5-9-2024

Novel bioinformatics tools for biomarker discovery in prostate cancer

Jace Webster
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Human and Statistical Genetics

Dissertation Examination Committee:
Christopher Maher, Chair
Aadel Chaudhuri
Malachi Griffith
Russell Pachynski
Jin Zhang

Novel Bioinformatics Tools for Biomarker Discovery in Prostate Cancer

by
Jace Webster

A dissertation presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2024
St. Louis, Missouri

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 2

Chapter 3

Chapter 4

Acknowledgments

I am immensely grateful to have had the opportunity to pursue my doctoral studies at Washington University in St. Louis. Any success I may have achieved here was only possible due to the never-ending support I have received, for which I will always be grateful.

First, I would like to thank my thesis advisor, Dr. Christopher Maher. Chris has been the best PI I could have asked for. He has been endlessly patient and encouraging, always giving timely and useful feedback on my projects. Chris always made time for me and was willing to talk about anything. I would also like to thank Dr. Nicole Maher, who has been very supportive and is a great source of strength in the lab. The two of them even laughed at my jokes sometimes.

I would also like to thank the many other faculty members I have worked with, especially those on my thesis committee. Drs. Aadel Chaudhuri, Malachi Griffith, Russell Pachynski and Jin Zhang have been very supportive throughout this process and have always been willing to provide valuable feedback. My work would not have been possible without them. I am also grateful for the guidance I received from faculty at BYU, particularly Dr. Marc Hansen, for giving me my first chance to work in a research lab, and Dr. Perry Ridge, for first introducing me to the field of bioinformatics and encouraging me to pursue it.

I am grateful for all the excellent members of the Maher Lab, both past and present. While each member of the lab has been helpful in their own way, I am especially grateful for Dr. Ha Dang, Dr. Sidi Zhao, Muheng Liao, Emily Coonrod, Gejae Jeffers and Hung Mai for the helpful support they gave me in my work and for inviting me to be involved in their exciting projects. I was able to learn a great deal by working with each of these wonderful people and my graduate school experience would not have been the same without them.

I am also grateful for the institutional support I have received from WashU, the Roy and Diana Vagelos Division of Biology & Biomedical Sciences, the Human and Statistical Genetics program and the NIH (including grant F31 CA265010-01). There have been many wonderful faculty members, administrative employees and staff who have made this journey possible. Although they may be unaware of their contributions to my graduate work, my research has also greatly benefited from the support to my morale and mental health that has been graciously offered by St. Louis City SC, Netflix, Disney+, the QuickTrip gas station at Gravois & Elgin Ave. and Tacos Wey & Grill, for which I am grateful.

Finally, I need to express my deep gratitude to my family. My wife Emily has been an ever-present support throughout this entire process. She has always encouraged me to be my best self, often seeing more potential in myself than I do. I will forever be grateful for how she has patiently helped me throughout grad school and for doing everything possible to allow me to focus on my academic work when needed. This work is as much her success as mine. I am also thankful for my amazing children. Being a father to Lucy, Theo and Ruth has been a truly humbling experience. They are a constant source of joy and a wonderful reminder that there is more to life than school or work. Their never-ending curiosity is what science is all about. I am grateful to my parents, Bret and Alisa, who have always encouraged me to never give up and to value and appreciate my education. I am also grateful to my in-laws, Paul and Tanya, who have helped St. Louis feel like home and have always been willing to offer their support.

I fear I am indebted to far too many people to be able to list them all here, but I am keenly aware that I could not have done this alone. The support and guidance I have received throughout my life from family, friends and mentors has been crucial in helping me become who I am today.

Jace Webster

Washington University in St. Louis

May 2024

ABSTRACT OF THE DISSERTATION

Novel bioinformatics tools for biomarker discovery in prostate cancer

by

Jace Webster

Doctor of Philosophy in Biology and Biomedical Sciences

Human and Statistical Genetics

Washington University in St. Louis, 2024

Professor Christopher Maher, PI

Professor Malachi Griffith, Chair

Prostate cancer (PCa) accounts for 29% of all expected cancer diagnoses in men in 2024, but patients presenting with different disease stages can have significantly different outcomes. Patients with indolent PCa may experience little to no impact on their quality of life and have a 5-year survival as high as 98%, but progression to aggressive disease causes 5-year survival to plummet to 30%. Patients with the most lethal form of the disease, metastatic castration-resistant PCa (mCRPC), have a median survival of only 5.5 months if they become resistant to treatment. Due to this clinical heterogeneity, it is critical to quickly and accurately stratify patients to match them with the appropriate treatment plans. To address this need, this thesis focuses on the development of novel tools that may be applied to diagnostic and prognostic biomarker detection in PCa by 1) creating a pipeline to aid in analysis of liquid biopsies, 2) developing a tool for discovering fusion-derived circular RNAs as potential biomarkers and 3) identifying an epigenetic signature for stratification of localized PCa.

Chapter 1: Introduction

1.1 Overview

This thesis focuses on the development of novel bioinformatics tools for biomarker discovery in prostate cancer. Prostate cancer is highly heterogeneous and therefore a variety of different considerations must be accounted for when attempting to identify biomarkers in early-stages vs late-stages of the disease. As a result, this work approaches biomarker discovery with a wide variety of strategies and includes methods relating to 1) structural variant detection in circulating tumor DNA, 2) machine learning algorithms applied to methylation profiles and 3) detection of circular RNAs formed by fusion transcripts. This chapter describes the heterogeneity found in prostate cancer and provides a foundation for understanding why each of these unique approaches may be particularly advantageous when considering specific aspects of the disease.

1.2 Prostate cancer

Prostate cancer (PCa) presents as a clinically heterogenous disease, accounting for 29% of all expected cancer diagnoses in men in 2024[1]. The five-year survival rate for localized, indolent PCa exceeds 98%[2], but the five-year survival for patients that advance to metastatic PCa declines to approximately 30%, resulting in PCa being the second leading cause of cancer-related deaths in men[3]. Metastatic castration resistant PCa (mCRPC), or disease that continues to progress despite surgery (castration) or hormone therapy (chemical castration), is considered the most lethal form of the disease. Virtually all mCRPC patients become resistant to standard treatments, at which point the median survival is only 5.5 months[3,4]. It has also been noted that disease prevalence differs among different regions and populations, with African Americans known to be more likely to be diagnosed and have worse clinical outcomes than Caucasian males[5]. Owing to the prevalence of PCa and considerable variation in outcomes across disease stages and populations, there is an urgent need for more comprehensive patient stratification methods.

1.2.1 Disease progression and treatment

Indolent and Aggressive Prostate Cancer

Approximately 80% of all PCa diagnoses occur when the disease is non-metastatic[6]. Those that present with indolent disease are often asymptomatic, with the most frequent complaints being difficulty with urination and nocturia[7]. Low-risk PCa patients are often monitored using strategies such as watchful waiting or active surveillance[8]. Active surveillance is a treatment strategy where surgery and other treatment options are deferred, instead focusing on proactive monitoring of disease progression. This typically involves regularly scheduled prostate-specific antigen (PSA) tests, digital rectal exams, biopsies, and/or imaging. This approach allows those with low-risk tumors to avoid possible side effects of more aggressive treatment. A watchful waiting plan is an even less aggressive treatment plan, often foregoing the use of frequent tests and generally used by those who do not want, or cannot have, more aggressive treatments such as instances where the patient has other life-threatening medical conditions.

For those under active surveillance or who have not received a PCa diagnosis, PSA tests are often used to monitor prostate health, but the use of PSA screening has still led to overdiagnosis and over-treatment leading to negative side effects and a drop in the quality of life of patients[9,10]. In light of this, global organizations offer variable guidelines regarding recommendations for PSA screening. For example, the United States Preventive Services Task Force suggests that screening in men aged 55 to 69 should be decided individually and suggests that routine PSA screening in men over 70 is not recommended[11]. Meanwhile the European Association of Urology recommends PSA screening only for those with a life expectancy of at least 10-15 years[12] and the Canadian Task Force on Preventive Health Care offers a weak recommendation against PSA screening, citing small and uncertain health benefits[13].

Although PSA screening may have benefits, formal diagnosis of PCa typically requires pathological analysis of a prostate biopsy. When evaluating the biopsy, samples are typically assessed using Gleason grade scores, which refers to the observed morphology of tumor cells. Cells exhibiting more aggressive characteristics, such as irregular gland formation and a lack of differentiation, receive scores >= 4, while less aggressive cells with normal morphology receive lower scores[14]. Conducting such biopsies is invasive, necessitating surgical procedures, which is not ideal for those who may be under active surveillance for years.

Various genomics-based methods have been developed to enhance this approach, but they typically rely on gene expression profiles obtained from solid tumor biopsies[15–17]. Introducing a non-invasive, non-PSA-based, molecular strategy for identifying aggressive tumors would offer two significant advantages: it would be less intrusive for patients and would facilitate repeated sampling for continuous monitoring. Similarly, the low recurrence of somatic SNVs and the delayed appearance of SVs in localized PCa means that any non-invasive molecular biomarker at this disease stage likely needs to extend beyond the detection of somatic nucleotide variants[18].

DNA methylation, which correlates with gene expression and occurs broadly throughout the genome, holds promise as a potential source of clinically useful biomarkers. Detectable non-invasively, DNA methylation may serve as a value metric for distinguishing between indolent and aggressive PCa[19,20].

Metastatic Prostate Cancer

Approximately 5% of patients are diagnosed with metastatic disease, many of whom will develop recurrence after treatment[21]. Those that progress to the most lethal form of the disease, mCRPC, have a median survival of only 5.5 months[3]. Fortunately, an increasingly large number of treatment options exist, but it is critical that patients are quickly matched to the most appropriate treatment.

In metastatic prostate cancer (mPCa), the primary treatment course includes androgen deprivation therapy (ADT), also known as hormone therapy. However, almost all patients eventually progress to mCRPC and receive AR-directed drugs, such as enzalutamide or abiraterone, or taxane-based chemotherapy[22,23]. Despite the expanding array of effective treatments for mPCa, there remains a notable absence of molecular biomarkers to assist in quickly pairing patients with their optimal treatment plans[23]. Recent advances in liquid biopsy techniques present a promising avenue to address this deficiency[24]. For example, the clinically validated circulating tumor cell (CTC)-based assay, which detects abnormal AR splice variants resulting in the truncation of the AR protein ligand-binding domain targeted by enzalutamide and abiraterone, illustrates the potential for non-invasive disease monitoring[25,26]. However, this established assay yields false-negative results in 80-90% of cases and exhibits 3% sensitivity when utilized prior to first-line treatment[3,25]. Consequently, there exists a clinical need for improved strategies for the non-invasive monitoring of mCRPC.

1.2.2 Common somatic variants

Single nucleotide variants

Patients diagnosed with indolent PCa typically present with low tumor mutation burdens, with a median of 0.5 missense and/or nonsense mutations per megabase (Mb)[27]. Notably, some tumors display no detectable exonic SNVs and the most recurrently modified gene (*SPOP*) undergoes mutation in only 8% of patients[27]. Mutation rates increase as patients progress to mCRPC, with an average tumor mutation burden of 2.7 SNVs per Mb and 10 oncogenes (*TP53, AR, FOXA1, SPOP, PTEN, ZMYM3, CDK12, ZFP36L2, PIK3CA* and *APC*) enriched for mutations when compared to indolent disease[28]. Other studies have pointed to mutations found in *BRCA1/*2 and *HOXB13* as high-risk factors[29,30]. The limited occurrence of recurrent SNVs and low tumor mutation burden, particularly in cases of indolent disease, make the use of SNVs as prognostic biomarkers difficult in PCa.

Copy number alterations and structural variants

In contrast to SNVs, as many as 70-87% of mCRPC patients harbor recurrent SVs[18]. Others have reported an average of 230 SVs per mCPRC patient[31]. While a wide variety of SVs and gene targets have been reported in PCa, some events are particularly prevalent. For example, the detection of *TMPRSS2::ERG* gene fusions has been proposed as a ‘gold standard biomarker for diagnosis’ in PCa due to its high prevalence early in tumor evolution, with a large variety of other fusions commonly observed as well[31,32]. Similarly, overexpression of the AR gene has long been known to drive castration resistance, with recent studies showing that this is a result not only of a previously described genomic amplification of the gene, but also due to increased copy numbers of an upstream enhancer[33–35]. Indeed, amplification of the enhancer alone (in the absence of gene amplification) has been reported[35].

1.2.3 Methylation landscape

Methylation is known to be associated with gene expression and to play a role in cancer progression[20,36–38]. In PCa, changes to methylation patterns have been previously documented to some extent, with hypermethylation and associated decreased expression of tumor suppressors such as *GSTP1* occurring very early in tumor formation, providing evidence that changes begin early in oncogenesis[39]. Others have shown that differential methylation of exon 3 of the oncogene *MYC* correlates with Gleason grade 3 and grade 4 in some cases, providing evidence that correlations between methylation and Gleason grade exist[40]. Taken together, these findings suggest that methylation profiling of Gleason grade 3 and 4 tumors may identify early changes associated with the transition from indolent to aggressive disease.

1.2.4 Liquid biopsies and prognostic assays

As opposed to invasive solid tumor biopsies, acquisition of liquid biopsies simply requires a sample of bodily fluid from the patient, often in the form of blood or urine[24]. Once collected, a wide variety of strategies can be applied for detection of different types of biomarkers including CTC analysis[26], ctDNA quantification[41], analysis of ctDNA/ctRNA for detection of genomic variants[22], ctDNA methylation profiling [19], and expression of mRNA-[15], lncRNA-[42] and circRNA-based[43,44] biomarkers. Detection of biomarkers using these various approaches have been used with success in a variety of cancer types, including PCa[15,22,26], non-small cell lung cancer[45], colorectal cancer[46] and pan-cancer studies[47]. The quickly growing list of successful applications of liquid biopsies is indicative of the clinical utility of this approach. However, the relatively low amount of material in patient plasma that originates from tumors can make the confident identification of some biomarkers difficult, so 1) standardized analytical strategies and 2) biomarkers based on molecules unlikely to quickly degrade are critical for further progress in this area.

Regarding PCa specifically, a number of liquid-based assays have been designed with variable amounts of success. As mentioned in section 1.2.1, a current clinically approved assay based on CTCs works by detecting AR splice variants that remove the protein domain targeted by common therapies[26]. This assay for identifying refractory mCRPC gives 80-90% of patients a false negative and has a sensitivity of only ~3% when used prior to first-line treatment[3,26]. Using a different approach, the Maher Lab, in collaboration with Drs. Chaudhuri and Pachynski, developed a targeted panel for detecting variants in cfDNA (EnhanceAR-Seq), which achieved a 100% positive predictive value for identifying resistance to AR-directed therapies and showed a significant association with patient survival (n=40)[22]. In addition to monitoring the AR gene and other oncogenes, the assay leveraged recent findings regarding the recurrent focal amplification of the AR enhancer[18,34,35]. Others have used an integrative approach based on both cfDNA- and CTC-monitoring of AR variants to monitor castration-resistant patients, while a different assay relied primarily on the abundance of ctDNA to evaluate castration-sensitive PCa[48,49].

1.3 Structural variation

Structural variants (SVs) are large-scale alterations to genomic sequences, typically defined as being >1kb in size, although no precise limit on the required size exists[50,51]. In addition to inherited or germline alterations, the somatic acquisition of SVs have been associated with various conditions and are particularly prevalent in specific cancer types, such as PCa[35,52]. A better understanding of these SVs and improved methods for detecting them may lead to improved treatment options.

