application, if we can calculate the probability of presence of any two genomic

events (heterozygous SNV, homozygous SNV, homozygous deletion etc.) in a set of cells, how do we measure their clonal relationship?

From a statistical perspective, we need to calculate a p-value that tells us how likely two mutations are colocalizing/mutually exclusive by random chance.

Suppose we have a probability matrix $P \in [0,1]^{n \times m}$ for n cells and m mutations, where $p_{i,j}$ is the probability that mutation j is present in cell i. For any two mutations j and j', let Y be the random variable that denote the number of observations of mutual exclusivity across the n cells, i.e. the number of observations of either $(0,1)$ or $(1,0)$ gametes in the mutation matrix. Assuming that the mutation states of cells are independent, Y is a Poisson binomial variable with parameters:

$$* \; p = \{p_{i,j}\left(1 - p_{i,j'}\right) + \left(1 - p_{i,j}\right)p_{i,j'} : \forall i \in [n]\}.$$

We define a null distribution that conserves the marginal expectation of the number of occurrences of each mutation across the n cells.

Specifically, we define probability matrix P' such that:

$$p'_{i,j} = \frac{\sum_{i=1}^{n} p_{i,j}}{n}.$$

Let X be the random variable that denote the number of observations of mutual exclusivity for mutations j and j' across the n cells under the null distribution.

X is a binomial variable with n trials and probability of success given by:

$$p = \frac{\sum_{i=1}^{n} p_{i,j}}{n}\left(1 - \frac{\sum_{i=1}^{n} p_{i,j'}}{n}\right) + \left(1 - \frac{\sum_{i=1}^{n} p_{i,j}}{n}\right)\frac{\sum_{i=1}^{n} p_{i,j'}}{n}.$$

The p-value for rejecting the null hypothesis is given by $Pr(X \geq Y)$.

Read count model

As mentioned above, we hope to statistically test for multiple types of genomic events. The simplest case is presence of a mutation, which has been well modeled in many previous works [149], [152] and also described in the CONDOR paper. We just need to vary it for other genomic events:

- A mutation being in a homozygous state is modeled as absence of the reference allele.
- A mutation being lost is modeled as absence of the alternate allele.
- An amplicon being homozygous deleted can be simply modeled as an indicator function with hard thresholding (e.g. total read count = 0).

The mutual exclusivity test

We formally define hypothesis testing problem as follows:

Given a probability matrix P and two mutation j and j', find the p-value $\Pr(X \geq Y)$, where Y and X is the number of occurrences of gametes $(0,1)$ and $(1,0)$ for mutation j and j' under the alternate and the null hypothesis, respectively.

If we want to test for occurrence of different gametes, we simply modify the indicators in the model: i.e. if we testing for colocalization, we would set the gametes to be (0,0) and (1,1).

3.3 the updated fast-CONDOR algorithm

Building the model

The constrained k −Dollo model integrates SNVs and SNPs with partial information about CNVs on the same set of cells during phylogeny inference. Specifically, the model incorporates CNVs via a clustering of cells, where all cells in the same cluster have the same copy-number profile (the copy number of all loci across the genome).

It involves the following four constraints:

1. a mutation is allowed to be gained at most once in the phylogeny. This constraint stems from the *infinite-sites assumption* which posits that it is very unlikely for the same position in the genome to get mutated multiple times independently.
2. a mutation can be lost at most k times in the phylogeny, where k is user-defined parameter.
3. since reversal of mutations *back mutations* in cancer are rare, we assume that SNVs and SNPs can only be lost due to overlapping CNVs. As such, mutation losses are only allowed between cells that belong to distinct copy-number clusters.

4. we assume that each copy-number profile describing the copy number states over the entire genome arises only once in the phylogeny. Consequently, all cells belonging to the same copy number cluster form a connected subtree in the phylogeny.

The original CONDOR algorithm suffers the scalability issue on thousands of cells, common in this PDAC snDNA-seq dataset. Herein we developed a scalable version "fast-CONDOR" to build coarse-grained tumor phylogenies *at the copy number cluster level*.

Specifically, each mutation can have one of the following four states for each copy number cluster -- (1) none of the cells contain the mutation (state 0), (2) some of the cells contain the mutation (state $(+)$), (3) all the cells contain the mutation (state 1), and (4) all the cells have lost the mutation (state 2).

Under the constrained k-Dollo model, since a mutation can occur at most once in the phylogeny, each mutation can be gained in at most one of the copy number clusters. As such, for each mutation, there can be at most one copy number cluster with state $(+)$.

Further model details, including construction of the objective function, optimization, are similar to as described in the CONDOR paper, therefore omitted here. They will be elaborated in our upcoming publication.

Creating the inputs

Furth Inputs to the model include:

1. CN clones from the previous section.
2. mutation signals, in the form of total read depth and alternative read count matrices.
3. annotation of the mutations, differentiating between germline heterozygous SNPs (which could undergo LOH) and somatic SNVs (which could undergo both gain and LOH).

To get 3, mutation calls from snDNA-seq were compared with bulk data of matched normal and tumor to get germline/somatic status. Homozygous SNPs, identified as those with pseudo-bulk VAF > 0.9 in snDNA-seq, were filtered out. Upon testing CONDOR, we found a set of loci recurrently called as LOH in nonneoplastic clones (i.e. cells without driver mutations such as KRAS, TP53) in multiple patients, likely arising due to artifactual allelic imbalance at corresponding amplicons caused by PCR. These loci were manually selected and filtered out.

Refining the outputs

After CONDOR optimizes a tree with above events attached to the branches, the CN clones as main leaves, and sub-CNV-clone SNV events and cells into subtrees, the outputs are refined with following steps:

1. CN clones fitting any of the following criteria are merged into the diploid clone at the root:
    a. smaller than 0.1% of total cells.
    b. have no additional SNV events than the diploid clone. This entailed CN clones with only LOH events being merged. This was implemented because based on

observation of our samples, such clones always had only likely false positive LOHs called, thus could not be confirmed.

2. internal nodes' CN profiles were inferred by Sankoff's algorithm [167].
3. edge lengths were calculated combining SNV/LOH events and CNV events:

$$\begin{aligned} &\mathbf{N}(\text{somatic SNV gains/LOHs}) \times p1 \\ +\ &\mathbf{N}(\text{germline SNP LOHs}) \times p2 \\ +\ &\text{CNV distance} \times p3 \end{aligned}$$

where:

- CNV distance was defined as the L1 norm of two clones' CNV profile vectors' difference. Only amplicons targeting genes covered by >= 3 amplicons were used.
- p1, p2, p3 were normalization factors determined empirically: p1 = 100/3, p2 = 100/3, p3 = 0.5.

Section 4: Genomic evolution of pancreatic cancer at single-cell resolution

1. Introduction

Armed with the wet-lab and dry-lab pipelines established in the previous sections, we can now proceed to the main goal of this work: to study the genomic evolution of pancreatic cancer at single-cell resolution.

2. Mutational and evolutionary landscape of pancreatic cancer at single-cell resolution

2.1 Cohort overview

We studied the single cell genomic features of 23 patients with a primary pancreas neoplasm. The final pathologic diagnosis of 22 patients was PDAC and one was diagnosed with acinar cell carcinoma (ACC) [**Figure 6A]**. We opted to study the latter patient given they contained a germline mutation in BRCA2; four of the 22 PDAC patients had gBRCA2 mutations as well (discussed further in later sections). For 20 patients, at least two spatially distinct samples were studied per patients to capture spatial heterogeneity. To capture single cell features at various clinical stages the cohort included samples from patients with both early-stage disease collected from surgical resections (n=8) and late-stage unresectable disease from research autopsy (n=11). Matched longitudinal biopsies taken

by 22-guage core needle biopsies before and after treatment were collected from 5 patients [**Supplementary Table 2]**.

To discern the genetic features of our cohort, we performed targeted snDNA-seq using a custom designed 596-amplicon panel targeting essential genes of PDAC; genes affected by both single nucleotide variant (SNV), small insertion/deletion (indel) and/or copy number variations (CNV) were considered [**Supplementary Table 3; Methods**]. In total 127,859 single-nucleus DNA libraries were profiled by this panel with a mean of 1827 single nuclei studied per sample (range 50-7540), and 5559 per patient (range 329 - 20123) **[Figure 7A].** Bulk whole-genome/exome sequencing (WGS/WES) data were collected for each tumor and matched normal sample simultaneously for quality control of germline/somatic event calling.

While resected/autopsied samples were large enough to be embedded in optimal cutting temperature (OCT) compound and inspected to ensure cellularity and tumor purity (cite your own paper), longitudinal biopsies were usually too small (<8mm^3) to be embedded and had to be extracted and sequenced without controlling for the two variables, thus resulting in many such samples to have lower than optimal throughput (thousands of nuclei). Nevertheless, even TP12 endoscopic ultrasound (EUS) biopsy with the lowest yield of 49 nuclei resulted in useful genotype information (**Figure 7B**), thus proving the feasibility of applying this snDNA-seq approach to routine pancreas biopsied samples.

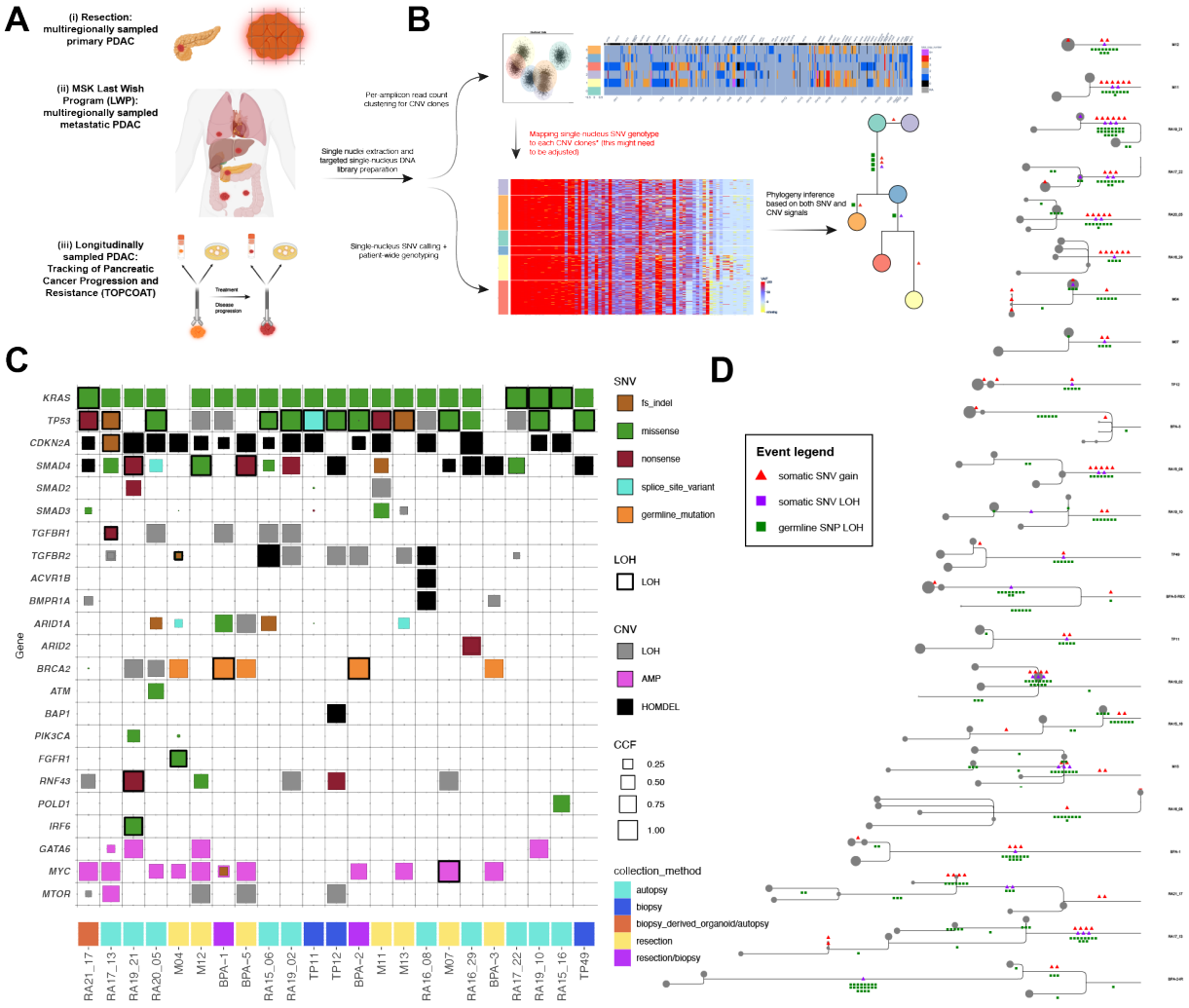

Figure 6. Study overview; genetic and phylogenetic features of the cohort.

a. sample collection procedure. Frozen primary tissues were first processed into single-nuclei suspension and targeted sequenced following our previously published protocol. **b.** overview of computational pipeline. The raw FASTQ data were demultiplexed and aligned as previously described. Single-nucleus BAM files then went through two orthogonal modules. (i) CNV clones were called by clustering the per-amplicon read count matrix (ii) SNVs in each nucleus were called. The two modalities were fed into CONDOR, the phylogeny inference module, to infer a consensus phylogeny. **c.** SnDNA-seq oncoprint. On the x-axis: cases are sorted by their respective number of mutational events in descending order from left to right. On the y-axis: genes are sorted by the number of alterations in the patient cohort in descending order from top to bottom. Within the grid, the square size is proportional to the cancer cell fraction (CCF) of each alteration; the square color corresponds to the type of alteration, as indicated by the legend on the right. **d.** All phylogenetic trees in condensed format. Each row represents on case's phylogeny with the case name labeled on top. The y-axis represents inferred copy number variation (CNV)

distances among clones (grey circles). Therefore the longer the branches of a phylogeny, the higher CNV complexity/heterogeneity its clones have. The size of the circle is proportional to the number of cells. For events on the edge: red triangles ▲ indicate gain of SNVs; purple triangles ▲ indicate LOH of SNVs; green squares ■ indicate LOH of germline SNPs.