1.3.1 Classes of structural variants

SVs can take a number of forms, each with different potential biological implications depending on the genomic context of the variant[51]. For example, translocations occur when a part of one chromosome becomes attached, or is substituted with, a different chromosome. Translocations may be considered balanced or imbalanced, based on the net gain or loss of genetic material. Other SV classes include inversions when the orientation of a region is flipped. Such events are balanced but can affect how the region interacts with upstream and downstream sequences. Largescale insertions and deletions, like small indels, may introduce or remove genetic sequences and alter the spatial configurations of a region. Many of these classes may result in a gene fusion, an event where two disparate genes become juxtaposed with one another resulting in a novel gene.[53]

1.3.2 Detection of structural variants

The study of SVs has traditionally lagged behind the study of smaller mutations in part due to the technical difficulties caused by their scale and complexity[50]. Most current methods for SV detection are based on the computational detection and interpretation of sequence alignment anomalies when aligning short, paired-end reads[54–56]. For example, mate pairs that align further or closer to each other than expected may indicate an insertion or deletion[57]. Alternatively, changes in read depth in different regions may suggest the gain or loss of a genomic region[58,59].

In most clinical situations, it is desirable to use a targeted sequencing panel to reduce costs and allow for increased coverage of clinically relevant regions. This presents a challenge for SV detection, as it is possible that a portion of an SV may not be covered by the panel, making it difficult to make precise calls. In the case of the EnhanceAR-Seq panel designed by our lab in collaboration with Drs. Chaudhuri and Pachynski for mCRPC patients, two particular SVs were of significant interest[22]. The first region of interest was the *AR/enhancer* locus. In order to capture the high variability present in potential SVs in this region, probes located throughout the *AR* gene as well as probes interspersed at regular intervals throughout the non-coding region of interest were used. The second region of interest included potential *TMPRSS2::ERG* gene fusions. Fusions between these genes are known to have a select number of common breakpoints, so no intergenic probes were required. Other possible approaches, though not employed by EnhancerAR-Seq, can include the use of probes that contain sequences that would only exist if a specific SV junction were formed.

When detecting SVs in cfDNA using a targeted panel, additional considerations must be made. For example, cfDNA is known to have non-random degradation based on nucleosome positioning and is often found in short fragments[60]. Additionally, ctDNA typically represents a very small fraction of all cfDNA, so highly sensitive methods that do not introduce poor specificity are required. Prior to our work, we identified three SV tools benchmarked for SV detection in cfDNA in their original publications, namely SVICT, Factera and Aperture[61–63]. While usable, each had specific, potential limitations that were thought to hinder their clinical adaptation. By default, SVICT targets SVs <2kb in size, smaller than many clinically relevant events[61,64]. Factera specializes in gene fusions but is not optimized for other SV classes. Aperture uses an alignment-free approach to overcome potential alignment issues but has the negative side effect of being unable to classify the types of SVs that are detected. Importantly, none of the tools allow for a matched control, which is usually standard for clinical applications and for minimizing noise when low amounts of ctDNA are expected.

1.4 Circular RNA

It has been well established that standard processing of RNA typically includes splicing mechanisms to remove introns and to allow for the creation of different isoforms (with different sets of exons)[65]. More recently, it has been shown that some RNAs undergo backsplicing, which is a specific type of splicing wherein a downstream region of a transcript becomes covalently bound to a region that would normally precede it, resulting in a circularized RNA[66].

1.4.1 Structure and biogenesis

Like linear RNA, circRNA consists of covalently bound nucleotides transcribed from a gene[66]. As a result of their circular structure, circRNAs cannot be modified to include a 5' cap or a poly-A tail, because they do not have a 5' or 3' end. Although this may prevent the circRNA from being transported to the cytoplasm using the same mechanisms employed by linear mRNA, it does provide the circRNA with protection from exonucleases[43]. This inherent structural protection from degradation is thought to increase the half-life of circRNAs.

The backsplicing events that result in these circular structures is thought to primarily be a function of standard spliceosome machinery, although trans-acting factors are also involved[43,67]. This is also usually facilitated by complimentary sequences located in the regions flanking the donor and acceptor splice sites, typically in the form of Alu-repeats[66].

1.4.2 Biological function

A variety of functions have been observed in circRNAs, suggesting that they are not sequencing artifacts. Such functions include mechanisms such as direct transcription regulation[68] and indirect regulatory functions facilitated by RNA-binding proteins and microRNAs[69–71]. They have also been shown to encode peptides, including novel peptides which would not have been possible by linear isoforms of the same gene[72].

1.4.3 Fusion-derived circular RNA

A subset of circular RNAs include those isoforms which consist of genetic sequences that come from two separate genes[73]. Such transcripts are referred to as fusion-derived circRNAs (fcircRNAs) and may be the result of a gene fusion or of a read-through transcript (although some refer to these as read-through circRNAs, or rtCircRNAs)[74]. For clarity and brevity, this work will refer to any circRNA composed of genetic sequences from two different genes, regardless of the proximity of the genes, as an fcircRNA.

Oncogenic Potential

Although little is known about fcircRNAs, recent studies suggest that may be functional and, in some cases, oncogenic. Indeed, fcircRNAs arising from *EML4*::*ALK* fusions were shown to promote cell migration and invasion in non-small cell lung cancer[45,75], while multiple different fcircRNAs have been suggested to have oncogenic functionality in leukemia[76–78]. Importantly, 62 fcircRNAs were recently reported in a cohort of PCa patients with localized disease, although no information was given about which fcircRNAs they were[79]. Despite this limited understanding about their functionality, the expected stability of circular transcripts and somatic nature of most clinically relevant gene fusions suggests that fcircRNAs could potentially be leveraged as cancer biomarkers[43–45].

Detection Methods

The study of fcircRNAs has been severely limited, in part because of two specific obstacles: 1) most RNA-Seq studies use Poly(A)-selection to enrich for coding transcripts which systematically removes circRNAs and 2) software limitations. Indeed, most studies that have investigated the existence and functionality of fcircRNAs began by identifying a gene fusion of interest and then performing targeted sequencing or other highly specific methods (such as Sanger sequencing of PCR products) to interrogate potential backsplices in the fusion of interest[45,76–78]. We are aware of only three previously published software tools with fcircRNA detection functionality. The first, Acfs, has systematic biases by requiring that fcircRNAs be composed of genes from different chromosomes or from different strands of the same chromosome[80]. The second, Fcirc, uses a built-in aligner to map reads to a user-supplied list of gene fusions, preventing the detection of novel events[81]. Finally, CircFusion, uses a similar workflow as Fcirc but relies on alignments from STAR rather than using its own aligner and again requires *a priori* knowledge of gene fusions[82]. The fact that the example fusion list provided by Fcirc is twice as long as the one provided by CircFusion highlights the limitations that this approach entails, as results will be directly influenced by the list used. No available tool allows for genome-wide, unbiased detection of fcircRNAs.

Chapter 2: Novel pipeline for analysis of circulating tumor DNA

Preface

This chapter has been adapted from the following publication:

Jace Webster, Ha X. Dang, Pradeep S. Chauhan, Wenjia Feng, Alex Shiang, Peter K. Harris, Russell K. Pachynski, Aadel A. Chaudhuri, Christopher A. Maher. PACT: A pipeline for analysis of circulating tumor DNA. *Bioinformatics*. 2023.

2.1 Introduction

Identification of genomic variants in ctDNA has emerged as a promising method for non-invasive monitoring of cancer progression and treatment response. This non-invasive monitoring is particularly beneficial in metastatic disease as ctDNA originating from both primary and secondary tumor sites can be found within a single blood sample. Despite low ctDNA abundance and expected allele frequencies, deep targeted sequencing has been successfully used to improve sensitivity for detecting single nucleotide variants (SNVs)[83]. Structural variants (SVs) are a major class of genomic drivers of cancer progression but their use in non-invasive applications remains limited due to the challenges of accurately detecting the wide variety of possible complex genomic rearrangements[22].

When attempting to overcome the limitations of current ctDNA SV callers, while also identifying copy number alterations (CNAs) and small mutations, users often resort to *ad hoc* or proprietary approaches. This leads to time-consuming analyses and inhibits reproducibility in the field, in part because of the variety of possible customizable tools and parameters required. A few tools have been developed for SV detection in ctDNA, however, each of the identified tools has crucial limitations inhibiting their ability to identify clinically relevant events. For example, none of them accept matched germline control data, which is critical for differentiating artifacts, germline and somatic events.

To address these limitations and promote accurate and reproducible detection of SVs, CNAs, and small mutations in ctDNA, we developed an open-source unified **P**ipeline for the **A**nalysis of **ct**DNA (PACT).

2.2 Results

Although SNV and CNA strategies for ctDNA are usually considered robust, SV detection has traditionally been challenging. While PACT integrates callers of various variant types, its major focus is on improving SV analysis in ctDNA (Figure 2-S1 and 2-S2). We benchmarked PACT using (i) patient data, (ii) an *in silico* simulation, and (iii) an *in vitro* dilution experiment, while comparing against other ctDNA SV callers including SViCT, Factera, and Aperture[61–63].

2.2.1 Analysis of published clinical and reference datasets

First, we applied each tool to ctDNA samples from a published cohort of 40 prostate cancer patients and found that only PACT and Aperture detected all published SVs (Figure 2-S3)[22]. Precision was not assessed due to the lack of gold standard positive controls. Similarly, only PACT and Aperture identified all expected SVs in a public cfDNA reference dataset (SRA: SRR8551545). However, Aperture reported 1,636 unvalidated SVs in the reference data (1,623 more than the next highest tool, Factera), suggesting poor precision (Figure 2-S4). We also found a high accuracy rate when applying the SNV and CNA portions of PACT to the reference dataset (Table 2-S1).

2.2.2 *In silico* simulation

Second, we performed an *in silico* simulation with tumor data from 4 prostate and 5 colorectal cancer patients (Table 2-S2)[22,84]. Sequencing reads from tumor and respective matched controls were combined to simulate ctDNA content ranging from 0.1-30%, bounded by the tumor purity of original samples. At each dilution, PACT achieved the highest sensitivity (Figure 2-1). Specificity was not assessed as validation of novel calls could not be performed, however, we observed that Aperture and Factera consistently had the most candidates (i.e., approximately 13x and 8x more than PACT at 7.5% tumor DNA content; Figure 2-S5), suggesting potentially poor precision.

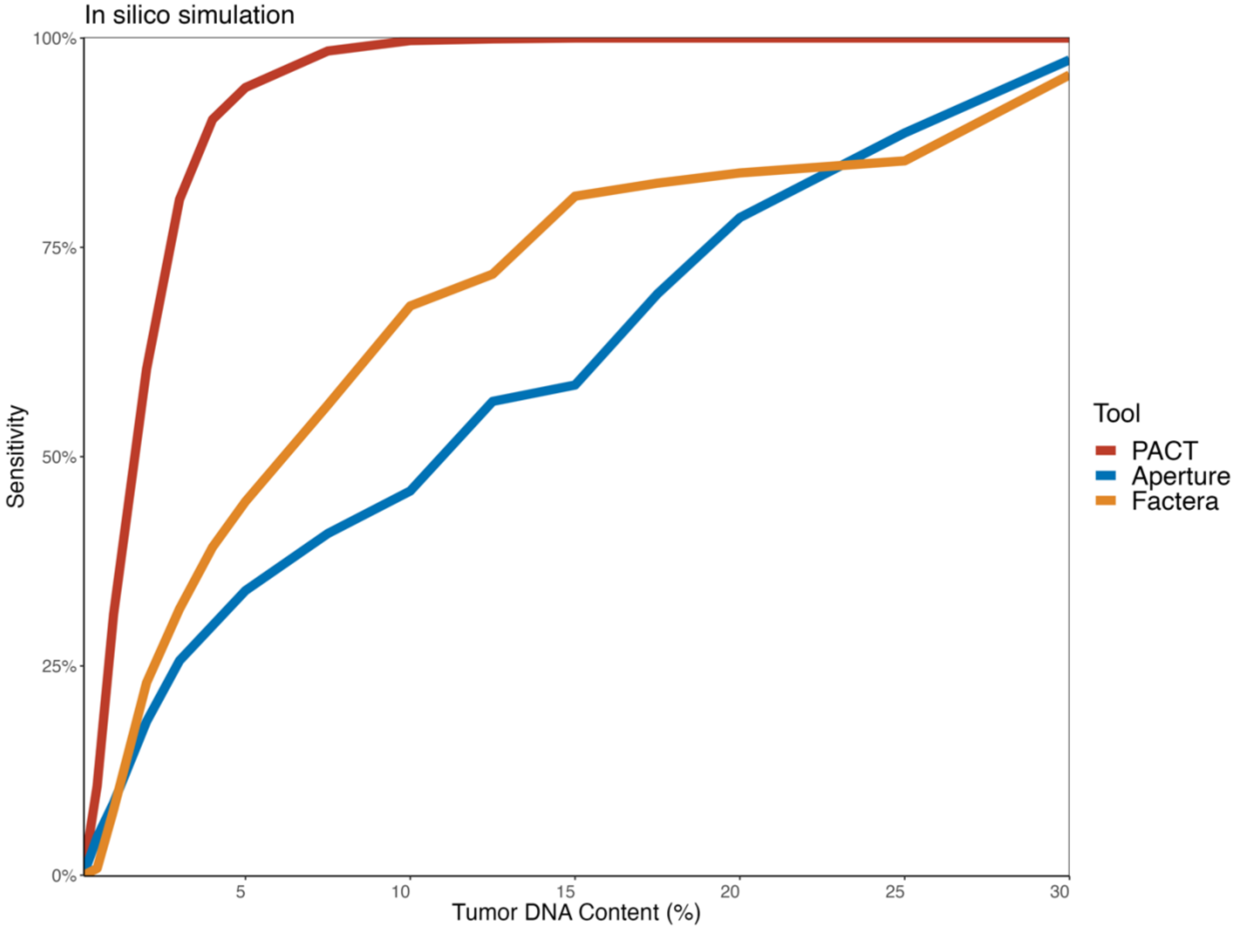


Figure 2-1. *In silico* benchmarking results. Observed sensitivity after *in silico* simulation of SVs across different tumor DNA content levels.

2.2.3 *In vitro* dilution experiment

Our third evaluation was performed using an *in vitro* dilution experiment of the well characterized breast cancer cell line (HCC1395) and its matched control cell line (HCC1395BL). We mixed these cells to created diluted samples with 0.1-100% tumor content. Targeted sequencing of 26 validated SVs was then performed. We found that PACT achieved the highest sensitivity and F1 accuracy scores and was the only tool to achieve >90% sensitivity at all dilutions >3% (Figures 2-2). At the lowest detectable level (0.12% tumor content), PACT, Aperture, and Factera achieved 15%, 10%, and 0% sensitivity, respectively.

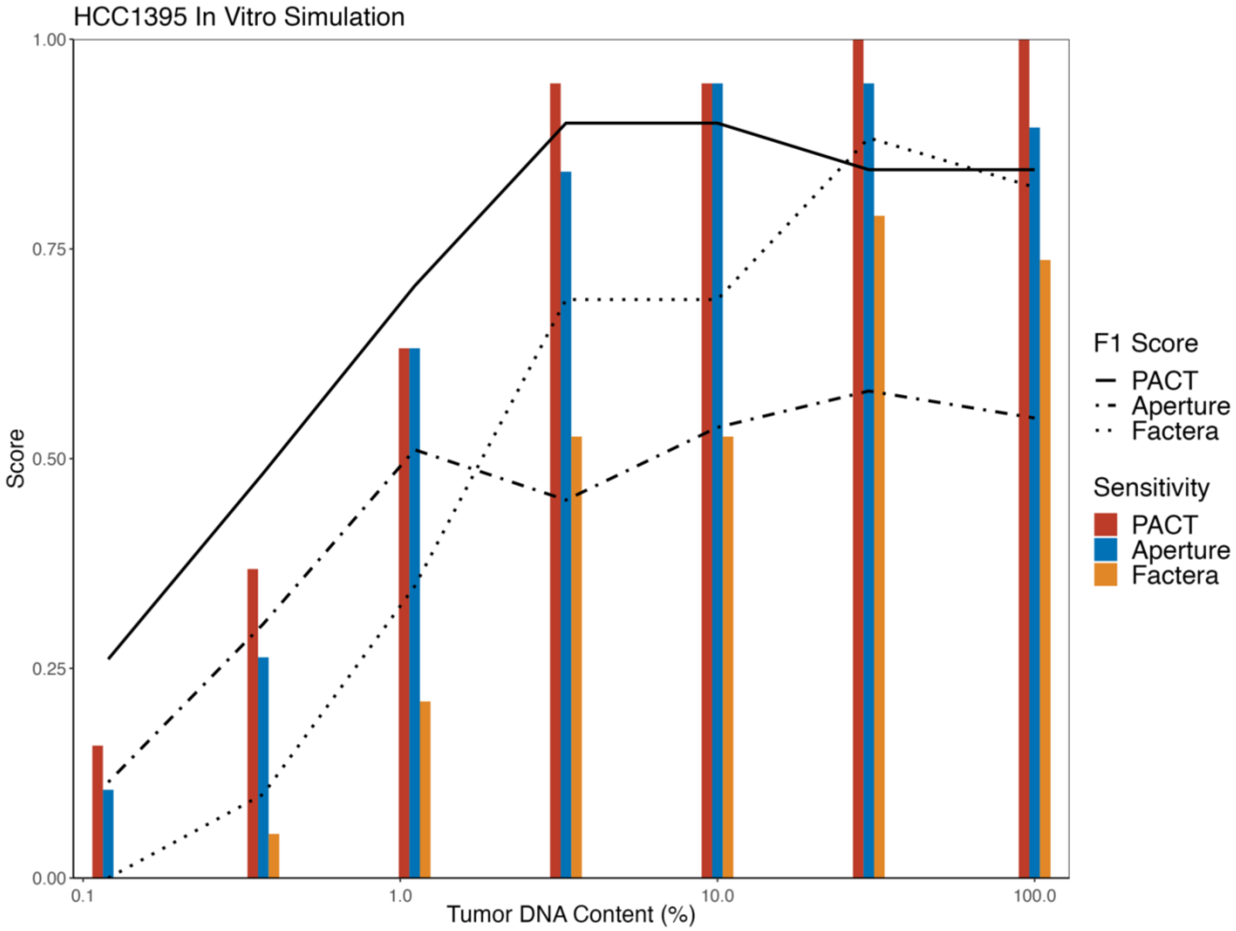


Figure 2-2. *In vitro* benchmarking results. Sensitivity and F1 accuracy scores achieved during an analysis of *in vitro* simulation data.

2.3 Discussion

Together, these results suggest PACT is both more sensitive and more precise than other ctDNA SV callers. By including SNV and CNA workflows within PACT and distributing the pipeline in a standardized workflow language (CWL), PACT is well suited for improving accuracy and reproducibility in ctDNA analysis, with potential clinical applications.

2.4 Methods

PACT is a standardized ctDNA pipeline for detection of SVs, CNAs, and small mutations. It is designed for reproducibility in high-performance computing environments capable of processing large numbers of samples and can be run by popular workflow management systems. PACT consists of methods for detection of small mutations, CNAs, and SVs independently. Briefly, each variant calling strategy begins with the creation of an initial list of candidates nominated using an ensemble of tools for variant calling. Where possible, each tool is run using relaxed filtering criteria to increase sensitivity. To obtain high specificity, normalization and/or filtering strategies are applied to all nominated variants based on expected noise in ctDNA caused in part by deep sequencing and low allele frequencies (often <1%)[83]. All workflows accept sequencing data from matched controls and from a panel of unmatched, healthy individuals to aid in removing non-somatic events. Each of these individual workflows is described in greater detail below.

2.4.1 SV Workflow

The PACT SV workflow can be broadly divided into two steps: 1) Creation of a broad list of SV candidates using relaxed filtering criteria and 2) filtering of candidates to reduce ctDNA-related noise. The first step in PACT beings by making initial somatic SV calls using Delly, Lumpy, and Manta in sensitive mode[54–56]. All three callers are commonly used in studies of SVs, although none of them were specifically designed to be used on cell-free DNA (cfDNA). However, we adapted these tools for cfDNA by using relaxed parameters that allow sensitive reporting of SVs with low levels of read support. Delly calls are generated using default settings and the *`delly call`* command, however this command is not followed by the *`delly filter`* command normally recommended in the tool's documentation, so that low-frequency SVs are not inadvertently removed at this stage. Lumpy calls are made using the *`lumpyexpress`* command, with the default minimum weight lowered to 3 by using the *`-m`* parameter. Manta is run using default settings, however only the *candidateSV.vcf* output file is used for downstream analysis, rather than relying on Manta's built-in filtering that gets applied to the final *somaticSV.vcf* and *diploidSV.vcf* output files.

Consensus initial SV calls are then identified by merging initial SV candidates using SURVIVOR[85]. In our analyses we modified SURVIVOR's default settings based on our experience applying SURVIVOR to clinically relevant SVs (*max-distance-to-merge=100, minimum-sv-size=200, same-strand=false, estimate-sv-distance=false*, all of which can be modified by the user of PACT). Initial consensus SVs represent highly sensitive collection of SVs supported by multiple callers. The vcf file containing consensus calls is then modified with a custom script to ensure compatibility with downstream tools.

To achieve a high level of specificity, PACT performs various filters to remove initial consensus SV calls that are likely false positives. First, targeted region-based filtering is performed to retain SV calls that originate from regions targeted by the targeted panel. A +/-

200bp wingspan is automatically added to targeted regions to ensure that SVs with breakpoints that are located immediately adjacent to targeted regions are also retained.

Additional region-based filtering is performed to remove consensus SV calls that originate from genomic regions that are difficult to align to and tend to have high false positive rates. This is done by 1) immediately removing any consensus SV with a breakpoint that falls in a "blacklisted" region and 2) removing consensus SVs that have >1 breakpoint that falls in a low-complexity genomic region. In our benchmarking, blacklisted regions were based on the blacklist bed file provided by 10x Genomics at http://cf.10xgenomics.com/supp/genome/hg19/sv_blacklist.bed and low-complexity regions were taken from https://github.com/lh3/varcmp/raw/master/scripts/LCR-hs37f5.bed.gz. We found that these inputs worked well in our benchmarking, but users can supply alternative regions using the "*neither_region*" and "*notboth_region*" parameters.

Next, PACT performs normal filtering to eliminate SV calls with evidence suggesting that they are likely germline events or systematic (sequencing/alignment) artifacts. To do this, consensus SV calls from individual cfDNA samples across a patient cohort are merged and re-genotyped across matched controls and cfDNA samples from healthy individuals. The "*svtools sort*" command is used for sorting and "*svtools lmerge"* command is used for merging SV calls across samples to generate cohort-wide SV calls[86]. Cohort-wide SV calls were then subsequently genotyped across all samples and matched controls using SVTyper[87]. Additionally, the user-supplied panel of healthy normals (sequenced with the same targeted panel) is genotyped using the cohort-wide vcf. The panel of healthy normal is expected to contain cfDNA sequencing data from healthy individuals, if available. In the case where a true panel of healthy normals are not available, the healthy normal panel may be substituted with a panel comprised of all available matched control samples (though they are not truly “healthy normal” samples). Candidate SVs are then filtered based on genotyping results to remove those with supporting reads found in the panel of normals or in each sample’s respective matched control. Additionally, PACT retains only consensus SV calls with multiple types of read support in cfDNA samples (at least 1 supporting split-read and 1 supporting discordant paired-end read), requiring at least >2 total supporting reads by default (customizable by the user). If either breakpoint from an SV overlaps with an optionally supplied “whitelist” region bed file, the requirement for two forms of evidence (split-read and discordant paired-end read) is waived, but the minimum read support threshold must still be met.

Finally, to help users interpret the SVs, PACT performs SV annotation using *snpEff*[88]. Final output also includes additional useful information, including which of the SV callers (Delly, Lumpy and/or Manta) originally reported the SV, whether the event corresponds to the optionally provided whitelist, and the number of supporting split-read and discordant read-pairs that were found. The final result of the SV workflow is a highly confident list of annotated SV calls presented in a standard bedpe format, designed to be compatible with downstream analysis tools for easy interpretation.

2.4.2 SNV/Indel Workflow

SNV calling begins by generating candidate somatic calls using Mutect, Strelka, VarScan, and Pindel[89–92]. Each variant caller accepts a variety of unique parameters. Where possible, the PACT workflow allows full customization of input values for parameters that are passed to each tool but attempts to provide reasonable defaults (documented at https://github.com/ChrisMaherLab/PACT) based on observed metrics in ctDNA where possible. Additionally, PACT accepts a list of whitelisted variants in VCF format which are genotyped using GATK's *HaplotypeCaller*[93]. In our testing, we used the whitelist VCF that can be downloaded from the DoCM database[94]. All candidate calls are then combined and then decomposed using Vt's *decompose* function[95].

The decomposed VCF is next annotated using *vep* and read counts are standardized by re-calculating the read depth for each SNV in both the tumor and matched control sample using the *bam-readcount* tool[96,97]. This allows for a standardized read depth measurement, rather than relying on reported read counts determined by individual callers. Filtering is then performed and is based on SNV frequeincy in gnomAD (https://gnomad.broadinstitute.org), mapping quality, read depth, and allele frequency. All thresholds include default values based on ctDNA quality control metrics observed by our group and can be found on the project's GitHub page. Finally, to address the large number of likely false positives that occur as a result of high sequencing depths combined with low expected variant allele frequencies, background error suppression is performed by genotyping candidate SNVs/Indels in the user-provided panel of cfDNA samples from healthy individuals using GATK's *HaplotypeCaller*[93]. If a panel of healthy individuals is unavailable, the user may instead provide a panel composed of all available matched controls for this step. Genotyping results are then supplied to GATK's *VariantFiltration* method for filtering out any call that has read support in more than a specified percentage of samples (default: 10% of the panel of normals)[93]. All resulting calls are converted to table format using GATK's *VariantsToTable* command with parameters that can be customized by the user.

Aspects of PACT's SNV/Indel workflow are comparable to the McDonnell Genome Institute's public analysis workflow which has also been distributed in CWL format (https://github.com/genome/analysis-workflows).

2.4.3 CNA Workflow

For CNA analysis, a read depth-based method is first employed to calculate the log-transformed ratio of depth between the patient cfDNA sample and the control panel based on cfDNA samples from healthy individuals or matched control normal samples, corrected for biased in GC content and repeat content. If a panel of healthy individuals is unavailable, the user may instead provide a panel composed of all available matched controls. We employ CNVKit for read depth ratio calculation, control panel construction and bias correction[58]. Next, log ratio of depth is recentralized using the CN control regions (chosen as the least CN altered regions via surveying existing whole genome sequencing data, if available) to account for depth bias often seen in targeted sequencing of cfDNA. Finally, regions with log ratio of depth that deviated from that of the CN controls (default: 3 standard deviations from the mean) are called as CNAs.

2.4.4 Application to published cfDNA prostate cancer cohort

Blood plasma cfDNA sequencing data from published prostate cancer patients that had been reported to contain either 1) tandem duplications of the Androgen Receptor (or its enhancer) or 2) deletions resulting in *TMPRSS2::ERG* gene fusions were selected for initial PACT testing[22]. Although matched tissue WGS was available for only one patient with a reported SV (WGS supported the SV found in cfDNA for this patient), calls were considered reliable as these two SVs are well recognized hallmarks in prostate cancer and, in this cohort, correlated with survival outcomes. Sensitivity was assessed based on the total number of previously reported SVs detected by each tool. False positive rates were not evaluated due to the lack of validation sequencing of novel calls. All tools (PACT, Factera, SViCT, Aperture) were run with default settings, with the exception of SViCT, which was run using '-M 6000000' (default: -M 2000). The -M parameter defines the max size of SVs evaluated by SViCT and was increased based on the reported size of the previously detected events.

2.4.5 Application to Horizon Discovery reference data

The Horizon Discovery reference dataset has been sequenced and is expected to contain *SLC34A2::ROS1* and *CCD6::RET* gene fusions, both at approximately 5% allele frequency (Catalog number: HD786, SRA: SRR8551544). Fastq files were downloaded and aligned by *bwa mem* with default settings and duplicate reads were marked using the Picard *MarkDuplicate* tool[98,99]. Outputs were then sorted using sambamba 0.6.8[100]. All tools were run using default settings. As no panel of healthy normals was available for use with PACT, we substituted it with a panel of matched controls. Unfortunately, SViCT repeatedly failed to run to completion, despite the standard bam format of the input files. Specifically, SViCT crashed with a "double free or corruption (!prev)" error message. With repeated testing with or without changes to the -M parameter, we repeatedly received either this error or at times an exit code 139 error (segmentation fault). Both kinds of errors are standard C++ error messages (SViCT is written in C++) and usually means that there was an error within the software itself that caused the program to mismanage its memory resources. We noted that multiple issues have been opened on the project's GitHub page relating to these and other memory related error messages. It is unclear why the tool ran successfully on the cfDNA prostate cancer cohort (suggesting the tool was installed correctly) but failed on these samples, even though these samples were compatible with all other tested tools. For these reasons, we excluded SViCT from this analysis.

Performance of the remaining tools was assessed by checking for the detection of the reported *SLC34A2::ROS1* and *CCDC6::RET* fusions and by determining the number of reported SVs that were called but were not expected to appear in this particular Horizon Discovery dataset.

2.4.6 *In silico* simulation

In silico simulation data was generated using sequencing data from solid tumor samples collected from two different cohorts. The first was the prostate cancer cohort described in Dang, Chauhan, et al., 2020. We selected 4 solid tumor prostate samples based on the criteria that *TMPRSS2::ERG* fusions had been detected 1) in the solid tumor data and 2) in cfDNA from the same patient. Although validation sequencing was not performed, the presence of this well documented fusion in multiple samples from the same patients were considered sufficient to be treated as true positives. The second cohort was a colorectal cancer cohort described in Dang, Krasnick, et al., 2020. In that study, whole genome sequencing was followed by targeted validation sequencing on a number of solid tumor samples. We selected 5 samples, which contained a total of 7 validated SVs, based on the criteria that 1) the same sample was used for both discovery and validation sequencing and 2) the SV had >50 supporting reads in the validation sequencing. All selected samples also has matched control data available.

Sequencing reads from all samples were aligned and processed as was done with the Horizon Discovery data. Reads from aligned tumor data were then systematically combined with reads from the aligned reads of each sample's respective matched control. Calculations for the number of reads used when generating samples of different tumor DNA content levels were performed based on the total number of reads in a sample (based on *samtools flagstat* output) and the previously annotated tumor purity of each tumor sample[22,84,101]. Reads were downsampled from each tumor sample using Picard DownsampleSam and then downsampled reads from each tumor were merged with their respective matched control using Picard MergeSamFiles[99]. Combined files were then labeled internally based on their expected dilution using the Picard AddOrReplaceReadGroups command and output bam files were check with samtools flagstat to confirm that they contained the expected number of reads[99,101]. This process was repeated 100 times for each sample at each tested tumor DNA content level, using a random seed each time for the DownsampleSam command.

All simulated samples were analyzed using the default settings of each tool and sensitivity was assessed by determining the number of previously selected (and validated) SVs detected at each tumor DNA content level across all 100 iterations. False positive rates were not assessed as we were unable to perform validation sequencing on potential novel calls. We were unable to assess SViCT's performance as it failed to run to completion on any of the simulated samples, failing on each sample with the same errors observed when the tool was applied to the Horizon Discovery dataset.

2.4.7 *In vitro* validation

We further evaluated PACT performance in an *in vitro* dilution experiment of a breast cancer cell line (HCC1395). First, published genomic breakpoints from 26 validated gene fusions in the HCC1395 breast cancer cell line were used to design a targeted panel using Roche's HyperDesign tool (https://hyperdesign.com)[102]. HCC1395 cancer cells were then combined *in vitro* with the matched control (HCC1395BL) following a standard serial dilution strategy to simulate different tumor DNA content levels (0.1%-100%) and then sequenced isolated DNA samples using our targeted panel. Sequencing reads were UMI tagged and aligned using *bwa* *mem*. Reads were grouped by UMI (fgibio GroupReadsbyUmi) and consensus reads were called (fgibio CallDuplexConsensusReads) to reconstruct the most likely read representing the corresponding DNA fragment[103]. Consensus reads were then re-aligned using *bwa mem* for final alignment[98].