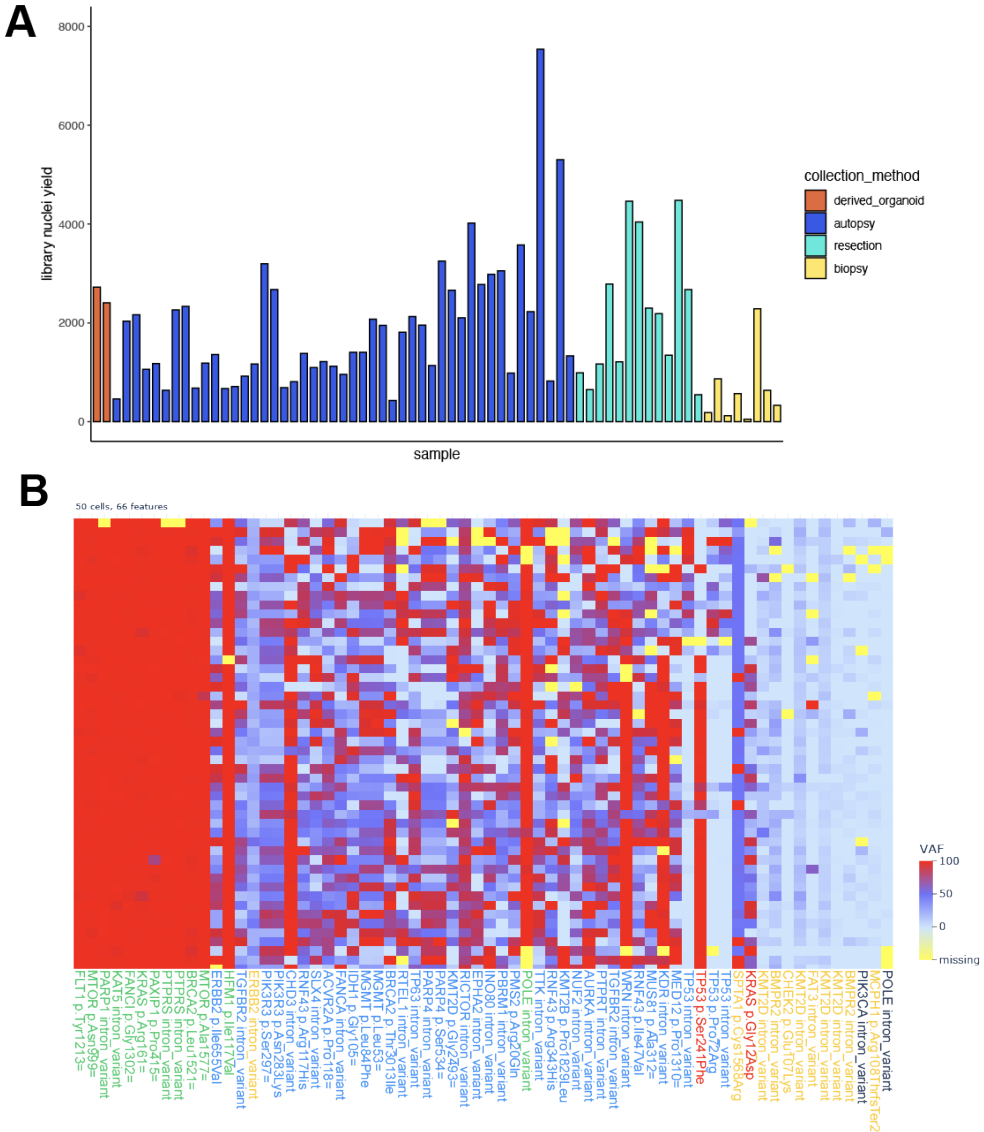

Figure 7. Technical details of the snDNA-seq pipeline.

a. SnDNA-seq library nuclei yield for each sample, colored by sample collection method. **b.** example single-cell mutation heatmap. Mutations (columns) are called and their variant allele frequency (VAF) in each single cell is plotted as the heat. Semi-continuous

({missing, [0, 100]}) VAF values are represented by the legend on the right. Based on their signals in orthogonal bulk sequencing of the matched normal and tumor, mutations are colored green (germline homozygous), blue (germline heterozygous), red (somatic), black (novel to snDNA-seq). The yellow SNVs are labeled as noisy loci that suffered from library preparation artifacts and are filtered out. The germline heterozygous SNPs and somatic SNVs are input for phylogeny inference.

2.2 Genetic features of cohort

We designed a novel bioinformatics pipeline for this unique snDNA-seq dataset [**Figure 6B)** and detected somatic SNVs, indels, monoallelic and/or biallelic losses in at least one known driver gene in all cases analyzed [**Figure 6C**]. Activating *KRAS* mutations were identified in all but one PDAC (95%). Alterations in *TP53* (78%), *CDKN2A* (74%), and *SMAD4* (65%) were also observed at high frequency, as expected [141]. Somatic alterations noted at lower frequency included *SMAD2, SMAD3, TGFBR1, TGFBR2, ARID1A, ARID2, BRCA2* and *RNF43.* Copy number alterations identified included amplifications in *MYC* (48%)*, GATA6* (17%) and *MTOR* (6%), as well as biallelic loss (homozygous deletion, homdel) in genes such as *CDKN2A, SMAD4* and *TGFBR2*. A technical strength of snDNA-seq compared to bulk sequencing is that it enables direct calculation of each alteration's corresponding cancer cell fraction (CCF); as a result we could distinguish between clonal drivers such as *TP53* (mean CCF = 0.85), *RNF43* (0.83), *CDKN2A* (0.79), *MYC* (0.78), and subclonal drivers such as *SMAD4* (0.71), *TGFBR2* (0.55), *ARID1A* (0.42) [**Figure 6C, square size**].

The snDNA-seq approach expands point mutation (SNV/indel) calling sensitivity from bulk sequencing's 2% to 0.5% single-cell prevalence. At this resolution, we observed a subset of genes with an increased frequency of point mutations than reported in bulk

sequencing studies. For example, *ARID1A* mutations were present in 26% of our cohort compared to 6% in The Cancer Genome Atlas (TCGA) (p=0.0015) and 14% in the International Cancer Genome Consortium (ICGC) (p=0.16) [55], [56] [**Figure 6B**].

Further, subclonal and subgene (affecting only a part of the gene) monoallelic/biallelic losses are difficult to detect with common bulk sequencing methods, particularly for PDAC where tumor purity is often limited. We leveraged our targeted snDNA-seq method by evaluating each mutant loci's variant allele frequency (VAF) and single amplicons' (200-300 base pairs) read count, after accounting for amplicon dropout rates, to detect such events in small groups of cells. For instance, we detected a *SMAD4* homozygous deletion in case TP12 by complete missing signal (zero read) of 1 of 11 amplicons targeting the gene, corresponding to its last exon, in a subset of cells (**Figure 8A)**. Thereby we detected a higher alteration frequency of *CDKN2A* and *SMAD4* (74% and 70%, respectively) than reported by TCGA (30%, 32%) and ICGC (30%, 27%). Specifically, we detected biallelic loss (homozygous deletion) of *CDKN2A* in 70% cases *and SMAD4* in 30% cases, which are comparable to the numbers (around 60% for *CDKN2A* and 40% for *SMAD4*) found in a mixed cohort of primaries/metastases using microdissection to enrich for tumor content [57]. We also detected high frequency of monoallelic losses (loss of heterozygosity, LOH) affecting *TGFBR1* (21%) and *TGFBR2* (39%). We interpret these findings to reflect the increase in sensitivity by our assay rather than more extensive sampling, as these rates are also higher than reported in in a recent study of 91 multiregion sampled PDAC research autopsies [to-be-published work from my lab].

2.3 Major patterns of evolutionary trajectory

In addition to novel methods of calling somatic variants, we developed a new computational approach to reconstruct the single-cell clonal phylogeny from each patient. Briefly, this method first clusters cells into copy number variation (CNV) clones, and uses germline single-nucleotide polymorphism (SNP) and somatic single-nucleotide (SNV) allelic information to build a clonal phylogeny at single-cell resolution [**Figure 6C**] [165].

Figure 6D illustrates the phylogenies generated from this method. A mean of 2.57 truncal driver SNV events were covered by our panel. Among all somatic driver gene SNVs identified, on average 74% were truncal in origin (i.e. occurred before the most recent common ancestor (MRCA) of all clones identified). Similarly, 78% of CNV alterations affecting driver genes occurred prior to the MRCA. Collectively, this pattern of early fixation of drivers followed by generation of intratumoral heterogeneity by copy number variation is consistent with multiregional bulk sequencing of larger PDAC cohorts [36], [122].

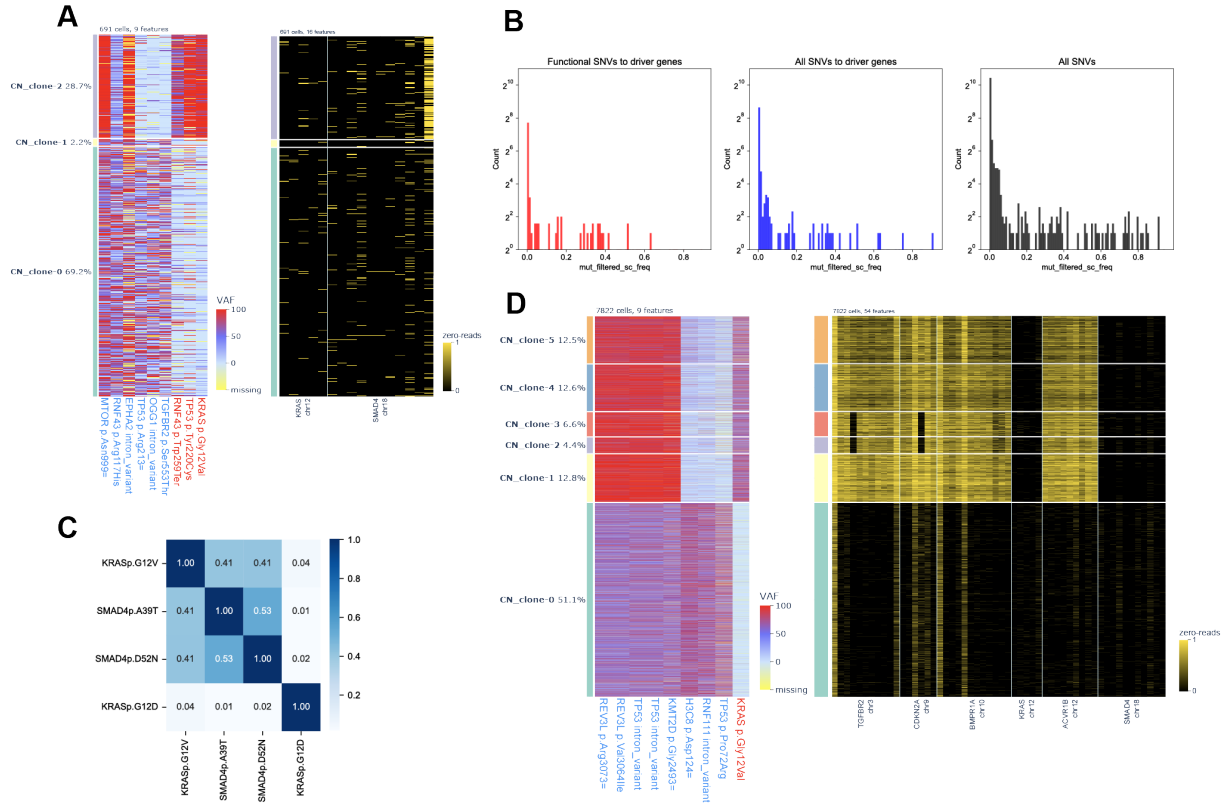

Figure 8. Novel features of the snDNA-seq dataset.

a. SnDNA-seq identified focal homozygous deletion (homdel) of *SMAD4.* Left, single-cell mutation VAF heatmap, with cells grouped into phylogenetic clones. The clustered heat of the *RNF43, TP53, KRAS* mutations indicate the tumor cells carrying these mutations. Right, single-cell amplicon-level binarized read count (zero or non-zero number of read) heatmap. The clustered heat at the last *SMAD4* amplicon indicates a focal Homdel. **b.** Distribution of mutations down to 0.5% single-cell prevalence as detected by snDNA-seq. **c.** mutual exclusivity test of the secondary *KRAS* p.G12D mutation against the main drivers of the RA17_22 PDAC. **d.** SnDNA-seq identified homdel of three receptor genes of the TGF-β pathway in RA16_08. The heatmaps are plotted as described for panel **a.**

3. Traces of convergent evolution

3.1 paucity of low-prevalence driver gene alterations

Based on bulk sequencing one or more pathways in PDAC may be affected by convergent evolution. For example, *KRAS* wildtype (WT) PDACs have been shown to have alternative genetic mechanisms of activating MAPK-ERK signaling[55], while the TGF-β pathway may accumulate genetic alterations in intracellular pathway components (i.e. *SMAD4, SMAD2*) or surface receptors (i.e. *TGFBR2*) [95]. With snDNA-seq's increased resolution, we asked if within the same tumor, additional clones exist to these or other PDAC-relevant molecular pathways that are otherwise unrecognized. Single cell populations of >= 1% CCF were included in the oncoplot in **Figure 6C**, and showed mutations missed by bulk-sequencing but captured by snDNA-seq: *PIK3CA* mutation (CCF=0.0375), a *SMAD3* mutation (CCF=0.0134) in M04 and a *SMAD4* mutation in TP11 (CCF=0.0113). Nonetheless, even at this resolution we found that most PDAC driver pathways contain a single genetic event indicating prior strong selection in association with a near total to complete clonal sweep. The only exception was M04, where three mutually exclusive clonal lineages carrying (1) *TGFBR2* p.K300Tfs*6 (2) *TGFBR2* p.A426G/p.M425I (3) *SMAD3* p.S425F mutations **[Figure 9D]** showed convergence towards TGF-β inactivation; in the same tumor, two mutually exclusively lineages carried (1) *FGFR1* p.T17K (2) *PIK3CA* p.E542Q respectively, demonstrating convergence towards MAPK–ERK activation.