All samples were analyzed with each SV caller using default settings. SVs targeted by potentially poorly designed probes were removed from analysis by filtering out any SV without any coverage at the 100% tumor DNA content level (8 SVs). Similarly, samples with <0.12% tumor DNA content were removed from analysis as no tool made a correct call at that level. Again, we were unable to get SViCT to run to completion on the HCC1395 samples, despite the files being compatible with all other tools and instead received the same error messages as previously described. As no panel of healthy normals was available, we used a panel of matched controls.

For evaluation, all calls that did not match previously validated SVs were labeled as false positives (FPs) and validated calls that were not detected were labeled as false negatives (FNs). Similarly, true positives were defined as calls that matched validated SVs (TPs). Sensitivity was used to assess the proportion of true events that were detected and precision was used to assess the proportion of all calls that corresponded to true events. The F1 accuracy score is the harmonic mean of the sensitivity and precision, such that a result with perfect precision and sensitivity would have an F1 score of 1.0 and poor performance would have a score that approaches 0.

2.4.8 Resource requirements

PACT is designed to be used in a high-performance computing environment. The pipeline itself contains a variety of published bioinformatics tools and therefore its minimum computing requirements are determined by the most resource intensive tool in the pipeline (CNVkit), which is set to require 64GB of RAM and 12 cores. Minimum requirements for any given tool can be manually changed by modifying the CWL file wrapper for the tool found in the *tools* directory on the project GitHub page, although we believe we have provided sensible default requirements for all tools. We found CPU time to be highly variable and it may be influenced by many factors including (but not limited to) sequencing depth, number of samples/matched controls, number of healthy normals, and number of variants identified.

2.5 Supplementary Materials

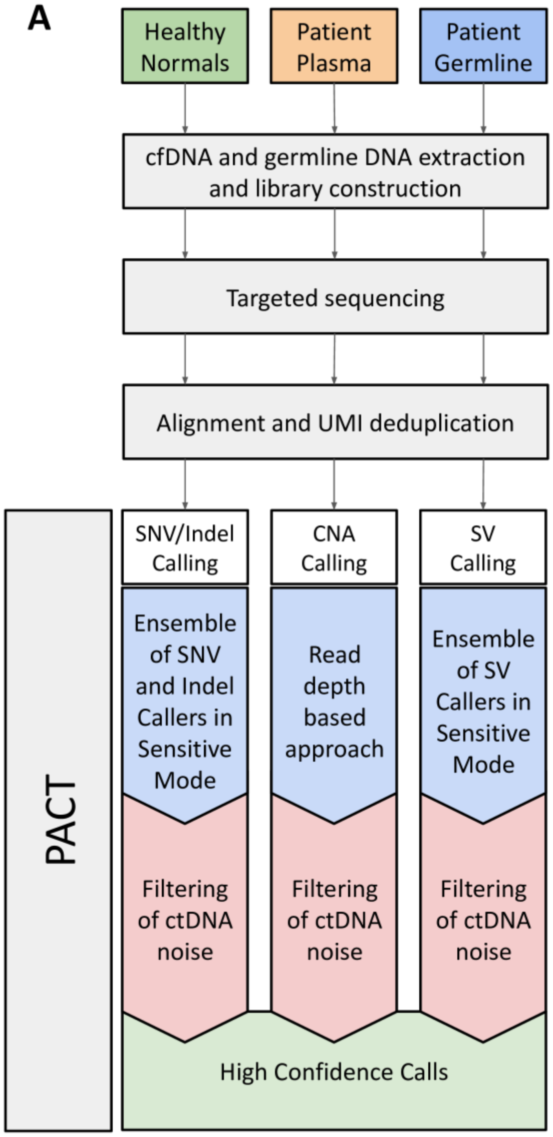

A
Healthy Normals
Patient Plasma
Patient Germline
cfDNA and germline DNA extraction and library construction
Targeted sequencing
Alignment and UMI deduplication
PACT
SNV/Indel Calling
CNA Calling
SV Calling
Ensemble of SNV and Indel Callers in Sensitive Mode
Read depth based approach
Ensemble of SV Callers in Sensitive Mode
Filtering of ctDNA noise
Filtering of ctDNA noise
Filtering of ctDNA noise
High Confidence Calls

Figure 2-S1. Overview of PACT workflow. Unmatched, health normal samples are compared against patient plasma and patient germline data to perform SNV, CNA and SV calling.

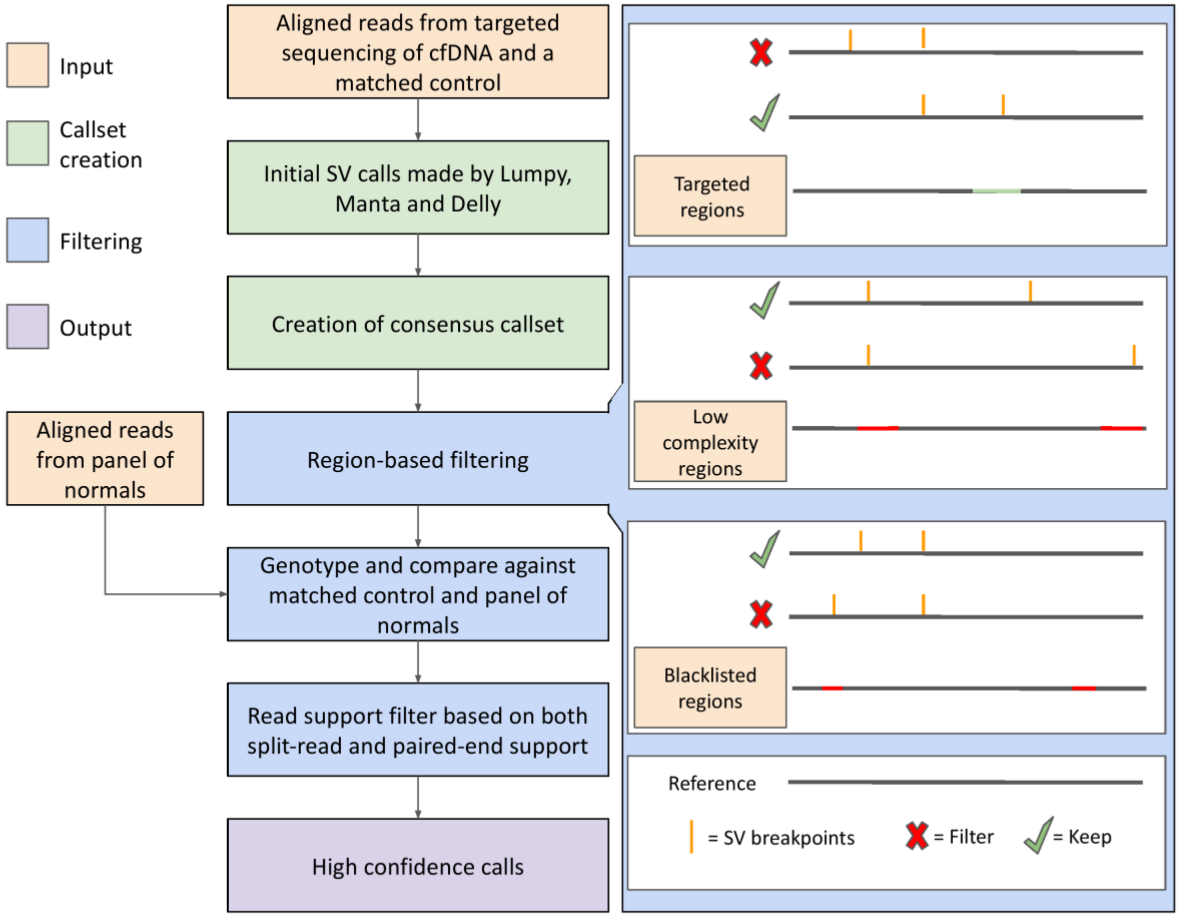

Input
Callset creation
Filtering
Output
Aligned reads from targeted sequencing of cfDNA and a matched control
Initial SV calls made by Lumpy, Manta and Delly
Creation of consensus callset
Region-based filtering
Aligned reads from panel of normals
Genotype and compare against matched control and panel of normals
Read support filter based on both split-read and paired-end support
High confidence calls
Targeted regions
Low complexity regions
Blacklisted regions
Reference
= SV breakpoints
= Filter
= Keep

Figure 2-S2. Overview of SV-calling portion of PACT. Aligned reads from targeted sequencing of cfDNA and a matched control are analyzed by an ensemble of SV callers using sensitive settings and consensus calls are then identified. A variety of filtering steps are then applied to reduce expected cfDNA noise. Region-based filters require that at least one breakpoint corresponds to a region targeted by the sequencing panel and filters out potential sequencing errors by removing SVs with more than one breakpoint that corresponds to low complexity genomic regions and SVs with any breakpoints that originate in blacklisted regions. Remaining candidates are then genotyped in a panel of healthy unmatched individuals to further remove potential artifacts and germline events, and then read support filtering is finally applied.

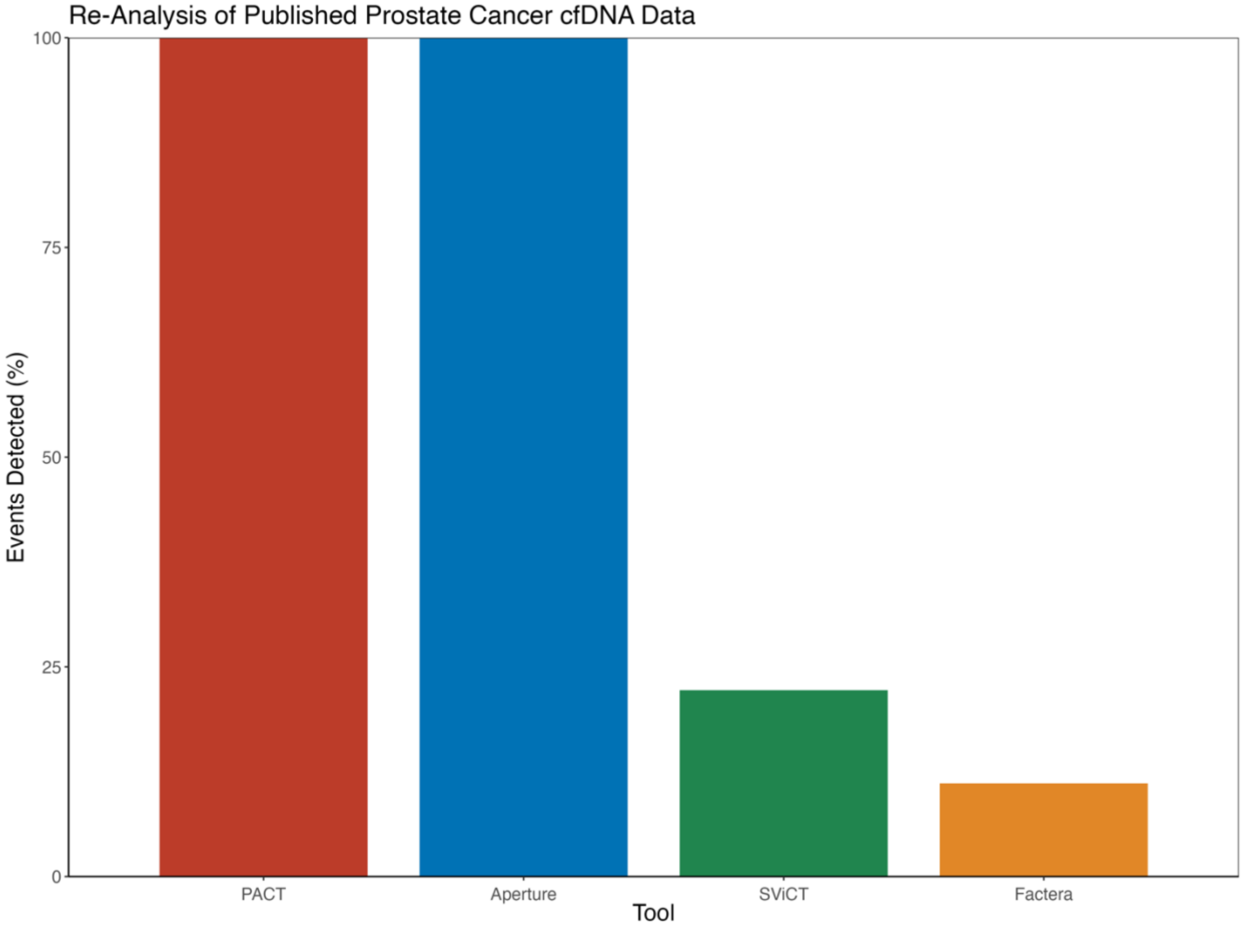


Figure 2-S3. Re-analysis of published cfDNA data. Events detected are based on 5 tandem duplications of the *AR* gene and/or its upstream enhancer and 4 deletions resulting in *TMPRSS2::ERG* gene fusions. Both events are considered hallmarks of PCa and correlated with survival in the original publication.

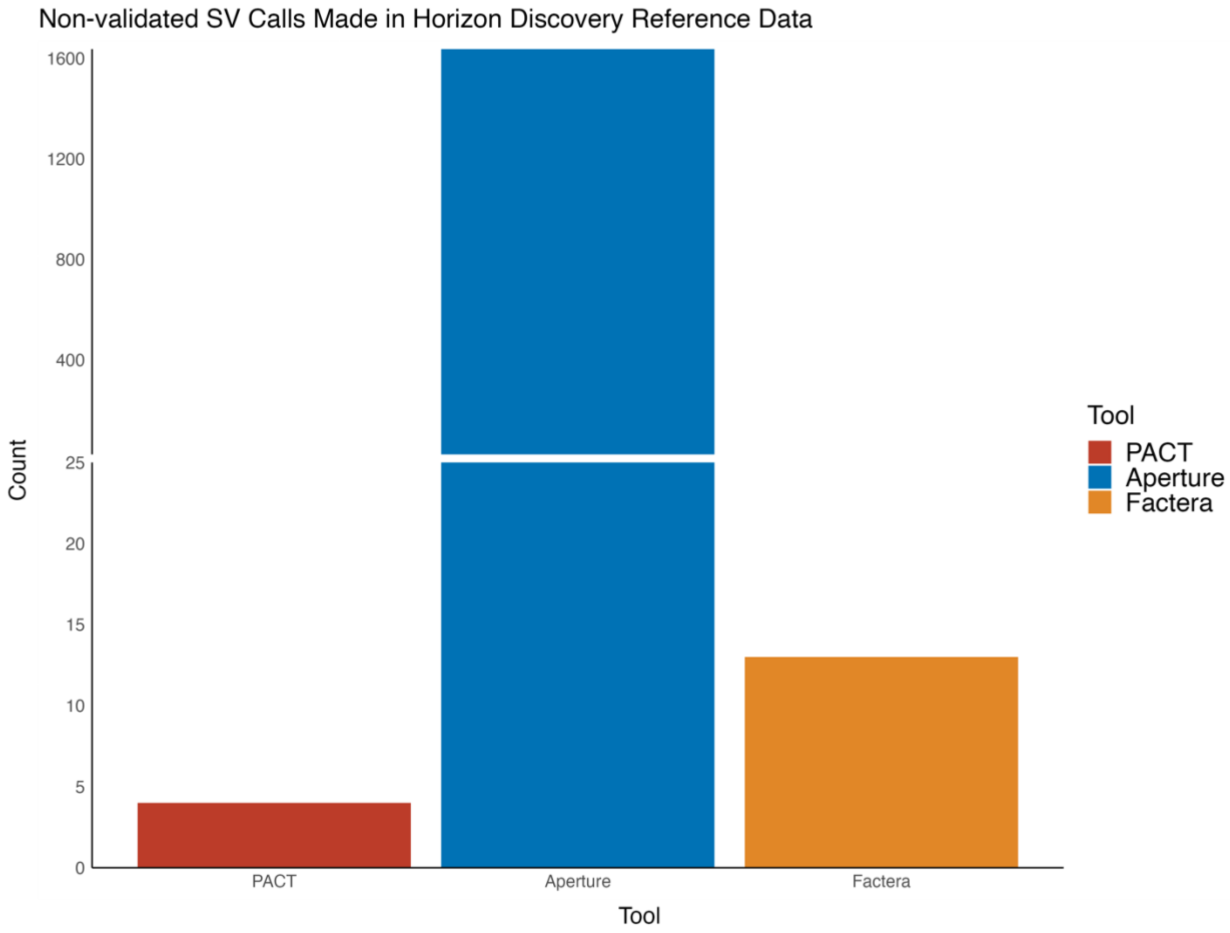


Figure 2-S4. Unreliable SV call count. Indicates the number of SVs reported that do not match previously validated SVs as reported by Horizon Discovery in their cfDNA reference dataset. SViCT results are not shown as it failed to run to completion.