As we extended the threshold to 0.5% single-cell prevalence (not necessarily limited to the cancer cell fraction) per sample, low-prevalence mutations to other PDAC drivers were detected but did not show strong positive selection (dN/dS = 1.06) **[Figure 8B]**. In fact, possible negative selection was observed when considering for all genes on the panel (dN/dS = 0.91), suggesting a purifying effect of cancer growth. One secondary *KRAS* alteration was identified in the primary site of patient RA17_22 (single-cell prevalence = 0.0083) but showed little correlation with the main tumor clone **[Figure 8C]**; we did not detect any additional secondary *KRAS* oncogenic alterations.

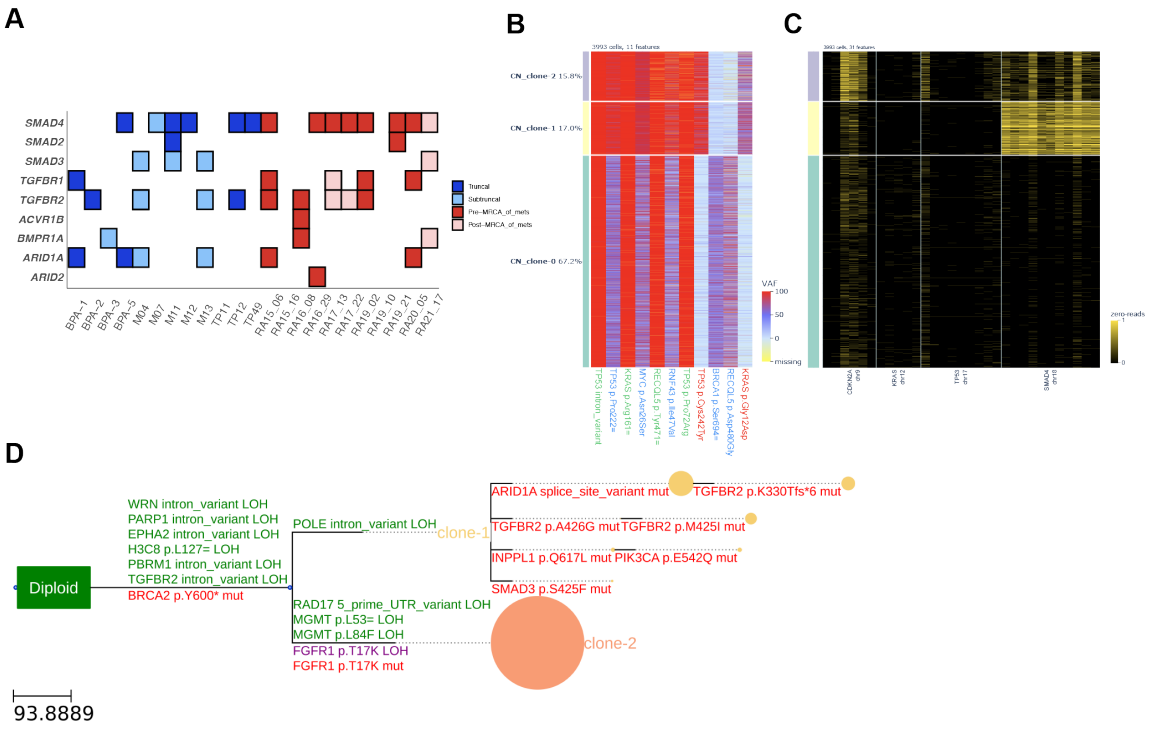

Figure 9. Convergence towards TGF-β inactivation.

a. TGF-β alteration status in phylogenies. X-axis: cases are sorted by stage at diagnosis from left (early-stage resectable PDAC) to right (metastatic PDAC). Y-axis:TGF-β pathway components covered by the snDNA-seq panel are ordered by canonical mediators (*SMAD4, SMAD2, SMAD3),* membrane receptors (*TGFBR1, TGFBR2,*

ACVR1B, BMPR1A) and downstream effectors (*ARID1A, ARID2*). **b.** for M07, single-cell mutation VAF heatmap. **c.** for M07, single-cell amplicon-level binarized read count (zero/non-zero) heatmap. **d.** single-cell clonal phylogeny of M04, with mutation events (green: germline SNP LOH; red: somatic SNV gain; purple: somatic SNV LOH) labeled on the edges. Clone sizes are indicated by the size of the circles.

3.2 abundance of inter-patient convergence towards TGF-β inactivation

Examples of inter-patient convergence for pathway activation were observed, for example the *FGFR1*/*PIK3CA* mutation in the *KRAS* WT PDAC M04. 6 of 23 cases had an *ARID1A* mutation, and in one PDAC that was WT for *ARID1A* (RA16_29) we found two mutations in *ARID2*, also a member of the SWI/SNF pathway **[Figure 6C]**.

Unlike these isolated examples, we again found pervasive inter-patient convergence towards the phenotype of tumor cell-intrinsic TGF-β unresponsiveness [**Figure 9A**]. The most common form of genetic inactivation of the pathway remains *SMAD4*, yet 6 other genes also have inactivating mutations that do or do not co-occur with *SMAD4*. The majority of these alterations, irrespective of the gene they occur in, are subtruncal in origin but occurred prior to the MRCA of the matched metastases. This suggests two features: (1) inactivating mutations of the TGF-β pathway have a strong selective advantage in the desmoplastic, nutrient-poor pancreas microenvironment; (2) these inactivating mutations have a selective advantage for survival in secondary sites. This is perhaps exemplified in PDAC M07 where *SMAD4* is deleted in one of the two clones localized to a single region of the primary tumor sampled [**Figure 9B, C**]. Some cases inactivate multiple components of the TGF-β pathway, but not *SMAD4*, the most prominent example being case RA16_08, which had three TGF-β family receptors focally deleted **[Figure 8D]**.

4. Continous evolution driven by CNVs

4.1 step-wise inactivation of a single gene

In addition to convergent evolution, we noted several examples where somatic alterations continued to accumulate within a single gene over the course of PDAC evolution. For instance, in case RA15_06, *TGFBR2*'s second amplicon was deleted in virtually all cancer cells **[Figure 10A, B**] whereas *CDKN2A* amplicons #2, 3, 5 were likely hemizygously deleted in clones 3 and 4 **[Figure 10C, arrow-pointed]**; subsequently the second allele was lost which were only localized to clones 1 and 2 corresponding to the left liver metastasis [**Figure 10D**]. We were not confident to call subclonal homdel in the other 3 *CDKN2A* amplicons because they had deletion signal in the normal tissue of not only this case but multiple other cases, which could likely attribute to primer failure due to high GC content in the region. Case RA17_22 (described as PA04 in Section 2) had two independent mutations to *SMAD4* on the same chromosome and then loss of the WT allele [**Figure 11A**]. Homdel to 10 of 11 amplicons of *SMAD4* took place in the MRCA of all tumor clones [**Figure 11B**]. The only intact amplicon **[Figure 11B, arrow-pointed]** was the one where the 2 point mutations located to. One subclone exclusively present in the liver metastasis and one slice of the primary **[Figure 11C, D]** had a *TGFBR2* LOH event, which potentially further contributed to the phenotype of TGF-β inactivation [168].

We also noted such a pattern at a *BRCA2* locus with germline mutation in BPA-2. The inferred tumor phylogeny started with *KRAS* p.G12D and *TP53* p.L344Q heterozygously mutated along with LOH at gBRCA2 locus **[Figure 10G]**. Then it diverged into two

branches- the more dominant branch carrying *TP53* LOH, *CDKN2A* homdel and many other LOH events; the other branch seemingly had *BRCA2* focal homdel at the germline mutated locus **[Figure 10H, I]**. Orthogonal statistical testing on raw single-cell read count data validated that the *BRCA2* homdel significantly colocalized with a heterozygous *TP53* state (p=0.0078).

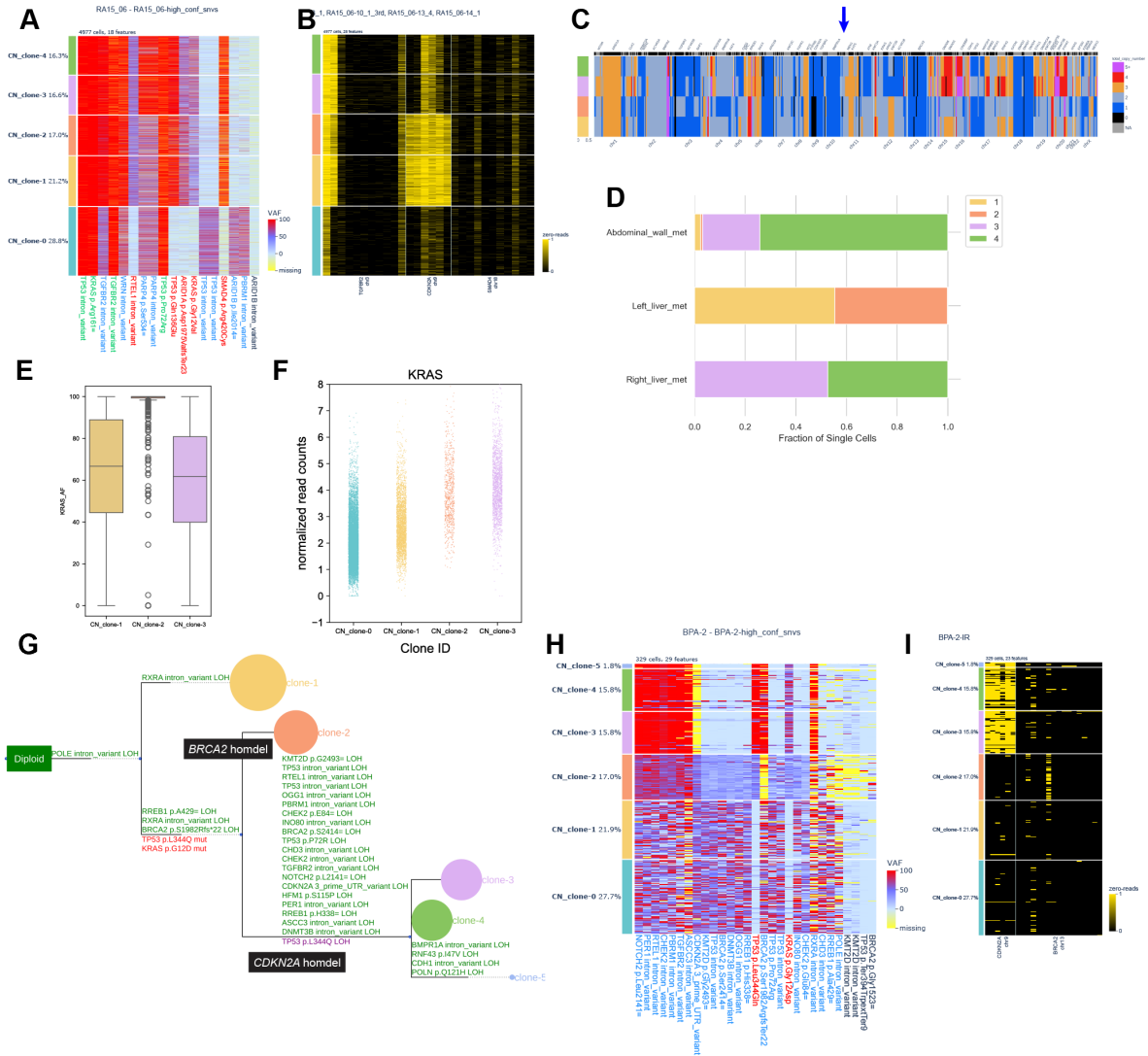

Figure 10. Step-wise inactivation of a single gene.

a. RA15_06 single-cell VAF heatmap for driver SNVs (left) and **b.** single-cell amplicon-level binarized (zero or nonzero) raw read count heatmap. **c.** Inferred total copy number

profiles of CNV clones found in RA15_06. For each clone, indicated by color blocks on the left, its per-gene total copy number (CN) states are shown, as indicated by the legend on the right. Genes (top of x-axis) are ordered by genomic location, while the black/grey blocks under each gene represent the number of amplicons covering it. Some amplicons were excluded from analysis due to significant noise, thus having "NA". While all amplicons targeting the same gene were constrained to have the same non-zero CN state, sub-gene homozygous deletion (0 total CN) was allowed. **d.** clone compositions at each primary/metastatic site sampled of RA15_06. **e.** single-cell VAF at the driver *KRAS* locus in RA17_22, across tumor clones. **f.** normalized read count of the *KRAS* gene in RA17_22, across normal/tumor clones. **g.** BPA-2 single-cell clonal phylogeny. h. BPA-2 single-cell mutation VAF heatmap. Note the yellow color of the heat indicates "missing" signal at the particular locus X cell, which corresponds to zero read at the amplicon and likely homdel in that genomic region. **i.** BPA-2 single-cell amplicon-level binarized read count heatmap. As cells are aligned between the two plots, the homdel signal appears to be shown by both the mutation heatmap and the read count heatmap: for *CDKN2A*, the cluster of cells with "missing" at the intron locus in the mutation heatmap correspond to those carrying zero read count across all 6 corresponding amplicons.