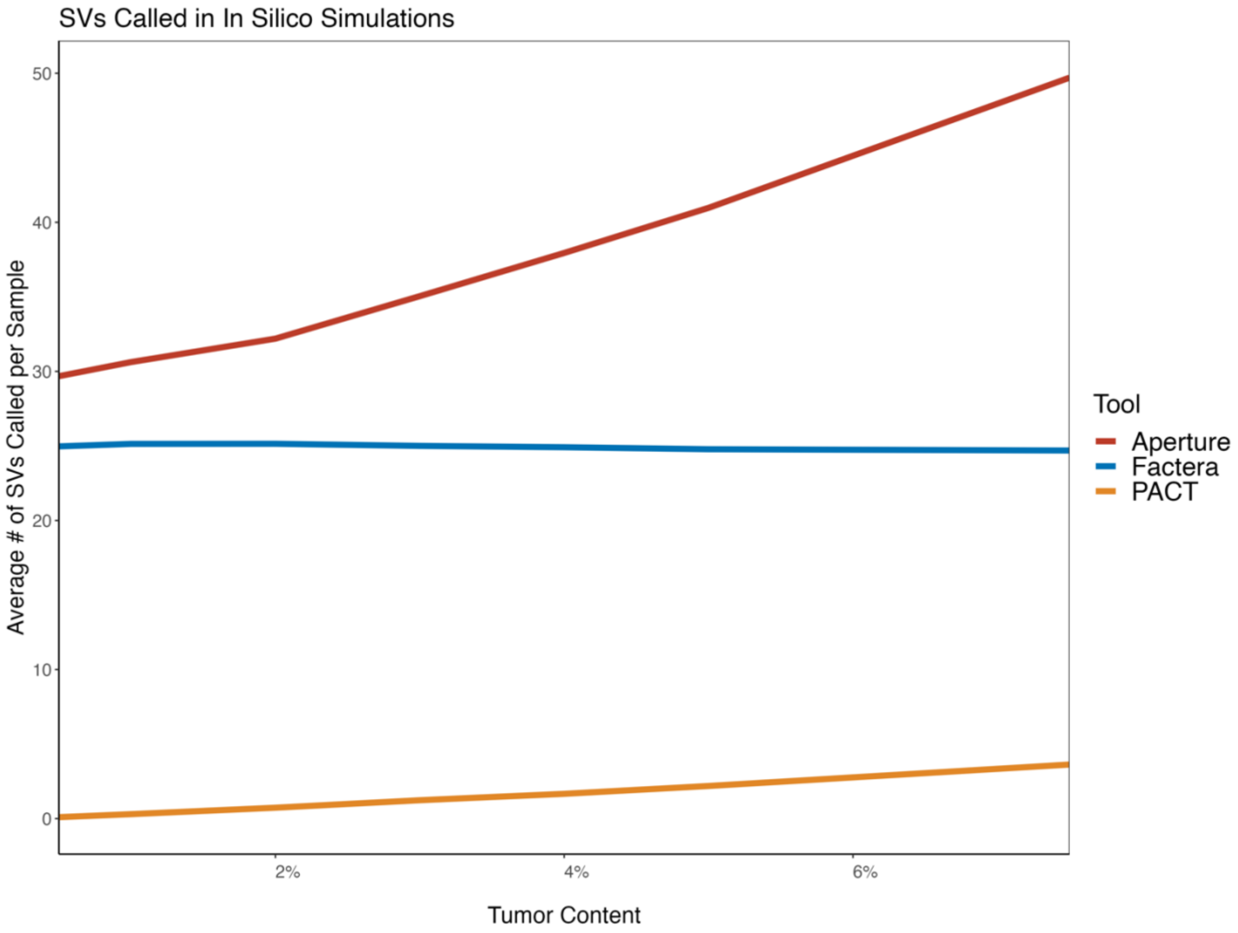


Figure 2-S5. Average number of SVs made during *in silico* analysis. Averages are based on 9 samples used across 100 iterations of the simulation. Only dilutions of <=7.5% tumor content are shown, as some samples were unable to be simulated at higher content levels due to the low tumor purity of the original samples being used as the basis for the simulation. Precision was not formally calculated due to the lack of validation sequencing of novel calls. However, PACT simultaneously achieved the highest sensitivity in these samples (Fig. 1B) and also reported the fewest total SVs, suggesting high precision.

Table 2-S1. Horizon Discovery SNVs and CNAs. All validated SNVs, INDELS and CNAs, as reported by Horizon Discovery, in the reference cfDNA dataset and their detection status based on the SNV/CNA workflows found in PACT. PACT reported an additional 21 non-synonymous variants in this dataset, all of which have been reported by Horizon Discovery in the genomic DNA that is meant to correspond to their ctDNA reference, suggesting that PACT was able to detect additional true SNVs that had been validated in the genomic DNA, but not ctDNA, version of this reference.

Gene	Mutation	Detected by PACT
GNA11	c.626A>T	Yes
AKT1	c.49G>A	Yes
PIK3CA	c.1633G>A	Yes
EGFR	c.2300_2308dup	Yes
EGFR	c.2235_2249del	Yes
MYC	Amplification	Yes
MET	Amplification	Yes

Table 2-S2. Samples used for *in silico* simulation.

Cancer of Origin	Structural Variant	# of Samples
Prostate	*TMPRSS2::ERG*	4
Colorectal	*VIT1A::TCF7L2*	2
Colorectal	*STRAP::DERA*	1
Colorectal	*PDE4D::SEC24A*	1
Colorectal	*IFT11::RHO*	1
Colorectal	*BIRC6::PLB1*	1
Colorectal	*ABR::NAALADL2*	1

Chapter 3: Novel method for detection of fusion-derived circRNA

Preface

This chapter has been adapted from the following publication:

Jace Webster, Hung Mai, Amy Ly, Christopher Maher. INTEGRATE-Circ and INTEGRATE-Vis: Unbiased Detection and Visualization of Fusion-Derived Circular RNA. *Bioinformatics*. 2023.

3.1 Introduction

Circular RNAs (circRNAs) occur when splicing mechanisms cause downstream exons to covalently bind to an upstream exon, referred to as a backsplice, resulting in a circular, rather than linear, transcript. Backsplicing events are thought to rely primarily on standard spliceosome machinery and are in part facilitated by complimentary sequences located within the introns that flank the donor and acceptor splice sites, although trans-acting factors are also involved[66]. CircRNAs have been shown to function through a variety of mechanisms, including direct regulation of transcription[68], indirect transcriptional regulation through interactions with microRNAs[69,70] or RNA-binding proteins[71], and by encoding peptides[72]. As circRNAs are not susceptible to degradation by exonucleases due to their circular structure, they are thought to be more stable than linear transcripts[43,44].

Fusion-derived circRNAs (fcircRNAs) are circRNAs that are generated by backsplicing within a gene fusion transcript and represent a recently discovered and poorly understood subset of circRNAs[73,76–78,104]. The gene fusion transcripts that form fcircRNAs are typically the result of genomic structural variation, such as translocations or deletions, that cause the 5’ end of a gene to become juxtaposed to the 3’ end of an independent gene. Such gene fusions are common in many cancers[53,64,105] and have been identified as druggable targets[106,107]. FcircRNAs can also have other sources, such as the backsplicing of read-through transcripts that contain multiple genes, although these have sometimes been referred to as read-through circRNAs[74,108]. For simplicity, we will refer to any circRNA that is composed of multiple independent genes as fcircRNAs.

While fcircRNAs remain poorly understood, recent studies have demonstrated they are functional. For example, fcircRNAs from *BCR::ABL1* fusions have shown oncogenic potential in leukemia[76] and *EML4::ALK* fcircRNAs were shown to promote cell migration and invasion in non-small cell lung cancer[75]. An additional 62 fcircRNAs have been reported within RNA-Seq data across a cohort of prostate cancer patients, but their potential functions were not investigated[79]. Considering the stability of circular transcripts and the somatic nature of most gene fusions, it is perhaps no surprise that early attempts have already been made to determine if fcircRNAs can be leveraged as cancer biomarkers[45].

Despite the oncogenic nature of some fcircRNAs and their potential as biomarkers, the study of fcircRNAs has been severely limited due to 1) the widespread use of Poly(A)-selection in RNA protocols which systematically removes circRNAs prior to sequencing and 2) a lack of software tools capable of detecting such events. As a result, most previously identified fcircRNAs were discovered through targeted sequencing of hypothetical backsplice junctions in gene fusions of interest[76]. We are aware of only three software tools developed for fcircRNA detection. The first published tool, Acfs[80], has systematic biases by algorithmically requiring fcircRNAs to be formed by fused genes originating from different chromosomes or from different strands of the same chromosome (removing the possibility of detecting an fcircRNA from a read-through transcript or from well-studied fusions like *TMPRSS2::ERG*). Acfs was initially developed only for circRNAs and although updated versions support fcircRNAs, default input parameters disable fcircRNA detection. The second tool, Fcirc[81], accepts unaligned reads as input and then uses a built-in aligner to map reads against custom reference sequences generated based on a user-supplied list of potential gene fusions. Finally, CircFusion[82] uses a nearly identical workflow as Fcirc but uses STAR[109] for performing read alignments.

Interestingly, both Fcirc and CircFusion require *a priori* knowledge via an input list of potential gene fusions thereby preventing unbiased fcircRNA discovery. Notably, the gene fusion list provided by Fcirc contains twice as many fusions as the list provided by CircFusion, highlighting an immediate discrepancy in the potential candidates that could be detected between tools. To our knowledge, there are no automated methods that allow the unbiased discovery of fcircRNAs throughout the full genome.

To address the need for improved fcircRNA detection methods, we have developed INTEGRATE-Circ. INTEGRATE-Circ is an open-source software tool capable of integrating both RNA and whole genome sequencing data to perform unbiased detection of novel gene fusions and report the presence of splice variants in gene fusion transcripts, including backsplicing events. We assessed the performance of INTEGRATE-Circ using simulated data and then demonstrate its utility through the analysis of leukemia and breast cancer cell lines. Additionally, we have released an update to our previously published tool, INTEGRATE-Vis, making it the first software capable of automatically generating publication-ready visualizations of fcircRNAs.

3.2 Results

3.2.1 INTEGRATE-Circ software

INTEGRATE-Circ leverages an algorithm originally developed for our highly accurate fusion discovery software, INTEGRATE[102]. The original INTEGRATE algorithm was designed to analyze RNA-Seq, and when available include whole genome sequencing (WGS), paired-end reads to detect high confidence, novel gene fusion events. A comparison with 8 gene fusion detection tools demonstrated that INTEGRATE was the most accurate method. As such, the methodology behind INTEGRATE serves as a strong starting point for developing tools that can detect junctions between fused genes.

A thorough explanation of the original INTEGRATE fusion detection algorithm is provided in the INTEGRATE publication, but a brief overview is provided here to give context for the changes that are implemented in INTEGRATE-Circ. The original workflow involves the creation of a gene graph such that each node consists of a gene and each edge is based on discordantly mapped read pairs that may encompass a fusion junction between the two genes. Initial pruning of the graph is performed, primarily through the re-alignment of discordant read pairs. Potential spanning reads and previously unmapped reads are then mapped to remaining gene nodes and their ‘neighboring node(s)’ in an attempt to identify spanning read support for putative fusions and reads that are aligned near each other are clustered together to identify potential fusion junctions. Fusion junctions that are supported by the mapped RNA-Seq spanning reads are then compared against WGS reads to allow for single-base pair resolution of the genomic breakpoints, if WGS data is provided.

INTEGRATE-Circ builds upon the INTEGRATE framework by using the location and orientation of detected junctions to infer the existence of unique isoforms generated by alternative splicing or backsplicing mechanisms. A general overview of this workflow is depicted in Figure 3-1. After identifying potential gene fusions, clusters of junction-spanning RNA-Seq reads are re-evaluated. For each potential fusion, each cluster of spanning RNA-Seq reads are compared to each other to determined which cluster has the highest read support, with the most well-supported junction being considered the primary fusion. The primary fusion is expected to correspond to the true genomic fusion junction and should be supported by WGS data, if available. All other spanning read clusters are then evaluated with respect to the primary

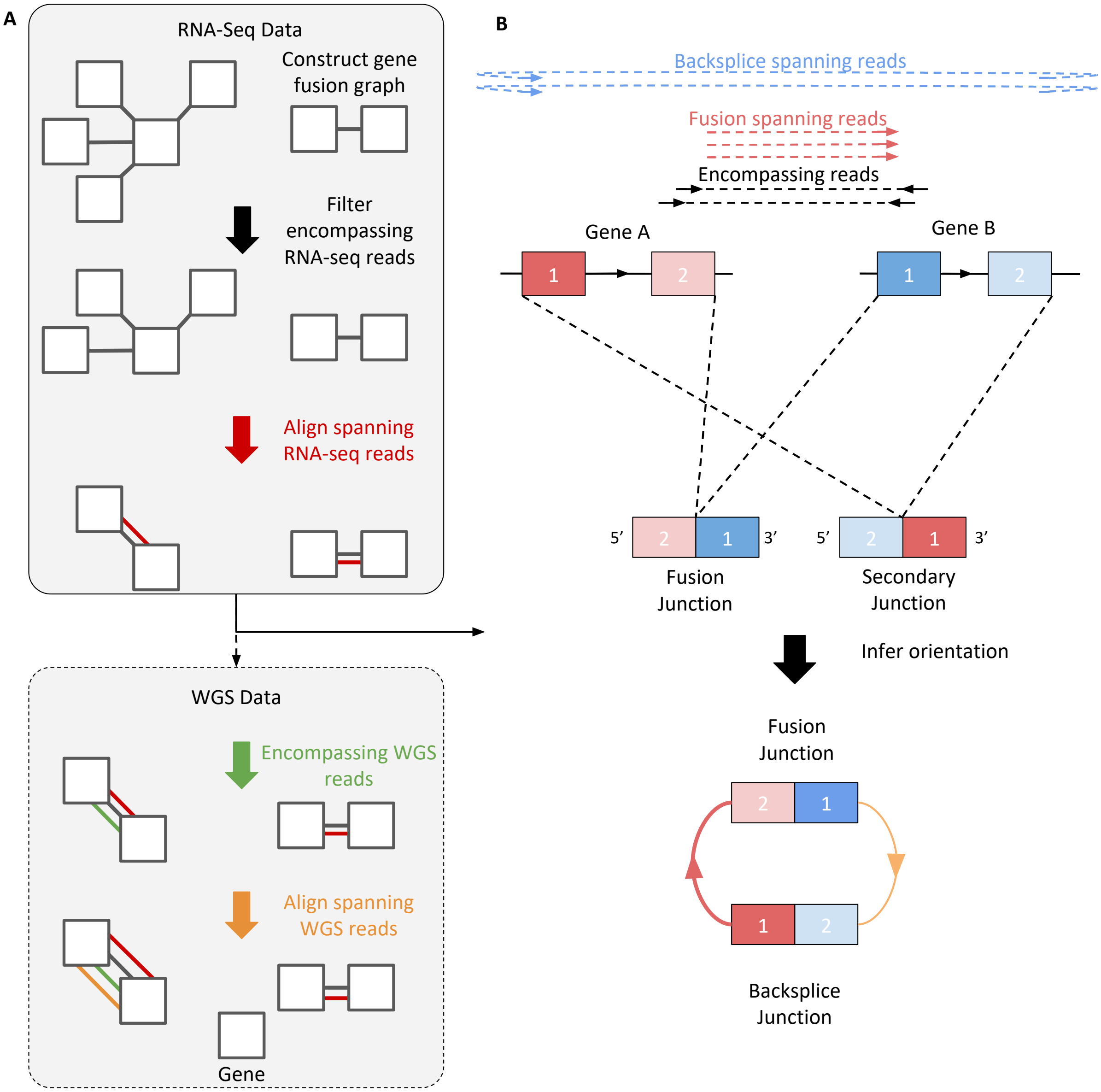


Figure 3-1. INTEGRATE-Circ workflow. A) INTEGRATE-Circ begins by creating a gene graph of potential fusions based on RNA-Seq data and removing nodes from the graph based on encompassing and spanning read support. If provided, encompassing and spanning WGS reads are then examined for additional evidence for fusions. B) Once gene fusion candidates have been identified, all RNA reads that span both genes are clustered together based on region to identify the locations of gene junctions. The junction with the most support is identified as the fusion and junction and all other junctions are then evaluated based on their orientation and positioning with respect to the fusion junction.

fusion junction. Since secondary junctions are thought to result from alternative splicing of transcripts are not expected to be a direct indication of genomic rearrangements, secondary junctions are not expected to have WGS support. A simplified schematic demonstrating how spanning read clusters are annotated based on their relative orientation to the primary junction is depicted in Figure 3-S1, although a much broader variety of potential secondary junction orientations, including those that do not match with canonical exon boundaries, are possible. INTEGRATE-Circ applies an extended version of the logic described in the schematic to all identified gene junctions. Where possible, junctions are compared based on canonical exon boundaries to aid in identifying reciprocal gene fusions. In cases where identified junctions are not located at annotated exon boundaries, relative locations and orientations are evaluated based on genomic base pair position for annotation purposes. By combining insights from RNA-Seq and WGS, INTEGRATE-Circ is designed to sensitively detect gene fusion junctions and be able to differentiate between genomic rearrangements and alternatively spliced transcripts, including backsplices.