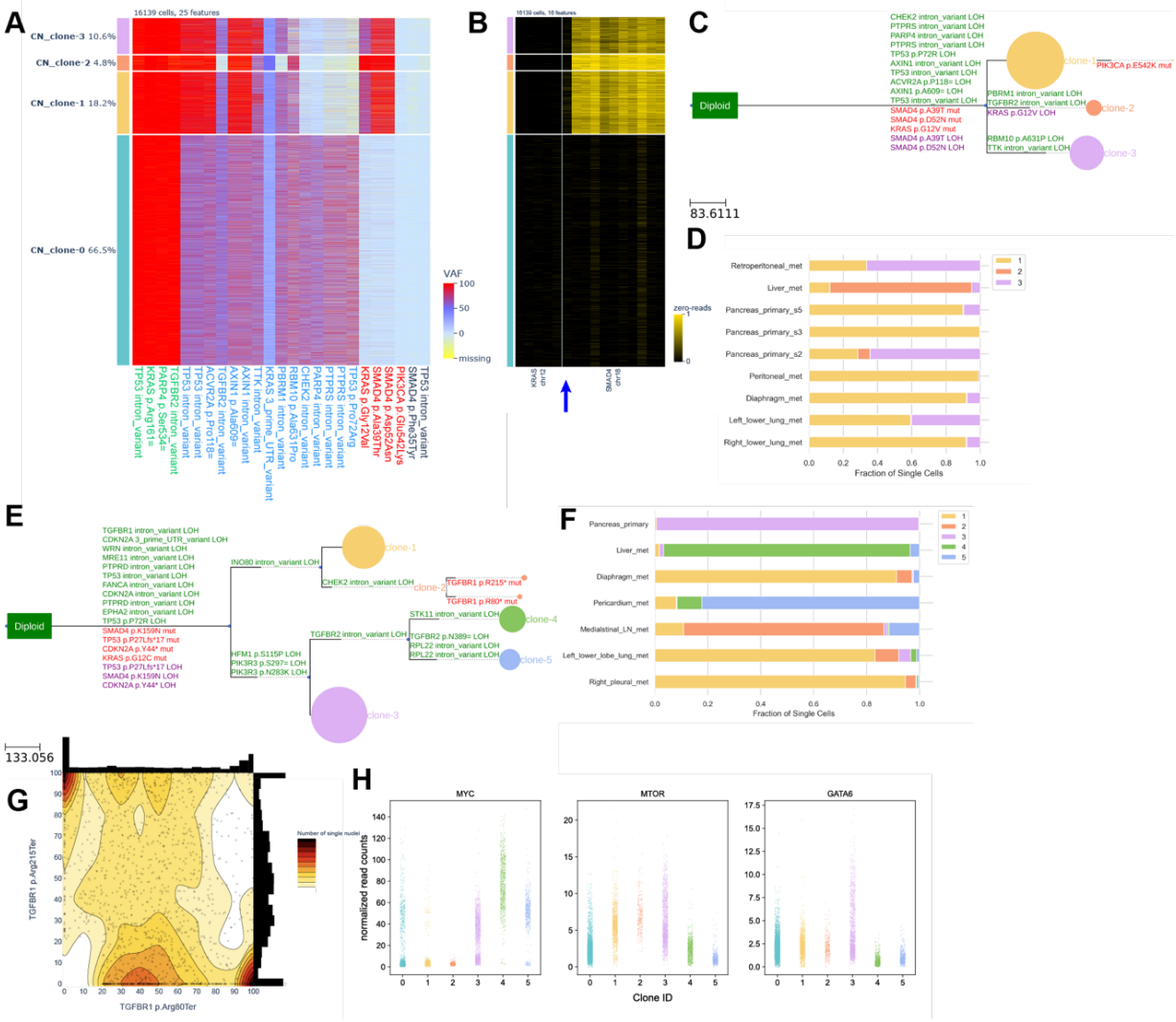

Figure 11. More continous evolution patterns.

a. RA17_22 single-cell VAF heatmap for driver SNVs **b.** RA17_22 single-cell amplicon-level binarized read count heatmap. The *intact SMAD4* amplicon carrying the two somatic mutations from **a.** is pointed by the blue arrow. **c.** RA17_22 single-cell clonal phylogeny**. d.** RA17_22 clone composition at each primary/metastatic site sampled. **e.** RA17_13 single-cell clonal phylogeny. **f.** RA17_13 clone composition at each primary/metastatic site sampled. **g.** 2D density contour plot of the two *TGFBR1* mutations' VAF distribution across single cells. **h.** stripplot of the distributions of normalized read counts of *MYC, MTOR, GATA6* genes in single cells of RA17_13, split by clones identified from the phylogeny.

4.2 mirrored loss of heterozygosity of driver mutations

Such continuous evolution also manifests as LOH of somatic driver mutation loci in RA17_13. Two *TGFBR1* nonsense mutations were subclonally gained in one of the two main lineages of metastatic clones [**Figure 11E]**. The fact that clones of that branch migrated to four spatially distant metastatic sites [**Figure 11F**] indicated that they likely came from a subclone selected for by the primary site rather than any metastatic site. Focal analysis of the subclone identified four sub-populations: (i) cells heterozygous for *TGFBR1* p. R80* (shortened as "A") (ii) cells heterozygous for both mutation A and *TGFBR1* p.R215* ("B") (iii) cells homozygous for A (iv) cells homozygous for B [**Figure 11G**]. Therefore a likely evolution path would be: some PDAC cell first gained A and clonally expanded into (i), gained B and expanded into (ii), and fractions of cells lost either copy to diverge into (iii) and (iv)

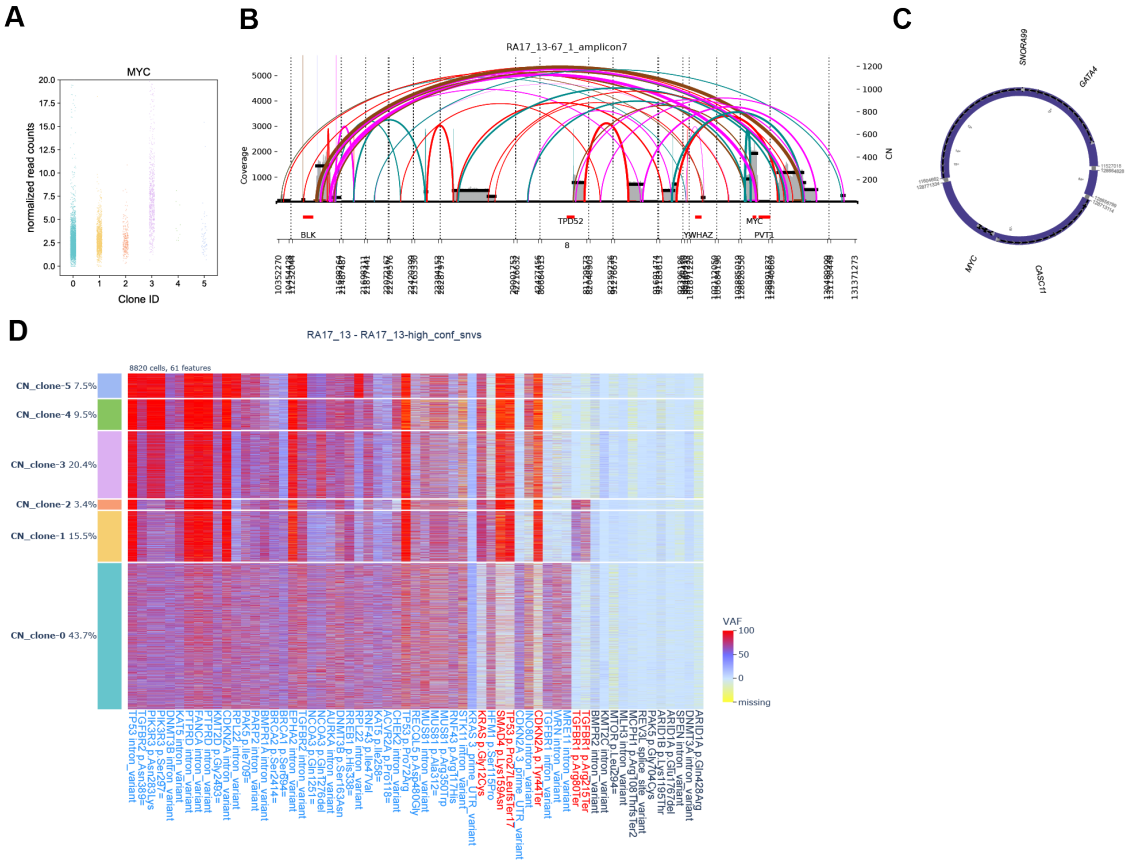

Figure 12. Detailed analysis of snDNA-seq and whole-genome sequencing data of RA17_13.

a. stripplot of the distribution of normalized read counts of *MYC* gene in single cells of RA17_13, split by clones, focused on clones with lower normalized total copy number **b.** structural variant view of an amplicon identified as ecDNA in RA17_13-67_1 by AmpliconArchitect. **c.** circular visualization of the ecDNA identified in RA17_13-67_1. **d.** RA17_13 single-cell mutation VAF heatmap.

4.3 focal amplifications

Aside from deletions, subclonal focal amplification was observed in advanced clones. In the tumor clones of case RA17_22, the *KRAS* mutant locus had allelic imbalance (single-cell VAF > 0.6); one clone had total loss of the WT allele **[Figure 10E].** Coupled with the total copy number gains **[Figure 10F]** this suggests that the mutant *KRAS* allele at least tripled in clone 2.

In case RA17_13, *MYC'*s total copy number (CN) state stayed in the normal range in clones 1 and 2 in the upper branch [**Figure 11E; Figure 12A]** for clones 3, 4, 5 in the lower branch, MYC was uniformly amplified to 20+ [**Figure 11H]**. Such an amplification pattern was inferred to be caused by extrachromosomal DNA (ecDNA) [**Figure 12B, C**] [169]. We interpret the single cells in the normal clone 0 with high *MYC* CN state to be tumor cells that our algorithm could not partition out due to the high total number of cells analyzed (n=8820), because they showed positive signal for the main tumor drivers as well **[Figure 12D].** Clones 1 and 2 in the upper branch had *MTOR* amplification instead. Intriguingly, clone 3, exclusively present in the primary tumor sampled, showed amplification signal for both *MYC* and *MTOR,* while also a uniquely high CN state in *GATA6*. Particularly, two populations, with and without high CN state of *MYC*, seem to coexist in clone 3. This suggests metastatic clonal heterogeneity is well represented at the primary site.

We caution readers that whole-genome doubling (WGD), prevalent in PDAC [170], is technically unobservable for this snDNA-seq technology due to its PCR basis. But paired whole-genome sequencing identified presence of clonal WGD in this case [to-be-published

work from my lab], therefore the real total copy number state of the abovementioned amplified genes could be higher than estimated here.

Taken together, through snDNA-seq we observed continuous genomic evolution through copy number loss and/or amplification towards particular genotypes. The driver events acquired and fixed early in tumor development likely culminated in increased genomic instability, which allowed for such CNV events to occur faster than clocklike SNV events [27] and allowed for rapid adaptation as the tumor invades and metastasizes to foreign environments.

5. Evolutionary dynamics through metastasis

Given the knowledge that CNVs are the main driver of ongoing evolution and thus clonal heterogeneity in pancreatic cancer, we next analyzed how this heterogeneity manifests in the context of metastasis. Among 11 patients of MSK's LWP program, we collected in total >= 3 samples per common metastatic site - peritoneum, liver, lung, lymph node (LN), diaphragm - and matched primary samples when available [**Figure 13A**].