All identified junctions from linear and circular transcripts (including read-throughs) are reported by INTEGRATE-Circ using a number of standardized formats, including bedpe, vcf and generic tsv formats with accompanying annotation information (including gene names, total RNA-Seq/WGS read support, a list of supporting reads, and whether the junction uses canonical exon boundaries). Although bedpe and vcf files are commonly used for annotating standard fusion breakpoints, no standardized file format exists to specifically describe fcircRNAs. Therefore, INTEGRATE-Circ reports fcircRNAs using a modified bedpe file, described in the README file of the project GitHub page. This modified bedpe format is consistent with the output file generated by Fcirc to help ensure consistency with other downstream applications in the future. This file format is accepted by INTEGRATE-Vis for fcircRNA visualization.

All benchmarking of INTEGRATE-Circ was performed on a big memory blade with 32 Intel Xeon CPU E5-2640s with 400G of memory. On our largest dataset (HCC1395 RNA-Seq with ~200 million paired-end reads), the run time for INTEGRATE-Circ on the big memory blade was approximately 1.5 hours.

INTEGRATE-Circ requires paired-end sequencing data in order to identify encompassing reads when creating the initial gene graph and therefore cannot accept single-end reads. Required sequencing depth varies based on how highly expressed a given fcircRNA may be. In our analysis of HCC1395, a total of ~200 million paired-end reads were generated and all validated fcircRNAs <5 supporting RNA-Seq reads. It is possible that deeper sequencing may be required for less abundant isoforms and to reduce the likelihood of false positives. In our testing, we found that INTEGRATE-Circ performed optimally on relatively short reads and we recommend the use of 2x75bp or 2x100bp read lengths.

3.2.2 INTEGRATE-Vis software

The fcircRNA visualization workflow within INTEGRATE-Vis consists of two primary steps: annotation and visualization. The annotation step uses a user-provided, standard GTF file to determine the exon boundaries of exons located immediately around the reported fusion and backsplice junctions. If a junction does not match canonical exon boundaries, the enarest upstream (for 5’ end of junctions) or downstream (for 3’ end of junctions) exon boundary is selected for visualization purposes. Additionally, genomic cytoband information for the chromosome(s) involved in the fcircRNA is extracted from the user-provided ideogram file in order to put the genomic location of the fcircRNA into context.

The second step in the workflow is the creation of the visualization using the annotation information generated during the previous step. The genomic locations of fusion genes are presented based on cytoband location and the resulting fusion gene transcript is presented. The presented fusion transcript contains a minimal number of exons (max of 3 per gene), but does not necessarily represent the full transcript length, nor are the exons presented to scale. The fcircRNA is then presented in relation to the fusion transcript. Optionally, the user may provide a bam file from which INTEGRATE-Vis will attempt to identify the number of spanning reads that support both reported junctions. As both INTEGRATE-Circ and Fcirc perform their own custom secondary alignment steps, it is possible that the read support values calculated by INTEGRATE-Vis will differ from those reported by INTEGRATE-Circ and/or Fcirc. Additional details can be found at https://github.com/ChrisMaherLab/INTEGRATE-Vis.

3.2.3 *In silico* simulation

To perform an initial comparison between the identified fcircRNA detection tools, a simulated dataset containing 30 linear fusion transcripts was generated based on the most frequent gene fusions reported in the COSMIC database[110]. Randomly generated backsplice junctions were then created for each of the fusion transcripts based on the exons present in the reported fusion transcript. RNA-sequencing reads for the fusion and backsplice junctions were simulated 100 separate times. An overview of this workflow and the resulting simulated inter- and intra-chromosomal events can be found in Figure 3-2 and Table 3-S1.

For benchmarking purposes, INTEGRATE-Circ, Fcirc, Acfs and CircFusion were applied to the simulated data with default settings (besides Acfs, which was given the 'Search_trans_splicing yes' parameter to support fcircRNA detection). Since Fcirc and CircFusion both require a list of potential gene fusions, the provided gene fusion lists (downloaded from GitHub for each tool on March 10, 2022 and February 7, 2023, respectively) were used as input (305 fusions for CircFusion and 773 fusions for Fcirc). CircFusion failed to run with default settings with exit warnings suggesting that the 305 gene fusion list was too large. This failure, combined with the fact that CircFusion accepts the expected transcript IDs, fusion breakpoints and backsplice junctions of potential fcircRNAs, suggests that CircFusion may be better optimized for validation of specific, previously identified events and led us to exclude CircFusion from the remaining benchmarking analyses. Similarly, although we were able to run Acfs successfully on the tool's provided example data, the tool reported no fcircRNAs in our simulated data (and reported no errors at runtime). It is possible that the fcircRNA detection portion of Acfs has poor sensitivity, as it failed to detect any fcircRNAs in real sequencing data in its original publication. Sensitivity, precision and F1 accuracy scores for results from INTEGRATE-Circ and Fcirc were then calculated for each of the 100 simulation iterations. We found that when comparing fcircRNA detection between INTEGRATE-Circ and Fcirc, INTEGRATE-Circ was superior in terms of sensitivity (mean: 87.3% +/- 4% vs 57.1% +/- 2%), precision (mean: 96.1% +/- 3% vs 74.2% +/- 4%) and F1 accuracy (mean: 91.5% +/- 4% vs 64.5% +/- 2%) (Figure 3-2C). Notably, if Acfs had consistently achieved the maximum sensitivity that its algorithm would allow (it systematically excludes 4 of the simulated backsplices because their contributing genes originated on the same strands of the same chromosomes), the maximum possible sensitivity of the tool in this simulation would be 86.6%, meaning that it could not have outperformed INTEGRATE-Circ's average sensitivity.

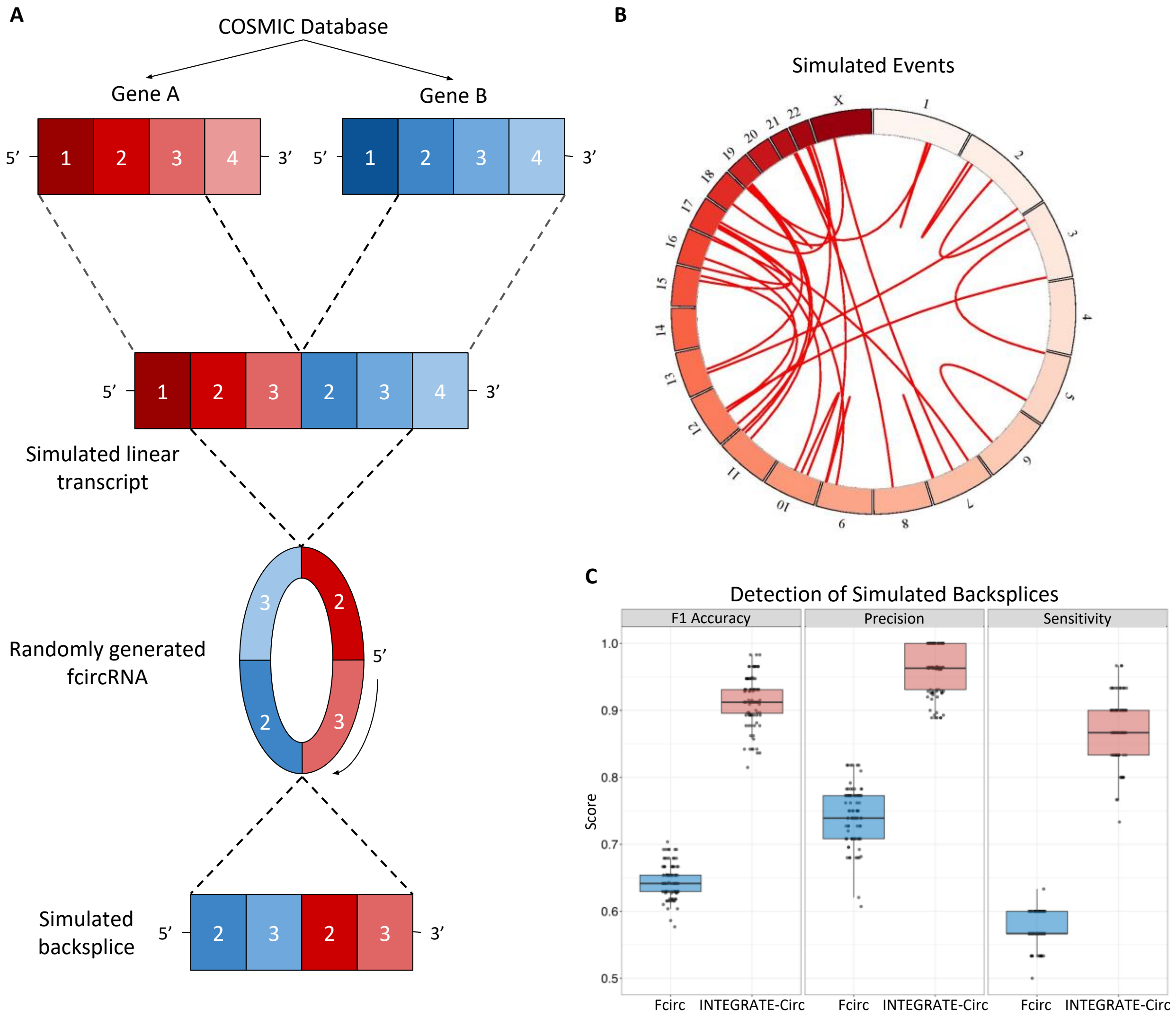


Figure 3-2. Benchmarking results based on *in silico* simulation. A) Schematic depicting the creation of simulated transcripts. Recurrent gene fusions were identified from the COSMIC database and theoretical backscplice junctions were then randomly introduced to the selected fusions. Linear fusion transcripts and a linearized version of the region that spans the simulated backsplice were used to simulate RNA-Seq reads. B) Circos plot representing all simulated events. C) F1 accuracy, precision and sensitivity scores after analysis of the simulated fcircRNAs by both tools across 100 iterations.

3.2.4 Application to public K562 cell line data

Next, we applied INTEGRATE-Circ, Acfs and Fcirc to the K562 lymphoblast cell line which contains four validated linear fusion transcripts, three of which have published support for the presence of fcircRNAs in either K562[76] or in a different context[74,108]. A summary of the results regarding the four previously validated fusion transcripts are shown in Table 3-1. For previously published junctions we required on or more reads. For novel junctions we required two or more independent reads. We found that Fcirc detected only on of the published linear gene fusions and no corresponding fcircRNAs while INTEGRATE-Circ detected all four linear fusions and reported fcircRNAs in three of the four fusions. Of the three fcircRNAs called by INTEGRATE-Circ, two have been previously reported (circ*PRKAA1*(5,6,7,8,9,10)*::TTC33*(1,2)[74] and circ*KANSL1*(3)*::ARL17A*(3)[108]) while one (circ*NUP214(25,26,27,28,29)::XKR3(2,3)*) was novel.

Table 3-1. Results of K562 analysis. Supporting reads for previously reported linear fusion transcripts in K562 and any fcircRNAs that may derive from those transcripts. Missing values indicate that no junction was reported.

	Read Support			
	Linear Fusion Junction		Backsplice Junction	
Fusion	INTEGRATE-Circ	Fcirc	INTEGRATE-Circ	Fcirc
BCR::ABL1	960	1146	-	-
PRKAA1::TTC33	12	-	1	-
KANSL1::ARL17A	35	-	9	-
NUP214::XKR3	356	-	3	-

Unfortunately, neither tool was able to detect the circ*BCR*(13,14)*::ABL1*(2,3) fcircRNAs that were previously reported in this cell line[76], however this result is consistent with a previous attempt to detect fcircRNAs using this same public sequencing data which also failed to detect the circ*BCR*(13,14)*::ABL1*(2,3) isoforms[74].

3.2.5 Application to HCC1395 cell line

For a final evaluation, we applied INTEGRATE-Circ, Acfs and Fcirc to the breast cancer cell line HCC1395. This cell line was chosen because it has significantly more validated fusions that K562 but has not previously been evaluated for the presence of fcircRNAs.

To ensure that each tool was working as intended, both Poly(A)-selected and total RNA sequencing data was analyzed. We expect that no fcircRNAs would be found in the Poly(A)-selected data due to the removal of circular transcripts during the poly(A) enrichment. As anticipated, INTEGRATE-Circ only reported fcircRNAs in the total RNA data (Figure 3-3A). In contrast, Fcirc unexpectedly nominated 16 fcircRNAs in the Poly(A)-selected data, nearly 3x more fcircRNAs than it reported in the total RNA data. None of the fcircRNAs called by Fcirc in the total RNA data were reported in the Poly(A)-selected data, or vice versa. We focused our remaining analysis on the total RNA data results since fcircRNAs observed in the Poly(A)-selected data were thought to be potential noise and are suggestive of Fcirc having a potentially high false positive rate.