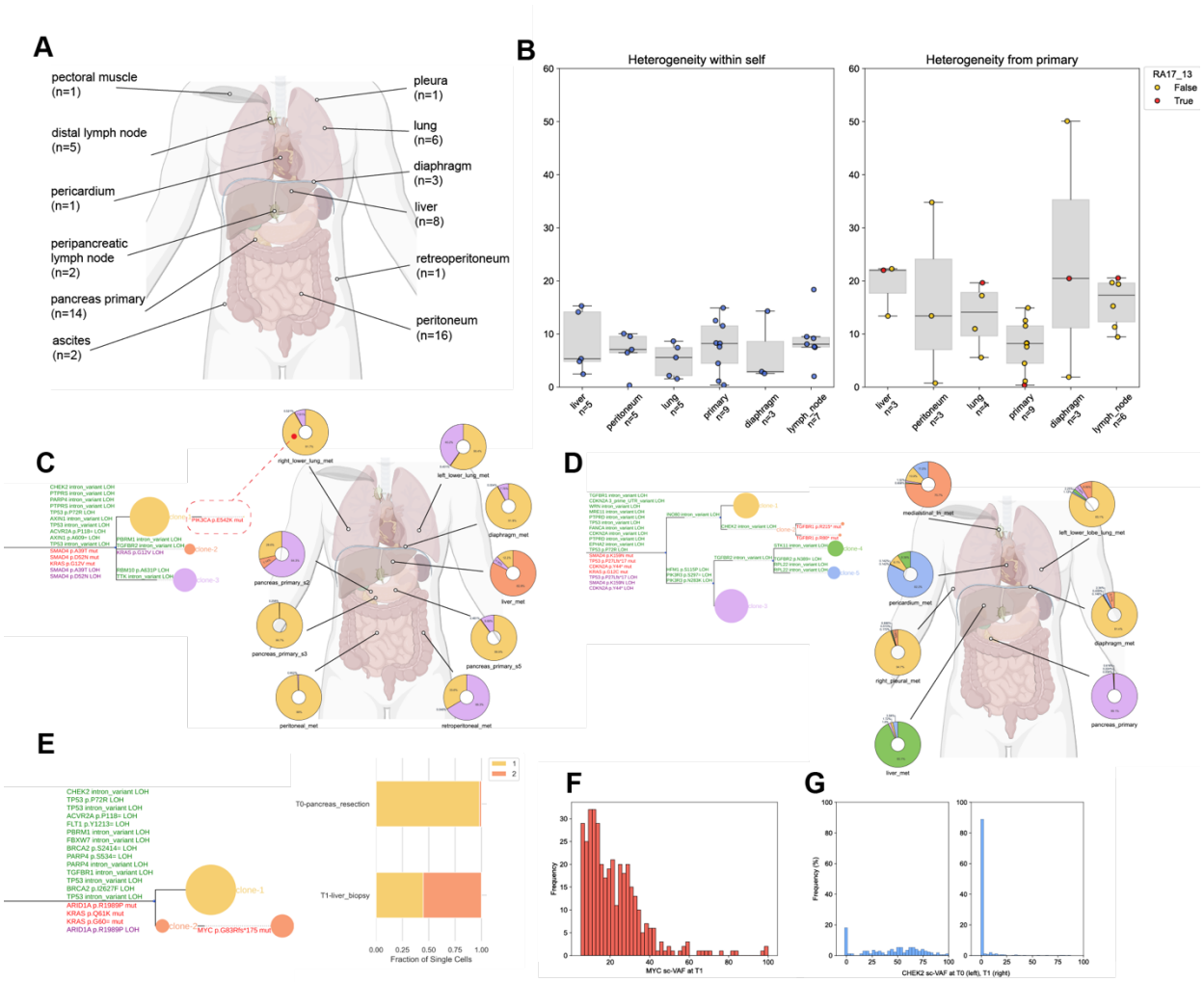

Figure 13. Evolutionary dynamics through metastasis and treatment.

a. Metastatic sites sampled in this study for the 9 autopsy cases. **b.** (left) intra-site heterogeneity and (right) primary-metastasis heterogeneity defined by average inter-cell CNV distances between the given one/two populations. For the right plot, the outlier case RA17_13 is colored as indicated in the legend. **c.** Left, RA17_22 single-cell clonal phylogeny. Right, clonal composition at each primary/metastatic site, mapped to spatial location (approximated). **d.** Left, RA17_13 single-cell clonal phylogeny. Right, clonal composition at each primary/metastatic site, mapped to spatial location (approximated). **e.** Left, BPA-1 single-cell clonal phylogeny. Right, clonal composition at each sample at each time point of collection. **f.** distribution of the somatic *MYC* mutation's single-cell VAF in the sample at T1. **g.** distribution of the germline *CHEK2* mutation's single-cell VAF in the sample at T0 (left) and T1 (right).

We first asked if there is any correlation between organ site and clonal pattern - specifically, if any site would have a stronger selective pressure and sculpt for (1) a more homogeneous clonal composition or (2) clones more advanced compared to the primary site. To answer these two questions, we calculated (1) intra-site heterogeneity and (2) primary-vs-metastasis difference by averaging inter-cell CNV distances between single cell populations from (1) the same site or (2) two separate sites [**Figure 13B**]. In the same patient, multiple samples of the same general site were combined (e.g. mesenteric nodule and omental nodule would be combined into peritoneal metastasis; multiple nodules at the liver would be combined), so that each data point would represent one site or primary-metastasis pair. There was significant inter-patient variability for both metrics [**Figure 13B**]. RA17_13, in particular, contributed high primary-metastasis heterogeneity across sites [**Figure 13B,** right]. This is likely due to the peculiar genomic instability which sculpted the complex CNV evolution paths described in the previous section [**Figure 11E**]. Two metastases in two other patients also had significant differences from their respective primary site: RA21_17's peritoneal metastasis and RA16_08's diaphragm metastasis. While the former was mostly caused by preservation of less evolved clones in the primary site, the latter was likely caused by elevated chromosomal instability generally [data now shown; will be included in our upcoming publication], which resulted in the many homdels described previously. These diverse storylines indicate that there is no general rule to organ-specific selective pressure which would make one organ prone to more homogeneous/advanced clonal composition.

Next, we investigated if organ sites proximal to each other tend to have more similar clonal composition, in the three most extensively sampled cases: RA16_29 (7 sites), RA17_13 (7 sites), RA17_22 (9 sites). RA16_29 and RA17_22 both presented with comb-like phylogeny indicating high-degree clonal sweep before metastasis and low inter-clone distance. This entailed high spatial homogeneity in these two cases and therefore no correlation between spatial location and clonal composition **[Figure 13C]**. An exception might be that RA17_22's right lung lower lobe metastasis had a unique subclonal *PIK3CA* p.E542K mutation **[Figure 13C, dashed red circle]**, although orthogonal bulk sequencing found it in both the lung site and another liver metastasis sample not present in this snDNA-seq cohort, suggesting that it is likely pre-metastatic [data not shown, will be included in our upcoming publication]. RA17_13, on the other hand, presented a clearly bifurcating phylogeny with two groups of metastatic sites each carrying clones from only one branch. Upper branch: mediastinal LN, right pleura, left lung lower lobe, diaphragm; lower branch: pericardium and liver [**Figure 13D**]. This again demonstrated no correlation between spatial proximity and clonal similarity.

6. Evolutionary dynamics through time

Finally, we performed a small pilot study of snDNA-seq as applied to longitudinally collected samples from four patients taken in association with the Tracking of Pancreatic Cancer Progression and Resistance (TOPCOAT) program at our institution. Three patients had metastatic disease (stage 4) at diagnosis (T0); phylogenies indicated a high clonal

homogeneity regardless of whether pancreas primary or the liver lesion were biopsied [**Figure 14A-C**], consistent with strong selection for the drivers present and one or more clonal sweeps prior to the biopsies taken. For a fourth patient BPA-1, samples were taken from resection (stage 2A, T0) and a recurrent liver lesion (stage 4, T1). Similar to the other three longitudinal cases, driver alterations including *KRAS* p.Q61Q and *ARID1A* p.R1989P SNVs, *CDKN2A* homdel and many LOH events involving *TP53, TGFBR1* and the germline heterozygous *BRCA2* p.I2627F locus, were fixed at T0 [**Figure 13E].** However, among 450 *KRAS* p.Q61K mutant cells at T1, 164 also showed a subclonal *MYC* p.G83Rfs*175 mutation (sample-specific CCF = 0.364), which was completely missing from T0 [**Figure 13E**]. The *MYC* mutation had a single-cell VAF (sc-VAF) densely distributed <0.2, which, coupled with a total copy number gain at the gene [**Figure 13F]**, suggests a mutation on the minor allele after total copy number gain. Although the germline CHEK2 SNP locus was inferred by our tree inference algorithm to undergo clonal LOH at T0, we noted that its sc-VAF distribution in tumor cells suggested a subclonal LOH at T0, which dominated at T1 **[Figure 13G]**, which likely corresponds to a focal CNV event after the large-scale CNVs that resulted in the many other LOHs mentioned above.

A caveat is none of these patient got targeted therapy which might better select for subclonal driver alteration as resistance mechanism. But the example BPA-1 showed that snDNA-seq would be able to capture such small subclonal events in routine biopsies.

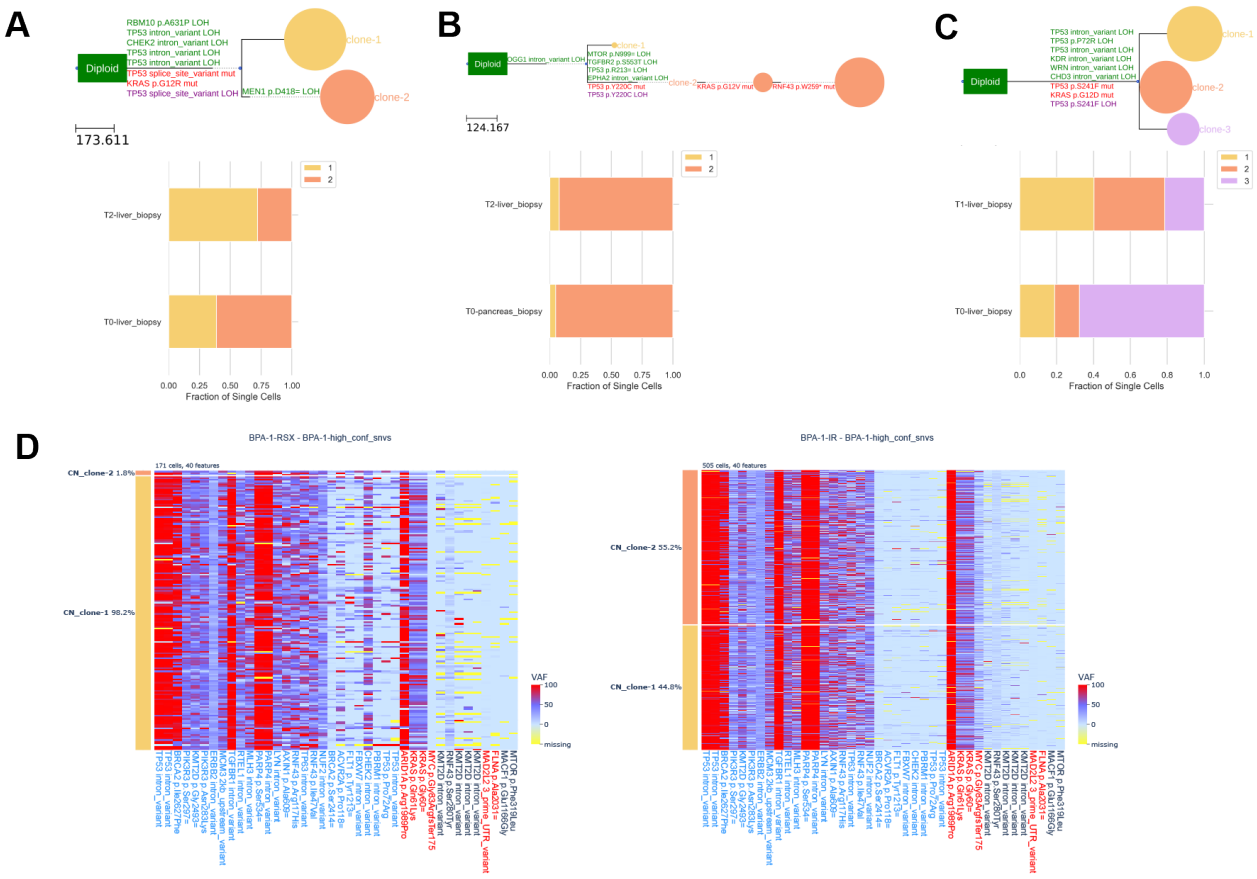

Figure 14. SnDNA-seq revealed longitudinal homogeneity in late PDAC diagnoses.

Single-cell clonal phylogenies of TP11 **(a)**, TP12 **(b)**, TP49 **(c)**.
d. BPA-1 single-cell mutation VAF heatmap at resection (RSX, left) and metastasis (interventional radiology, IR, right). Note the appearance of the subclonal *MYC* mutation at the IR biopsy.

Section 5. Conclusions, next steps, and outlook for the future

1. Technical advancement

We believe our targeted snDNA-seq approach's clinical utility is higher than other well-based, single-cell whole-genome sequencing methods for two major reasons. (1) As demonstrated by success of MSK-IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) [171], high-depth sequencing of most significant genes of diseases yields more clinically relevant information than low-depth sequencing of the genome, whose value is still mostly for basic biological research [50]. (2) The wet-lab workflow is largely automated, fast, and accommodates a wide range of sample types – from freshly dissociated cells to small amounts of frozen nuclei extracted from archival routine biopsy tissues – with the exception of formalin-fixed tissues possibly due to the DNA damage caused by the fixation process [these were considered negative data therefore not shown]. In comparison, well-based single-cell whole-genome sequencing [143], [144] is more demanding with respect to workflow complexity, sample quality (debris proportion, clumping proportion etc.) and quantity which limits its clinical application.

Notably, the combined wet/dry-lab workflow presented here is the first to apply such single-nucleus DNA sequencing to solid tumor, call SNVs and CNVs *de novo*, and resolve high-resolution single-cell clonal phylogeny from the data. Therefore, we believe this workflow carries high technical significance and has wide applicability.