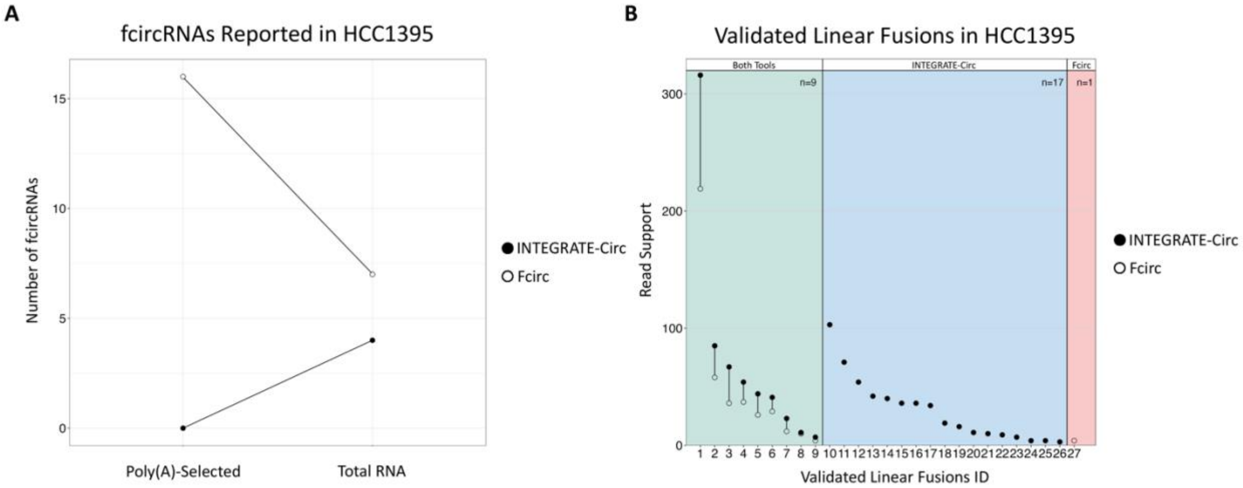
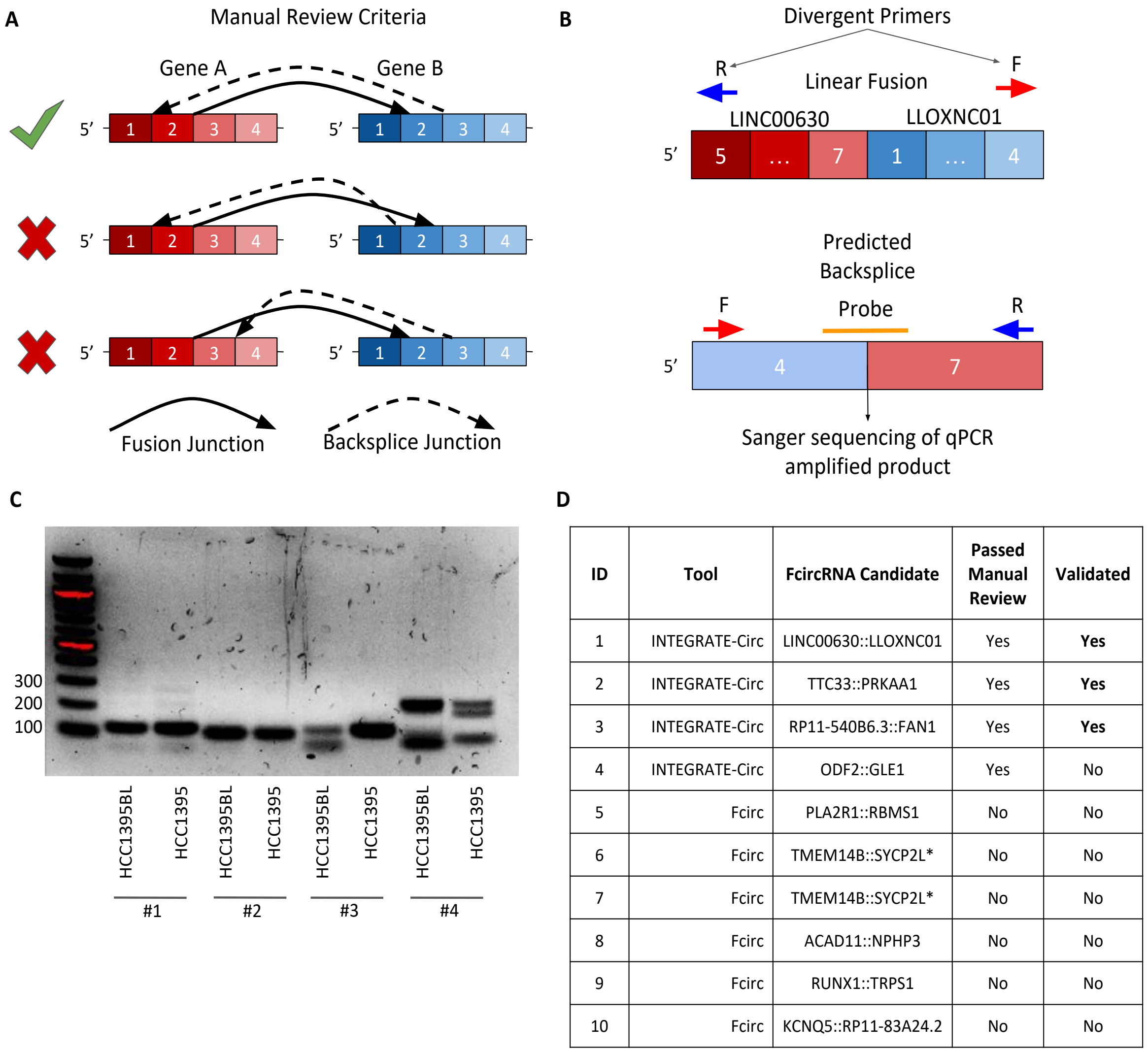


Figure 3-3. Analysis of HCC1395 cell line data. A) Number of fcircRNAs reported by INTEGRATE-Circ and Fcirc when applied to Poly(A)-selected and total RNA sequencing data. B) Reported read support for previously validated HCC1395 linear fusions.

As all fcircRNAs must, by definition, be a subset of the detected fusion transcripts, we next compared INTEGRATE-Circ and Fcirc linear fusion calls made using the total RNA data against a published list of validated fusions in HCC1395[102], requiring >2 supporting independent reads. In the 9 validated fusions that were reported by both tools, we found that INTEGRATE-Circ reported greater read support in 100% of the fusions (Figure 3-3B). Additionally, 17 previously reported fusions were found by INTEGRATE-Circ alone while only 1 published event was found solely by Fcirc. Two additional validated gene fusions called by Fcirc failed to meet our filtering criteria and were missed by INTEGRATE-Circ because the reads were either not mapped to the gene of interest and/or no encompassing reads were detected (which is required by INTEGRATE-Circ).

3.2.6 *In vitro* validation of novel fcircRNAs

Finally, we attempted to validated predicted fcircRNAs using PrimeTime Probe reverse transcription quantitative PCR (PrimeTime Probe qPCR) amplification of putative backsplices. Divergent primers for all fcircRNA candidates that passed manual review (Figure 3-4A) from either tool were designed using the strategy depicted in Figure 3-4B, as has been described previously for both circRNA[111] and fcircRNA[78] validation. We also performed the PrimeTime Probe qPCR assay on the HCC1395 B Lymphocyte (HCC1395BL) cell line, which serves as a matched control cell line. PrimeTime Probe qPCR amplified products were run on a gel (Figure 3-4C) and purified products from the HCC1395 cell line were excised from the gel and Sanger sequenced, confirming the presence of the circ*TTC33*(1,2,3)::*PRKKA1*(3,4,5), circ*LINC00630*(5,6,7)*::LLOXNC01-237H1.2*(1,2,3,4) and circ*RP11-540B6.3*(1)*::FAN1*(1) fcircRNAs reported by INTEGRATE-Circ in HCC1395 (Figure 3-4D). Notably, PrimeTime Probe qPCR products consistent with the size of circ*TTC33(1,2,3)::PRKKA1*(3,4,5) and circ*LINC00630*(5,6,7)*::LLOXNC01-237H1.2*(1,2,3,4) were detected in the HCC1395BL cell line as well as the cancer cell line (Figure 3-S2). As these fcircRNAs appear to be derived from read-through transcripts and are not the result of somatic structural variation, it is perhaps unsurprising that evidence for them was found in both cell lines, rather than only in the cancer cell line. We were unable to confirm the presence of any of the fcircRNAs reported by Fcirc. Similarly, as each validated fcircRNA was composed of genes from the same strands of the same chromosomes, the Acfs algorithm would have been able to detect any of the validated fcircRNAs.



ID	Tool	FcircRNA Candidate	Passed Manual Review	Validated
1	INTEGRATE-Circ	LINC00630::LLOXNC01	Yes	**Yes**
2	INTEGRATE-Circ	TTC33::PRKAA1	Yes	**Yes**
3	INTEGRATE-Circ	RP11-540B6.3::FAN1	Yes	**Yes**
4	INTEGRATE-Circ	ODF2::GLE1	Yes	No
5	Fcirc	PLA2R1::RBMS1	No	No
6	Fcirc	TMEM14B::SYCP2L*	No	No
7	Fcirc	TMEM14B::SYCP2L*	No	No
8	Fcirc	ACAD11::NPHP3	No	No
9	Fcirc	RUNX1::TRPS1	No	No
10	Fcirc	KCNQ5::RP11-83A24.2	No	No

Figure 3-4. Validation of HCC1395 fcircRNAs. A) Manual review process for fcircRNAs. While the top example represents the expected relative locations of fusion and backsplice junctions, the other schematics represent scenarios where the backsplice donor would not be present in the fusion transcript (middle example) or the backsplice acceptor would not be present in the fusion transcript (bottom example). Reported fcircRNAs that follow the middle or bottom examples are physically impossible as an fcircRNA must be a subset of the sequence present in the fusion transcript and were therefore excluded from PrimeTime Probe qPCR validation. B) Design of divergent forward and reverse primers that face away from the fusion junction and placement of PrimeTime Probes to span the reported backsplice junction. The design for LINC00630::LLOXNC01 is shown, but an identical procedure was used for each

candidate that was evaluated. C) Gel of the amplified PrimeTime Probe qPCR products for both the HCC1395 cancer cell line and the matched normal tissue HCC1395BL. Bands of the expected size were excised and sent for Sanger sequencing. D) All reported fcircRNA candidates. A validation status of "Yes" denotes that the Sanger sequencing of PrimeTime Probe qPCR products matched the expected backsplice junction sequence. The * indicates that multiple fcircRNA isoforms were reported to result from the same fusion transcript.

3.2.6 Visualization of novel fcircRNAs using INTEGRATE-Vis

Currently there are no publicly available tools for visualizing fcircRNAs. Most studies have relied on the manual creation of schematics to convey their findings, which can be time consuming, leads to highly variable figure quality between studies, and can cause confusion when trying to accurately depict complex fcircRNA isoforms which, until recently, lacked a formalized nomenclature[104]. To improve the dissemination of information in this relatively new field, we implemented an updated version of INTEGRATE-Vis (v1.1.0). In addition to the visualizations of linear fusion transcripts which were supported by earlier version of INTEGRATE-Vis[112], the tool now supports the visualization of detected fcircRNAs and is compatible with both INTEGRATE-Circ and Fcirc output files. Example outputs using default settings are shown in Figure 3-5, which depicts the three novel, validated HCC1395 fcircRNAs identified by INTEGRATE-Circ.

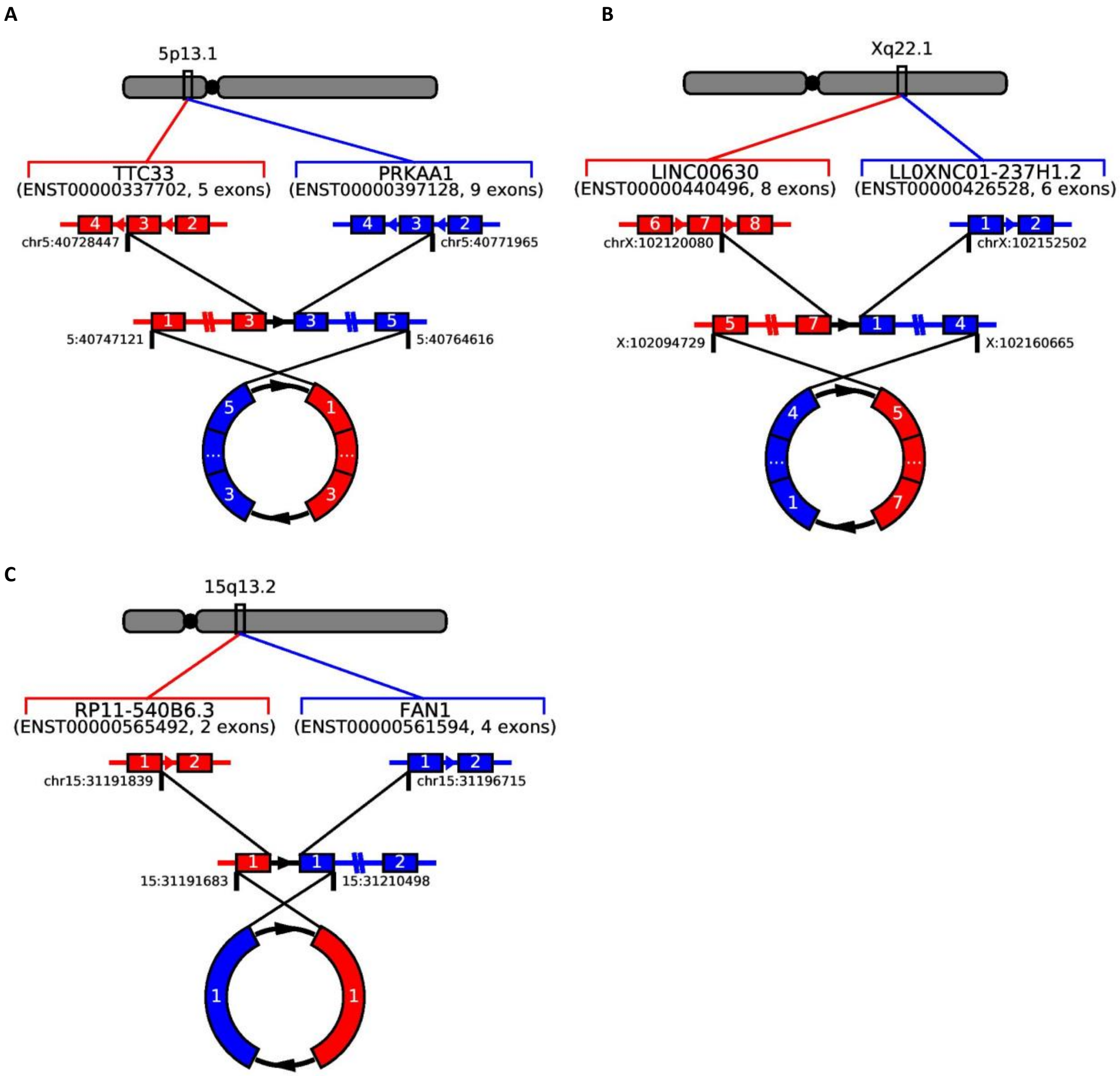


Figure 3-5. Validated HCC1395 fcircRNAs visualized with INTEGRATE-Vis. Default output from INTEGRATE-Vis depicting the validated A) TTC33::PRKAA1, B) LINC00630::LLOXNC01-237H1.2 and C) RP11-540B6.3::FAN1 fcircRNAs.

3.3 Discussion

Here we present both the novel tool, INTEGRATE-Circ, and v1.1.0 of INTEGRATE-Vis. Together, these open-source software tools allow for unbiased detection and visualization of novel fcircRNAs. Through the use of (1) simulated data, (2) publicly available cell line data, and (3) experimental validation in a paired breast and normal cell line, we have demonstrated the ability of INTEGRATE-Circ to accurately identify linear fusion transcripts and fcircRNAs using short-read, paired-end sequencing data in an unbiased fashion.

One potential limitation of the INTEGRATE-Circ algorithm is that all annotations assume that the junction with the most spanning read support is the true fusion junction. This assumption may be false in situations where an alternative splice variant of a fusion transcript is more abundant than the transcript that represents the full genomic fusion. The inclusion of WGS data in the INTEGRATE-Circ algorithm is meant to minimize the likelihood of incorrectly designating an alternatively spliced junction as the primary junction, as the WGS reads should only support the true genomic fusion. Although users can run INTEGRATE-Circ without WGS data, including this information is likely to improve performance when trying to avoid such scenarios.