2. New insights on genomic evolution of pancreatic cancer

Despite disappointing improvement in the past decades compared to other cancers, pancreatic cancer is set to enter the age of precision medicine with the advent of targeted therapies spearheaded by various *KRAS*/relevant pathway inhibitors [172], [173] along with better biomarkers for treatment selection, such as *BRCA* mutation/HRD phenotype for PARP inhibitor, immunogenicity for immunotherapy [174], [175]. It is the perfect time for such a high-resolution method as single-nucleus DNA-sequencing to elucidate the closest-to-ground-truth ***baseline*** genomic profile of pancreatic cancer in such contexts as early/late diagnosis, metastasis, across treatments. The main takeaways are:

1. driver SNVs such as oncogenic *KRAS* mutations, loss-of-function *TP53* mutations, are likely fixed by the time of invasion meaning that there is such strong selection force for them that alternate mutations to the same gene are rare and clonal sweep is expansive in the primary tumor. This is encouraging for *KRAS* targeting therapies as it suggests that resistant clones are at least not preexistent. Allele-specific (e.g. G12D-targeting) agents could potentially work well as the *KRAS* mutant allele is unique per tumor based on our dataset.
2. driver CNVs, such as *CDKN2A* and *SMAD4* homozygous deletion, *MYC* amplification, could be subclonal, meaning that they are not fixed in the tumor at invasion and are only present in a subset of cells.
3. further CNVs, likely due to increased genomic instability in the tumor, form the backbone for continuous evolution across space (metastasis) and time (treatment, or simply genetic drift). They would likely correlate with transcriptomic or

epigenomic changes that enable phenotypic changes as the tumor adapts to the new environment at distant sites and treatments.

As more trial data of precision medicine in pancreatic cancer accumulate, we foresee our snDNA-seq approach to be critical for identifying resistance mechanism with high sensitivity, and the data presented in this work would form a useful basis for comparison. Together, they would facilitate precision medicine efforts in PDAC and improve its clinical management.

3. Application of snDNA-seq in more diseases

The snDNA-seq workflow clearly has more applications than studying pancreatic cancer. At the time of writing this thesis I have already started collaboration with biologists from a variety of fields to apply the method. With Etay Ziv, MD, PhD (MSKCC) and colleagues, we are using snDNA-seq to dissect the clonal dynamics of low-grade and high-grade neuroendocrine tumors (NET) of the pancreas. With Rocio Vicario, PhD (MSKCC), Frederic Geissmann, MD, PhD (MSKCC) and colleagues, we are using snDNA-seq to study the clonal relationship of pathogenic microglia in the context of histocytosis and Alzheimer's disease. With Alexandra Miller, MD, PhD (New York University), Ingo Mellinghoff, MD (MSKCC) and colleagues, we are using snDNA-seq to study glioblastoma's clonal evolution across surgery and recurrence.

Bibliography

[1] "Cancer Facts & Figures 2023 | American Cancer Society." Accessed: May 19, 2024. [Online]. Available: https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/2023-cancer-facts-figures.html

[2] L. Rahib, M. R. Wehner, L. M. Matrisian, and K. T. Nead, "Estimated Projection of US Cancer Incidence and Death to 2040," *JAMA Netw. Open*, vol. 4, no. 4, p. e214708, Apr. 2021, doi: 10.1001/jamanetworkopen.2021.4708.

[3] "Annual Report to the Nation 2022," SEER. Accessed: May 19, 2024. [Online]. Available: https://seer.cancer.gov/report_to_nation/

[4] G. M. Petersen, "Familial pancreatic cancer," *Semin. Oncol.*, vol. 43, no. 5, pp. 548–553, Oct. 2016, doi: 10.1053/j.seminoncol.2016.09.002.

[5] J. F. Griffin, K. E. Poruk, and C. L. Wolfgang, "Pancreatic cancer surgery: past, present, and future," *Chin. J. Cancer Res.*, vol. 27, no. 4, pp. 332–348, Aug. 2015, doi: 10.3978/j.issn.1000-9604.2015.06.07.

[6] C. J. Halbrook, C. A. Lyssiotis, M. Pasca Di Magliano, and A. Maitra, "Pancreatic cancer: Advances and challenges," *Cell*, vol. 186, no. 8, pp. 1729–1754, Apr. 2023, doi: 10.1016/j.cell.2023.02.014.

[7] W. Park, A. Chawla, and E. M. O'Reilly, "Pancreatic Cancer: A Review," *JAMA*, vol. 326, no. 9, pp. 851–862, Sep. 2021, doi: 10.1001/jama.2021.13027.

[8] H. Jin, L. Wang, and R. Bernards, "Rational combinations of targeted cancer therapies: background, advances and challenges," *Nat. Rev. Drug Discov.*, vol. 22, no. 3, pp. 213–234, Mar. 2023, doi: 10.1038/s41573-022-00615-z.

[9] "Expanding the Reach of Precision Oncology by Drugging All KRAS Mutants - PubMed." Accessed: May 19, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/35046095/

[10] L. A. Rojas *et al.*, "Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer," *Nature*, vol. 618, no. 7963, pp. 144–150, Jun. 2023, doi: 10.1038/s41586-023-06063-y.

[11] E. A. Collisson, P. Bailey, D. K. Chang, and A. V. Biankin, "Molecular subtypes of pancreatic cancer," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 16, no. 4, pp. 207–220, Apr. 2019, doi: 10.1038/s41575-019-0109-y.

[12] "Genomics-Driven Precision Medicine for Advanced Pancreatic Cancer: Early Results from the COMPASS Trial | Clinical Cancer Research | American Association

for Cancer Research." Accessed: May 19, 2024. [Online]. Available: https://aacrjournals.org/clincancerres/article/24/6/1344/474/Genomics-Driven-Precision-Medicine-for-Advanced

[13] W. J. Ho, E. M. Jaffee, and L. Zheng, "The tumour microenvironment in pancreatic cancer — clinical challenges and opportunities," *Nat. Rev. Clin. Oncol.*, vol. 17, no. 9, pp. 527–540, Sep. 2020, doi: 10.1038/s41571-020-0363-5.

[14] B. T. Grünwald *et al.*, "Spatially confined sub-tumor microenvironments in pancreatic cancer," *Cell*, vol. 184, no. 22, pp. 5577-5592.e18, Oct. 2021, doi: 10.1016/j.cell.2021.09.022.

[15] M. H. Sherman and G. L. Beatty, "Tumor Microenvironment in Pancreatic Cancer Pathogenesis and Therapeutic Resistance," *Annu. Rev. Pathol.*, vol. 18, pp. 123–148, Jan. 2023, doi: 10.1146/annurev-pathmechdis-031621-024600.

[16] V. P. Balachandran *et al.*, "Identification of unique neoantigen qualities in long-term survivors of pancreatic cancer," *Nature*, vol. 551, no. 7681, pp. 512–516, Nov. 2017, doi: 10.1038/nature24462.

[17] M. Łuksza *et al.*, "Neoantigen quality predicts immunoediting in survivors of pancreatic cancer," *Nature*, vol. 606, no. 7913, pp. 389–395, Jun. 2022, doi: 10.1038/s41586-022-04735-9.

[18] J. Fan, M.-F. Wang, H.-L. Chen, D. Shang, J. K. Das, and J. Song, "Current advances and outlooks in immunotherapy for pancreatic ductal adenocarcinoma," *Mol. Cancer*, vol. 19, no. 1, p. 32, Feb. 2020, doi: 10.1186/s12943-020-01151-3.

[19] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nat. Med.*, vol. 10, no. 8, pp. 789–799, Aug. 2004, doi: 10.1038/nm1087.

[20] C. J. Der, T. G. Krontiris, and G. M. Cooper, "Transforming genes of human bladder and lung carcinoma cell lines are homologous to the ras genes of Harvey and Kirsten sarcoma viruses.," *Proc. Natl. Acad. Sci.*, vol. 79, no. 11, pp. 3637–3640, Jun. 1982, doi: 10.1073/pnas.79.11.3637.

[21] S. H. Friend *et al.*, "A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma," *Nature*, vol. 323, no. 6089, pp. 643–646, Oct. 1986, doi: 10.1038/323643a0.

[22] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz, and K. W. Kinzler, "Cancer genome landscapes," *Science*, vol. 339, no. 6127, pp. 1546–1558, Mar. 2013, doi: 10.1126/science.1235122.

[23] Druker Brian J. *et al.*, "Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia," *N. Engl. J. Med.*, vol. 355, no. 23, pp. 2408–2417, 2006, doi: 10.1056/NEJMoa062867.

[24] P. C. Nowell, “The Clonal Evolution of Tumor Cell Populations,” *Science*, vol. 194, no. 4260, pp. 23–28, Oct. 1976, doi: 10.1126/science.959840.

[25] L. M. F. Merlo, J. W. Pepper, B. J. Reid, and C. C. Maley, “Cancer as an evolutionary and ecological process,” *Nat. Rev. Cancer*, vol. 6, no. 12, pp. 924–935, Dec. 2006, doi: 10.1038/nrc2013.

[26] L. B. Alexandrov *et al.*, “Signatures of mutational processes in human cancer,” *Nature*, vol. 500, no. 7463, pp. 415–421, Aug. 2013, doi: 10.1038/nature12477.

[27] L. B. Alexandrov *et al.*, “The repertoire of mutational signatures in human cancer,” *Nature*, vol. 578, no. 7793, pp. 94–101, Feb. 2020, doi: 10.1038/s41586-020-1943-3.

[28] C. A. Iacobuzio-Donahue *et al.*, “DPC4 gene status of the primary carcinoma correlates with patterns of failure in patients with pancreatic cancer,” *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, vol. 27, no. 11, pp. 1806–1813, Apr. 2009, doi: 10.1200/JCO.2008.17.7188.

[29] A. Makohon-Moore and C. A. Iacobuzio-Donahue, “Pancreatic cancer biology and genetics from an evolutionary perspective,” *Nat. Rev. Cancer*, vol. 16, no. 9, pp. 553–565, Sep. 2016, doi: 10.1038/nrc.2016.66.

[30] S.-J. Dawson *et al.*, “Analysis of circulating tumor DNA to monitor metastatic breast cancer,” *N. Engl. J. Med.*, vol. 368, no. 13, pp. 1199–1209, Mar. 2013, doi: 10.1056/NEJMoa1213261.

[31] A. P. Makohon-Moore *et al.*, “Precancerous neoplastic cells can move through the pancreatic ductal system,” *Nature*, vol. 561, no. 7722, pp. 201–205, Sep. 2018, doi: 10.1038/s41586-018-0481-8.

[32] A. M. Braxton *et al.*, “3D genomic mapping reveals multifocality of human pancreatic precancers,” *Nature*, vol. 629, no. 8012, pp. 679–687, May 2024, doi: 10.1038/s41586-024-07359-3.

[33] S. Yachida *et al.*, “Distant metastasis occurs late during the genetic evolution of pancreatic cancer,” *Nature*, vol. 467, no. 7319, pp. 1114–1117, Oct. 2010, doi: 10.1038/nature09515.

[34] F. Notta *et al.*, “A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns,” *Nature*, vol. 538, no. 7625, pp. 378–382, Oct. 2016, doi: 10.1038/nature19823.

[35] N. McGranahan *et al.*, “Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution,” *Cell*, vol. 171, no. 6, pp. 1259-1271.e11, Nov. 2017, doi: 10.1016/j.cell.2017.10.001.

[36] A. P. Makohon-Moore *et al.*, "Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer," *Nat. Genet.*, vol. 49, no. 3, pp. 358–366, Mar. 2017, doi: 10.1038/ng.3764.

[37] M. Al Bakir *et al.*, "The evolution of non-small cell lung cancer metastases in TRACERx," *Nature*, vol. 616, no. 7957, pp. 534–542, Apr. 2023, doi: 10.1038/s41586-023-05729-x.

[38] Gerlinger Marco *et al.*, "Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing," *N. Engl. J. Med.*, vol. 366, no. 10, pp. 883–892, 2012, doi: 10.1056/NEJMoa1113205.

[39] Jamal-Hanjani Mariam *et al.*, "Tracking the Evolution of Non–Small-Cell Lung Cancer," *N. Engl. J. Med.*, vol. 376, no. 22, pp. 2109–2121, 2017, doi: 10.1056/NEJMoa1616288.

[40] A. M. Frankell *et al.*, "The evolution of lung cancer and impact of subclonal selection in TRACERx," *Nature*, vol. 616, no. 7957, pp. 525–533, Apr. 2023, doi: 10.1038/s41586-023-05783-5.

[41] N. McGranahan, F. Favero, E. C. de Bruin, N. J. Birkbak, Z. Szallasi, and C. Swanton, "Clonal status of actionable driver events and the timing of mutational processes in cancer evolution," *Sci. Transl. Med.*, vol. 7, no. 283, p. 283ra54, Apr. 2015, doi: 10.1126/scitranslmed.aaa1408.

[42] R. H. Hruban *et al.*, "K-ras oncogene activation in adenocarcinoma of the human pancreas. A study of 82 carcinomas using a combination of mutant-enriched polymerase chain reaction analysis and allele-specific oligonucleotide hybridization," *Am. J. Pathol.*, vol. 143, no. 2, pp. 545–554, Aug. 1993.

[43] S. A. Hahn *et al.*, "DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1," *Science*, vol. 271, no. 5247, pp. 350–353, Jan. 1996, doi: 10.1126/science.271.5247.350.

[44] M. Schutte *et al.*, "Identification by representational difference analysis of a homozygous deletion in pancreatic carcinoma that lies within the BRCA2 region.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 92, no. 13, pp. 5950–5954, Jun. 1995.

[45] C. Caldas *et al.*, "Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma," *Nat. Genet.*, vol. 8, no. 1, pp. 27–32, Sep. 1994, doi: 10.1038/ng0994-27.

[46] J. C. Venter *et al.*, "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, Feb. 2001, doi: 10.1126/science.1058040.

[47] S. Jones *et al.*, "Core signaling pathways in human pancreatic cancers revealed by global genomic analyses," *Science*, vol. 321, no. 5897, pp. 1801–1806, Sep. 2008, doi: 10.1126/science.1164368.

[48] T. J. Ley *et al.*, "DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome," *Nature*, vol. 456, no. 7218, pp. 66–72, Nov. 2008, doi: 10.1038/nature07485.

[49] D. C. Koboldt *et al.*, "Comprehensive molecular portraits of human breast tumours," *Nature*, vol. 490, no. 7418, pp. 61–70, Oct. 2012, doi: 10.1038/nature11412.

[50] L. A. Aaltonen *et al.*, "Pan-cancer analysis of whole genomes," *Nature*, vol. 578, no. 7793, pp. 82–93, Feb. 2020, doi: 10.1038/s41586-020-1969-6.

[51] M. L. Hoang *et al.*, "Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 35, pp. 9846–9851, Aug. 2016, doi: 10.1073/pnas.1607794113.

[52] F. Abascal *et al.*, "Somatic mutation landscapes at single-molecule resolution | Nature," *Nature*, vol. 593, no. 7859, pp. 405–410, May 2021, doi: 10.1038/s41586-021-03477-4.

[53] A. V. Biankin *et al.*, "Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes," *Nature*, vol. 491, no. 7424, pp. 399–405, Nov. 2012, doi: 10.1038/nature11547.

[54] A. K. Witkiewicz *et al.*, "Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets," *Nat. Commun.*, vol. 6, no. 1, p. 6744, Apr. 2015, doi: 10.1038/ncomms7744.

[55] Cancer Genome Atlas Research Network. Electronic address: andrew_aguirre@dfci.harvard.edu and Cancer Genome Atlas Research Network, "Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma," *Cancer Cell*, vol. 32, no. 2, pp. 185-203.e13, Aug. 2017, doi: 10.1016/j.ccell.2017.07.007.

[56] N. Waddell *et al.*, "Whole genomes redefine the mutational landscape of pancreatic cancer," *Nature*, vol. 518, no. 7540, Art. no. 7540, Feb. 2015, doi: 10.1038/nature14169.

[57] A. A. Connor *et al.*, "Integration of Genomic and Transcriptional Features in Pancreatic Cancer Reveals Increased Cell Cycle Progression in Metastases," *Cancer Cell*, vol. 35, no. 2, pp. 267-282.e7, Feb. 2019, doi: 10.1016/j.ccell.2018.12.010.

[58] A. Hayashi *et al.*, "Evolutionary Dynamics of Non-Coding Regions in Pancreatic Ductal Adenocarcinoma," Sep. 12, 2020, *bioRxiv*. doi: 10.1101/2020.09.11.294389.

[59] C. J. Tokheim, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, and R. Karchin, "Evaluating the evaluation of cancer driver genes," *Proc Natl Acad Sci USA*, vol. 113, 2016.

[60] L. A. Garraway and E. S. Lander, "Lessons from the cancer genome," *Cell*, vol. 153, 2013.

[61] T. Kurosaki, M. W. Popp, and L. E. Maquat, "Quality and quantity control of gene expression by nonsense-mediated mRNA decay," *Nat Rev Mol Cell Biol*, vol. 20, 2019.

[62] S. Cherry and K. W. Lynch, "Alternative splicing and cancer: insights, opportunities, and challenges from an expanding view of the transcriptome," *Genes Dev*, vol. 34, 2020.

[63] M. Luksza, "A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy," *Nature*, vol. 551, 2017.

[64] J. L. Humphris, "Hypermutation in pancreatic cancer," *Gastroenterology*, vol. 152, 2017.

[65] D. T. Le, "PD-1 blockade in tumors with mismatch-repair deficiency," *N Engl J Med*, vol. 372, 2015.

[66] M. E. Feigin, "Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma," *Nat Genet*, vol. 49, 2017.

[67] M. C. Whittle, "RUNX3 controls a metastatic switch in pancreatic ductal adenocarcinoma," *Cell*, vol. 161, 2015.

[68] S. Karmakar, "MicroRNA regulation of K-Ras in pancreatic cancer and opportunities for therapeutic intervention," *Semin Cancer Biol*, vol. 54, 2019.

[69] A. A. Connor, "Association of distinct mutational signatures with correlates of increased immune activity in pancreatic ductal adenocarcinoma," *JAMA Oncol*, vol. 3, 2017.

[70] H. Sakamoto, "The evolutionary origins of recurrent pancreatic cancer," *Cancer Discov*, vol. 10, 2020, doi: 10.1158/2159-8290.CD-19-1508.

[71] A. Hayashi, "Evolutionary dynamics of non-coding regions in pancreatic ductal adenocarcinoma. Preprint at," *bioRxiv*, 2020, doi: 10.1101/2020.09.11.294389.

[72] P. J. Stephens *et al.*, "Massive Genomic Rearrangement Acquired in a Single Catastrophic Event during Cancer Development," *Cell*, vol. 144, no. 1, pp. 27–40, Jan. 2011, doi: 10.1016/j.cell.2010.11.055.

[73] C. Almoguera, D. Shibata, K. Forrester, J. Martin, N. Arnheim, and M. Perucho, "Most human carcinomas of the exocrine pancreas contain mutant c-K-ras genes," *Cell*, vol. 53, no. 4, pp. 549–554, May 1988, doi: 10.1016/0092-8674(88)90571-5.

[74] L. Buscail, B. Bournet, and P. Cordelier, "Role of oncogenic KRAS in the diagnosis, prognosis and treatment of pancreatic cancer," *Nat Rev Gastroenterol Hepatol*, vol. 17, 2020.

[75] P. A. Hart, "Type 3c (pancreatogenic) diabetes mellitus secondary to chronic pancreatitis and pancreatic cancer," *Lancet Gastroenterol Hepatol*, vol. 1, 2016.

[76] C. Shi, "KRAS2 mutations in human pancreatic acinar-ductal metaplastic lesions are limited to those with PanIN: implications for the human pancreatic cancer cell of origin," *Mol Cancer Res*, vol. 7, 2009.

[77] M. Kanda, "Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia," *Gastroenterology*, vol. 142, 2012.

[78] M. R. Burgess, "KRAS allelic imbalance enhances fitness and modulates MAP kinase dependence in cancer," *Cell*, vol. 168, 2017.

[79] S. Mueller, "Evolutionary routes and KRAS dosage define pancreatic cancer phenotypes," *Nature*, vol. 554, 2018.

[80] E. M. O'Reilly and J. F. Hechtman, "Tumour response to TRK inhibition in a patient with pancreatic adenocarcinoma harbouring an NTRK gene fusion," *Ann Oncol*, vol. 30, 2019.

[81] C. Heining, "NRG1 fusions in KRAS wild-type pancreatic cancer," *Cancer Discov*, vol. 8, 2018.

[82] M. R. Jones, "NRG1 gene fusions are recurrent, clinically actionable gene rearrangements in KRAS wild-type pancreatic ductal adenocarcinoma," *Clin Cancer Res*, vol. 25, 2019.

[83] J. Gil and G. Peters, "Regulation of the INK4b-ARF-INK4a tumour suppressor locus: all for one or one for all," *Nat Rev Mol Cell Biol*, vol. 7, 2006.

[84] W. Y. Kim and N. E. Sharpless, "The regulation of INK4/ARF in cancer and aging," *Cell*, vol. 127, 2006.

[85] M. Schutte, "Abrogation of the Rb/p16 tumor-suppressive pathway in virtually all pancreatic carcinomas," *Cancer Res*, vol. 57, 1997.

[86] W. Hosoda, "Genetic analyses of isolated high-grade pancreatic intraepithelial neoplasia (HG-PanIN) reveal paucity of alterations in TP53 and SMAD4," *J Pathol*, vol. 242, 2017.

[87] B. Vogelstein, D. Lane, and A. J. Levine, "Surfing the p53 network," *Nature*, vol. 408, 2000.

[88] S. Yachida, "Clinical significance of the genetic landscape of pancreatic cancer and implications for identification of potential long-term survivors," *Clin Cancer Res*, vol. 18, 2012.

[89] J. P. Morton, "Mutant p53 drives metastasis and overcomes growth arrest/senescence in pancreatic cancer," *Proc Natl Acad Sci USA*, vol. 107, 2010.

[90] J. P. Morton, D. S. Klimstra, M. E. Mongeau, and B. C. Lewis, "Trp53 deletion stimulates the formation of metastatic pancreatic tumors," *Am J Pathol*, vol. 172, 2008.

[91] Z. Shu, S. Row, and W. M. Deng, "Endoreplication: the good, the bad, and the ugly," *Trends Cell Biol*, vol. 28, 2018.

[92] C. J. David and J. Massagué, "Contextual determinants of TGFβ action in development, immunity and cancer," *Nat Rev Mol Cell Biol*, vol. 19, 2018.

[93] Y. Togashi, "Homozygous deletion of the activin A receptor, type IB gene is associated with an aggressive cancer phenotype in pancreatic cancer," *Mol Cancer*, vol. 13, 2014.

[94] W. Huang, "Pattern of invasion in human pancreatic cancer organoids is associated with loss of SMAD4 and clinical outcome," *Cancer Res*, vol. 80, 2020.

[95] J. Hong *et al.*, "Convergence for Inactivation of TGFβ Signaling Is a Common Feature of Advanced Pancreatic Cancer," Jan. 30, 2024, *medRxiv*. doi: 10.1101/2024.01.30.24301554.

[96] G. M. O'Kane, "GATA6 expression distinguishes classical and basal-like subtypes in advanced pancreatic cancer," *Clin Cancer Res*, vol. 26, 2020.

[97] K. L. Aung, "Genomics-driven precision medicine for advanced pancreatic cancer: early results from the COMPASS trial," *Clin Cancer Res*, vol. 24, 2018.

[98] R. C. Centore, G. J. Sandoval, L. M. Mendes Soares, C. Kadoch, and H. M. Chan, "Mammalian SWI/SNF chromatin remodeling complexes: emerging mechanisms and therapeutic strategies," *Trends Genet*, vol. 36, 2020.

[99] L. Wang and A. Shilatifard, "UTX mutations in human cancer," *Cancer Cell*, vol. 35, 2019.

[100] A. H. Shain, "Convergent structural alterations define SWItch/Sucrose NonFermentable (SWI/SNF) chromatin remodeler as a central tumor suppressive complex in pancreatic cancer," *Proc Natl Acad Sci USA*, vol. 109, 2012.

[101] C. R. Clapier, J. Iwasa, B. R. Cairns, and C. L. Peterson, "Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes," *Nat Rev Mol Cell Biol*, vol. 18, 2017.

[102] H. Rafati, "Repressive LTR nucleosome positioning by the BAF complex is required for HIV latency," *PLoS Biol*, vol. 9, 2011.

[103] C. C. Sze and A. Shilatifard, "MLL3/MLL4/COMPASS family on epigenetic regulation of enhancer function and cancer," *Cold Spring Harb Perspect Med*, vol. 6, 2016.

[104] A. Hayashi, "A unifying paradigm for transcriptional heterogeneity and squamous features in pancreatic ductal adenocarcinoma," *Nat Cancer*, vol. 1, 2020.

[105] T. D. Somerville, "Squamous trans-differentiation of pancreatic cancer cells promotes stromal inflammation," *eLife*, vol. 9, 2020.

[106] T. D. D. Somerville, "TP63-mediated enhancer reprogramming drives the squamous subtype of pancreatic ductal adenocarcinoma," *Cell Rep*, vol. 25, 2018.

[107] M. Chan-Seng-Yue, "Transcription phenotypes of pancreatic cancer are driven by genomic events during tumor evolution," *Nat Genet*, vol. 52, 2020, doi: 10.1038/s41588-019-0566-9.

[108] Y. Zhong, "GATA6 activates Wnt signaling in pancreatic cancer by negatively regulating the Wnt antagonist Dickkopf-1," *PLoS ONE*, vol. 6, 2011.

[109] M. H. Bailey, "Comprehensive characterization of cancer driver genes and mutations," *Cell*, vol. 173, 2018, doi: 10.1016/j.cell.2018.02.060.

[110] C. V. Dang, "MYC on the path to cancer," *Cell*, vol. 149, 2012.

[111] N. E. Baker, "Emerging mechanisms of cell competition," *Nat Rev Genet*, vol. 21, 2020.

[112] N. J. Roberts, "ATM mutations in patients with hereditary pancreatic cancer," *Cancer Discov*, vol. 2, 2012.

[113] M. A. Lowery, "Prospective evaluation of germline alterations in patients with exocrine pancreatic neoplasms," *J Natl Cancer Inst*, vol. 110, 2018.

[114] C. Tomasetti, B. Vogelstein, and G. Parmigiani, "Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation," *Proc Natl Acad Sci USA*, vol. 110, 2013.

[115] J. Kleeff, "Chronic pancreatitis," *Nat Rev Prim*, vol. 3, 2017.

[116] J. Ling, "Kras G12D-induced IKK2/β/NF-κB activation by IL-1α and p62 feedforward loops is required for development of pancreatic ductal adenocarcinoma," *Cancer Cell*, vol. 21, 2012.

[117] J. L. Kopp, "Identification of Sox9-dependent acinar-to-ductal reprogramming as the principal mechanism for initiation of pancreatic ductal adenocarcinoma," *Cancer Cell*, vol. 22, 2012.

[118] F. McAllister, "Oncogenic Kras activates a hematopoietic-to-epithelial IL-17 signaling axis in preinvasive pancreatic neoplasia," *Cancer Cell*, vol. 25, 2014.

[119] Y. Luo, "Oncogenic KRAS reduces expression of FGF21 in acinar cells to promote pancreatic tumorigenesis in mice on a high-fat diet," *Gastroenterology*, vol. 157, 2019.

[120] H. F. Dvorak, "Similarities between tumor stroma generation and wound healing," *N Engl J Med*, vol. 315, 1986.

[121] G. C. Gurtner, S. Werner, Y. Barrandon, and M. T. Longaker, "Wound repair and regeneration," *Nature*, vol. 35, 2008.

[122] J. G. Reiter, "Minimal functional driver gene heterogeneity among untreated metastases," *Science*, vol. 361, 2018.

[123] J. G. Reiter, "An analysis of genetic heterogeneity in untreated cancers," *Nat Rev Cancer*, vol. 19, 2019, doi: 10.1038/s41568-019-0185-x.

[124] O. G. McDonald, "Epigenomic reprogramming during pancreatic cancer progression links anabolic glucose metabolism to distant metastasis," *Nat Genet*, vol. 49, 2017.

[125] C. A. Iacobuzio-Donahue, C. Michael, P. Baez, R. Kappagantula, J. E. Hooper, and T. J. Hollman, "Cancer biology as revealed by the research autopsy," *Nat. Rev. Cancer*, vol. 19, no. 12, pp. 686–697, Dec. 2019, doi: 10.1038/s41568-019-0199-4.

[126] N. Navin, "Tumour evolution inferred by single-cell sequencing," *Nature*, vol. 472, 2011, doi: 10.1038/nature09807.

[127] C. Gawad, W. Koh, and S. R. Quake, "Single-cell genome sequencing: current state of the science," *Nat. Rev. Genet.*, vol. 17, no. 3, Art. no. 3, Mar. 2016, doi: 10.1038/nrg.2015.16.

[128] D. Liu *et al.*, "Evolution of delayed resistance to immunotherapy in a melanoma responder," *Nat. Med.*, vol. 27, no. 6, pp. 985–992, Jun. 2021, doi: 10.1038/s41591-021-01331-8.

[129] B. Lim, Y. Lin, and N. Navin, “Advancing Cancer Research and Medicine with Single-Cell Genomics,” *Cancer Cell*, vol. 37, no. 4, pp. 456–470, Apr. 2020, doi: 10.1016/j.ccell.2020.03.008.

[130] H. Zhang *et al.*, “Application of high-throughput single-nucleus DNA sequencing in pancreatic cancer,” *Nat. Commun.*, vol. 14, no. 1, p. 749, Feb. 2023, doi: 10.1038/s41467-023-36344-z.

[131] F. Lan, B. Demaree, N. Ahmed, and A. R. Abate, “Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding,” *Nat. Biotechnol.*, vol. 35, no. 7, pp. 640–646, Jul. 2017, doi: 10.1038/nbt.3880.

[132] M. Pellegrino *et al.*, “High-throughput single-cell DNA sequencing of acute myeloid leukemia tumors with droplet microfluidics,” *Genome Res.*, vol. 28, no. 9, pp. 1345–1352, Sep. 2018, doi: 10.1101/gr.232272.117.

[133] Y. Gao, “V211D mutation in MEK1 causes resistance to MEK inhibitors in colon cancer,” *Cancer Discov*, vol. 9, 2019, doi: 10.1158/2159-8290.CD-19-0356.

[134] O. Marin-Bejar, “Evolutionary predictability of genetic versus nongenetic resistance to anticancer drugs in melanoma,” *Cancer Cell*, vol. 39, 2021, doi: 10.1016/j.ccell.2021.05.015.

[135] Y. Zhao, “Diverse alterations associated with resistance to KRAS(G12C) inhibition,” *Nature*, vol. 599, 2021, doi: 10.1038/s41586-021-04065-2.

[136] L. A. Miles, “Single-cell mutation analysis of clonal evolution in myeloid malignancies,” *Nature*, vol. 587, 2020, doi: 10.1038/s41586-020-2864-x.

[137] C. M. Mcmahon, “Clonal selection with RAS pathway activation mediates secondary clinical resistance to selective FLT3 inhibition in acute myeloid leukemia secondary resistance to selective FLT3 inhibition in AML,” *Cancer Discov*, vol. 9, 2019, doi: 10.1158/2159-8290.CD-18-1453.

[138] K. Morita, “Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics,” *Nat Commun*, vol. 11, 2020, doi: 10.1038/s41467-020-19119-8.

[139] M. Slyper, “A single-cell and single-nucleus RNA-Seq toolbox for fresh and frozen human tumors,” *Nat Med*, vol. 26, 2020, doi: 10.1038/s41591-020-0844-1.

[140] “Mission Bio. Performance of the Tapestri ® Platform for Single-Cell Targeted DNA Sequencing. Tapestri White Paper https://missionbio.com/wp-content/uploads/2019/10/WhitePaper_MissionBio_TapestriPlatform_RevA.pdf (2019).”, [Online]. Available: https://missionbio.com/wp-content/uploads/2019/10/WhitePaper_MissionBio_TapestriPlatform_RevA.pdf

[141] A. Hayashi, J. Hong, and C. A. Iacobuzio-Donahue, “The pancreatic cancer genome revisited,” *Nat Rev Gastroenterol Hepatol*, vol. 18, 2021, doi: 10.1038/s41575-021-00463-z.

[142] C. G. Fischer, “Intraductal papillary mucinous neoplasms arise from multiple independent clones, each with distinct mutations,” *Gastroenterology*, vol. 157, 2019, doi: 10.1053/j.gastro.2019.06.001.

[143] E. Laks, “Clonal decomposition and DNA replication states defined by scaled single,” *Cell Genome Seq. Cell*, vol. 179, 2019.

[144] D. C. Minussi, “Breast tumours maintain a reservoir of subclonal diversity during expansion,” *Nature*, vol. 592, 2021, doi: 10.1038/s41586-021-03357-x.

[145] R. Satija, J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev, “Spatial reconstruction of single-cell gene expression data,” *Nat Biotechnol*, vol. 33, 2015, doi: 10.1038/nbt.3192.

[146] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species,” *Nat Biotechnol*, vol. 36, 2018, doi: 10.1038/nbt.4096.

[147] J. Kuipers, K. Jahn, B. J. Raphael, and N. Beerenwinkel, “Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors,” *Genome Res*, vol. 27, 2017, doi: 10.1101/gr.220707.117.

[148] H. Zafar, N. Navin, K. Chen, and L. Nakhleh, “SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data,” *Genome Res*, vol. 29, 2019, doi: 10.1101/gr.243121.118.

[149] J. Singer, J. Kuipers, K. Jahn, and N. Beerenwinkel, “Single-cell mutation identification via phylogenetic inference,” *Nat Commun*, vol. 9, 2018, doi: 10.1038/s41467-018-07627-7.

[150] “Kuipers, J., Tuncel, M. A., Ferreira, P., Jahn, K. & Beerenwinkel, N. Single-cell copy number calling and event history reconstruction. Preprint at bioRxiv.https://doi.org/10.1101/2020.04.28.065755 (2020).”.

[151] S. Malikic, K. Jahn, J. Kuipers, S. C. Sahinalp, and N. Beerenwinkel, “Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data,” *Nat Commun*, vol. 10, 2019, doi: 10.1038/s41467-019-10737-5.

[152] G. Satas, S. Zaccaria, G. Mon, and B. J. Raphael, “SCARLET: single-cell tumor phylogeny inference with copy-number constrained mutation losses,” *Cell Syst*, vol. 10, 2020, doi: 10.1016/j.cels.2020.04.001.

[153] E. Sollier, J. Kuipers, K. Takahashi, N. Beerenwinkel, and K. Jahn, "COMPASS: joint copy number and mutation phylogeny reconstruction from amplicon single-cell sequencing data," *Nat. Commun.*, vol. 14, no. 1, p. 4921, Aug. 2023, doi: 10.1038/s41467-023-40378-8.

[154] B. Demaree, "Joint profiling of DNA and proteins in single cells to dissect genotype-phenotype associations in leukemia," *Nat Commun*, vol. 12, 2021, doi: 10.1038/s41467-021-21810-3.

[155] F. Mölder *et al.*, "Sustainable data analysis with Snakemake," Apr. 19, 2021, *F1000Research*: 10:33. doi: 10.12688/f1000research.29032.2.

[156] I. Martincorena and P. J. Campbell, "Somatic mutation in cancer and normal cells," *Science*, vol. 349, no. 6255, pp. 1483–1489, Sep. 2015, doi: 10.1126/science.aab4082.

[157] M. Kimura, "The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations," *Genetics.*, vol. 61, 1969, doi: 10.1093/genetics/61.4.893.

[158] J. Demeulemeester, S. C. Dentro, M. Gerstung, and P. Loo, "Biallelic mutations in cancer genomes reveal local mutational determinants," *Nat Genet*, vol. 54, 2022, doi: 10.1038/s41588-021-01005-8.

[159] A. McPherson, A. Roth, E. Laks, T. Masud, A. Bashashati, and A. W. Zhang, "Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer," *Nat Genet*, vol. 48, 2016, doi: 10.1038/ng.3573.

[160] M. El-Kebir, "SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error," *Bioinformatics.*, vol. 34, 2018, doi: 10.1093/bioinformatics/bty589.

[161] S. Ciccolella, M. Soto Gomez, M. D. Patterson, G. Della Vedova, I. Hajirasouliha, and P. Bonizzoni, "gpps: an ILP-based approach for inferring cancer progression with mutation losses from single cell data," *BMC Bioinformatics*, vol. 21, 2020.

[162] J. S. Farris, "Phylogenetic analysis under Dollo's Law," *Syst Biol*, vol. 26, 1977, doi: 10.1093/sysbio/26.1.77.

[163] Z. Chen, F. Gong, L. Wan, and L. Ma, "BiTSC 2: Bayesian inference of tumor clonal tree by joint analysis of single-cell SNV and CNA data," *Brief Bioinforma*, vol. 23, 2022, doi: 10.1093/bib/bbac092.

[164] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, and M. D. Robinson, "Eleven grand challenges in single-cell data science," *Genome Biol*, vol. 21, 2020, doi: 10.1186/s13059-020-1926-6.

[165] P. Sashittal, H. Zhang, C. A. Iacobuzio-Donahue, and B. J. Raphael, "ConDoR: tumor phylogeny inference with a copy-number constrained mutation loss model," *Genome Biol.*, vol. 24, no. 1, p. 272, Nov. 2023, doi: 10.1186/s13059-023-03106-5.

[166] M. A. Myers, B. J. Arnold, V. Bansal, K. M. Mullen, S. Zaccaria, and B. J. Raphael, "HATCHet2: clone- and haplotype-specific copy number inference from bulk tumor sequencing data," Jul. 15, 2023, *bioRxiv*. doi: 10.1101/2023.07.13.548855.

[167] D. Sankoff, "Minimal Mutation Trees of Sequences," *SIAM J. Appl. Math.*, vol. 28, no. 1, pp. 35–42, 1975.

[168] H. Ijichi *et al.*, "Aggressive pancreatic ductal adenocarcinoma in mice caused by pancreas-specific blockade of transforming growth factor-beta signaling in cooperation with active Kras expression," *Genes Dev.*, vol. 20, no. 22, pp. 3147–3160, Nov. 2006, doi: 10.1101/gad.1475506.

[169] K. M. Turner *et al.*, "Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity," *Nature*, vol. 543, no. 7643, pp. 122–125, Mar. 2017, doi: 10.1038/nature21356.

[170] C. M. Bielski *et al.*, "Genome doubling shapes the evolution and prognosis of advanced cancers," *Nat. Genet.*, vol. 50, no. 8, Art. no. 8, Aug. 2018, doi: 10.1038/s41588-018-0165-1.

[171] A. Zehir *et al.*, "Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients," *Nat. Med.*, vol. 23, no. 6, pp. 703–713, Jun. 2017, doi: 10.1038/nm.4333.

[172] U. N. Wasko *et al.*, "Tumour-selective activity of RAS-GTP inhibition in pancreatic cancer," *Nature*, pp. 1–10, Apr. 2024, doi: 10.1038/s41586-024-07379-z.

[173] M. Holderfield *et al.*, "Concurrent inhibition of oncogenic and wild-type RAS-GTP for cancer therapy," *Nature*, pp. 1–8, Apr. 2024, doi: 10.1038/s41586-024-07205-6.

[174] W. Park *et al.*, "Genomic Methods Identify Homologous Recombination Deficiency in Pancreas Adenocarcinoma and Optimize Treatment Selection," *Clin. Cancer Res.*, vol. 26, no. 13, pp. 3239–3247, Jul. 2020, doi: 10.1158/1078-0432.CCR-20-0418.

[175] W. Park *et al.*, "Phase 2 trial of pembrolizumab and olaparib (POLAR) maintenance for patients (pts) with metastatic pancreatic cancer (mPDAC): Two cohorts B non-core homologous recombination deficiency (HRD) and C exceptional response to platinum-therapy.," *J. Clin. Oncol.*, vol. 41, no. 16_suppl, pp. 4140–4140, Jun. 2023, doi: 10.1200/JCO.2023.41.16_suppl.4140.