While fcircRNAs are composed of sequences from different genes, there are multiple ways for disparate gene sequences to become part of the same transcript, such as gene fusions and read-throughs. Each of the isoforms validated in the HCC1395 cell line in this study were the result of read-through transcripts, sometimes referred to as read-through circRNAs (rt-circRNAs) instead of fcircRNAs. Some events fir poorly into any current characterization, such as the circ*KANSL1*(3)*::ARL17A*(3) transcript identified by INTEGRATE-Circ in the K562 cell line and previously reported in a medulloblastoma patient and other cell lines[74,108]. *ARL17A* is immediately upstream of the adjacent *KANSL1* on chromosome 17, but their positions are inverted as *KANSL1* becomes the 5' gene partner of the *KANSL1::ARL17A* fusion transcript that later gives rise to the associated circRNA[108]. The resulting circRNA is therefore not a typical read-through event, nor does it necessarily involve a genomic alteration. Indeed, *KANSL1::ARL17A* circularized transcripts have been referred to as both fcircRNAs[108] and as rt-circRNAs[74] in published literature. In either case, the capability of INTEGRATE-Circ to detect circRNAs resulting from both read-throughs and larger intra-/inter-chromosomal fusions, as evidenced by our analysis of cell line data and a variety of simulated fusion events, is indicative of the broad utility of our unbiased approach.

As demonstrated by the performance of INTEGRATE-Circ in both breast cancer and leukemia cell lines, this approach has broad applicability independent of the cancer type. Indeed, as seen in our analysis of the healthy normal HCC1395BL cell line, fcircRNAs caused by read-throughs can be present even in healthy normal tissue. It is possible that fcircRNAs are more prevalent in diseases where structural variation is a common feature, but prior limitations have prevented comprehensive studies. Similarly, it is possible that their prevalence increases later in disease development due to the accumulation of somatic mutations. By providing improved detection and visualization methods, we hope that future work will be able to address such questions.

In summary, we have demonstrated that the novel software tool, INTEGRATE-Circ, can sensitively and accurately identify both linear fusion transcripts and fcircRNAs with single-base pair resolution, in an unbiased manner, across a variety of datasets. Additionally, the companion tool INTEGRATE-Vis is the first to provide automated visualization of fcircRNAs. We anticipate that the combined use of these tools will facilitate a wide variety of future studies to better understand the basic and clinical significance of fcircRNAs.

3.4 Methods

3.4.1 Simulated data generation

The Gene Fusion Curation portion of the COSMIC v96 database[110] was used to identify 30 recurrent gene fusions for use as a basis for the *in silico* simulation. Fusions were first ranked by the number of mutated samples and then the single most common isoform for each fusion was selected for simulation. No gene was permitted to appear in more than one selected fusion, meaning that some fusions were skipped because promiscuous genes were listed as frequently having multiple gene partners (for example, the *SS18*::*SSX1* and *SS18*::*SSX2* fusions were both highly recurrent, but only the *SS18*::*SSX1* fusion was selected). The positions of each individual exon present in the selected fusions were then identified using Ensembl's hg19 annotation and individual exon sequences were then isolated using *bedtools getfasta -name -s -fi <hg19.fa> -bed <all_exons.bed> > exons.fa*[113]. Backsplices with randomly generated junctions (using canonical exon boundaries) were then designed for each of the 30 fusions. All linear fusion isoforms selected from COSMIC and their corresponding fcircRNAs were then assembled by grouping together the necessary exons from the *exons.fa* file to form multi-exon transcripts.

Next, RNA-Seq reads were generated using the simReads() function from the Rsubread R library (R version 4.0.0, Rsubread version 2.4.3)[114]. The initial random seed was set to 42 and then reads were simulated over 100 iterations using the following parameters: *library.size = 100000, read.length = 100, paired.end = True, simulate.sequencing.error = True*. No WGS data was simulated.

3.4.2 K562 data

K562 cell line sequencing data was downloaded from the Sequence Read Archive using accession number SRR8587462. No WGS data was used for analysis. Data was prepared for analysis as described in section 3.4.6.

3.4.3 HCC1395 sequencing data

Poly(A)-selected HCC1395 sequencing data was downloaded from the public Sequence Read Archive, accession number SRR892423. For total RNA sequencing of HCC1395, a total RNA input of 1μg was used to generate the RNA-Seq library using the New England BioLabs NEBNext Ultra II Directional RNA Library Prep for Illumina kit with rRNA Depletion module and NEBNext Multiplex Oligos for Illumina (Unique Dual Index UMI Adaptors RNA Set 1) per manufacturer's protocol. Paired-end sequencing was performed on the NovaSeq platform.

3.4.4 Manual review of fcircRNA calls in HCC1395

Manual review was performed based on the fusion and backsplice junctions reported in the final output files of INTEGRATE-Circ and Fcirc to ensure that the junctions were in an orientation that was capable of forming an fcircRNA. An example of some of the patterns observed are present in FIGURE. As the orientation of the original genes and the junctions are known, it is possible to predict the potential outcomes of any combination of junctions. The top example in FIGURE depicts the expected pattern during the manual review and describes an fcircRNA that would contain only the 2nd exon of Gene A and the 3rd exon of Gene B. The middle example from FIGURE depicts a fusion junction connecting Gene A exon 2 to Gene B exon 2, but suggests that Gene B exon 1 is a backsplice donor. As Gene B exon 1 is not contained in the fusion transcript, it is not possible for Gene B exon 1 to be involved in the backsplice junction. Similarly, the final example in Figure 4A depicts a backsplice acceptor (Gene A exon 4) that would not be present in the linear fusion transcript. Scenarios such as there were excluded from PrimeTime qPCR validation as they were considered impossible and representative of a software error.

3.4.5 Validation of fcircRNA calls in HCC1395

Potential candidates were precisely validated using PrimeTime qPCR Probe Assays (Integrated DNA Technologies) with forward and reverse primer sets covering approximately 50nt each side of the backsplice junctions, and qPCR probes that specifically spanned the backsplice junctions themselves. Amplification was performed using the manufacturers recommended protocols. Amplified PrimeTime Probe qPCR products were purified using DNA Clean & Concentrator-5 (ZYMO RESEARCH) and purified DNA was analyzed with Sanger sequencing. Output from Sanger sequencing was compared to predicted backsplice junction sequences to assess the presence of the fcircRNA.

Selection and excision of DNA from the gel was done based on an expected size of approximately 100bp for each fcircRNA (based on the distance of primers from backsplice junctions, FIGURES). Non-specific bands were thought to be caused by non-specific binding to different transcripts, in part due to non-ideal PCR conditions. For example, due to having a large number of primers being used during the PCR reactions, the annealing temperature was sub-optimal for a small number of primers. Similarly, due to the low expected abundance of fcircRNA, a relatively large number of PCR cycles were performed (45 cycles), which may also introduce non-specific binding.

3.4.6 Alignment of all sequencing data

Fcirc performs its own sequence alignment and accepts unaligned fastq files as input. Therefore, all reads from the *in silico* simulation, public K562 data and HCC1395 sequencing were provided to Fcirc in an unaligned format and were then aligned to hg19 by Fcirc.

In the original INTEGRATE publication[102], it was shown that performance can vary slightly based on the aligner used and that optimal performance was achieved by using GSNAP for initial alignment. For this reason, all reads analyzed by INTEGRATE-Circ were aligned by GSNAP (version 2021-03-08)[115] using the following parameters: *-d hg19, --novelsplicing=1,* --*read-group-platform=Illumina, --extend-soft-clips*. Aligned reads were then sorted using *samtools sort* (v1.7) (CITATION) and processed by INTEGRATE-Circ using default parameters. Although INTEGRATE-Circ accepts WGS data, no WGS data was used for benchmarking purposes in order to provide a fair comparison between the different tools and because WGS data is not necessary for fcircRNA detection.

For Acfs, reads were prepared using the recommended steps for paired-end sequencing data provided on the project's GitHub page (https://github.com/arthuryxt/acfs).

3.5 Supplementary Materials

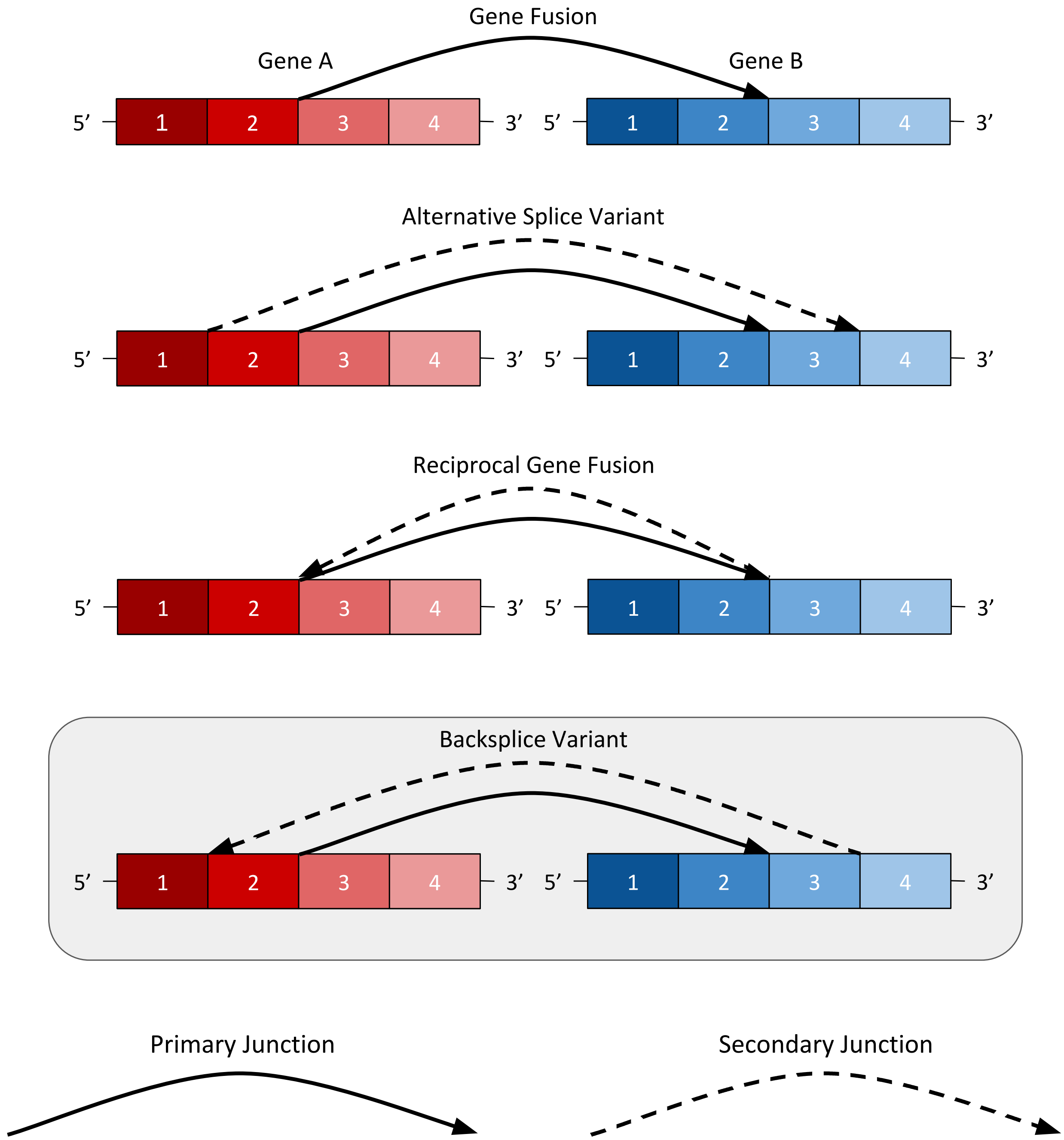


Figure 3-S1. Junction annotation strategy used by INTEGRATE-Circ. Simplified schematic of how INTEGRATE-Circ compares secondary junctions to the primary fusion junction in order to assign annotations.

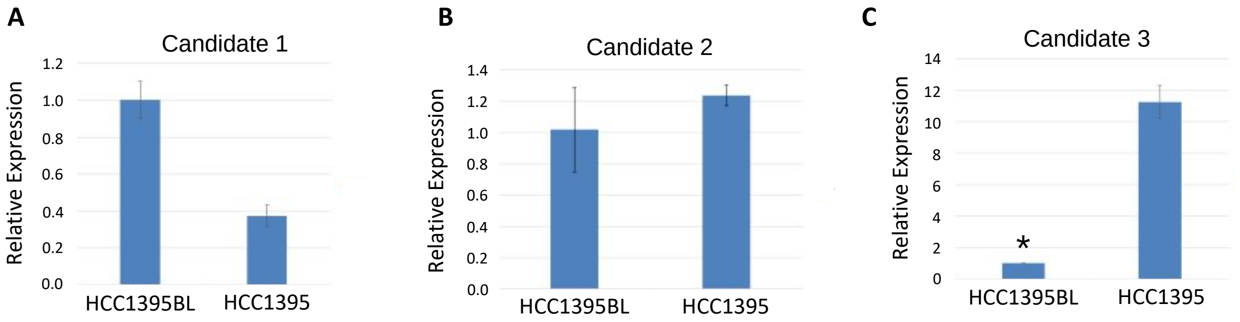


Figure 3-S2. Quantification of fcircRNA expression in HCC1395 and HCC1395BL. A-C) Relative expression of fcircRNA candidates 1-3 based on PrimeTime Probe qPCR. Expression values for each candidate are normalized to the expression found in HCC1395BL. *-In the case of Candidate 3, no expression was detected for HCC1395BL. For visualization purposes, HCC1395BL expression was therefore set to a "hypothetical" Tm value of 36.

Table 3-S1. Fusions used for *in silico* simulation

5’ Fusion Gene	3’ Fusion Gene
BCR	ABL1
TMPRSS2	ERG
EWSR1	FLI1
PML	RARA
EML4	ALK
KIAA1549	BRAF
CCDC6	RET
SS18	SSX1
RUNX1	RUNX1T1
PAX3	FOX01
FUS	DDIT3
COL1A1	PDGFB
CRTC1	MAML2
NAB2	STAT6
ETV6	NTRK3
CBFA2T3	GLIS2
KMT2A	MLLT1
PAX8	PPARG
ASPSCR1	TFE3
HMGA2	LPP

JAZF1	SUZ12
SET	NUP214
CD74	ROS1
TPM3	NTRK1
CTNNB1	PLAG1
TAF15	NR4A3
CKDN2D	WDFY2
YWHAE	NUTM2B

Chapter 4: Methylation biomarkers of aggressive prostate cancer

4.1 Introduction

Although localized, indolent PCa has a 98% 5-year survival rate, survival drops to 30% in metastatic disease over that same time period[1,116]. For this reason, being able to reliably distinguish between indolent and aggressive tumors is critical. Traditionally, this has been done using Gleason grade scores[14]. The Gleason score is assigned based on morphological differences in tumor cells and is reported as the sum of the score assigned to two separate regions of the tumor on a 1-5 scale, with indolent regions receiving a 3 and more aggressive tumors receiving 4-5. More recently, measures of prostate specific antigens (PSA) are also used to measure tumor progression[9,10,13]. Unfortunately, over-treatment is a widely observed problem in early PCa and improved stratification strategies could lead to direct improvements in treatment strategies and the quality of life of the patient[9,10,13].

Compared to other cancers, PCa has a relatively low tumor mutation burden and few recurrent SNVs and SVs have been identified in localized disease[27]. For this reason, biomarkers based on genetic mutations are not tractable. In lieu of this, a number of RNA expression-based assays have been developed in this space with varying degrees of success[15,17]. Unfortunately, RNA is less stable than DNA which can make quality control more complicated. Similarly, since PCa may remain indolent for years, a liquid biopsy-based assay would be beneficial in this space to allow for non-invasive, long-term monitoring of disease progression, but RNA is difficult to detect in blood[43].

Methylation profiling has emerged as an attractive strategy for biomarker detection because it does not rely on tumor mutation burden and because it can be measured directly from DNA (including cfDNA)[19]. A number of methylation changes in aggressive PCa have already been well documented, such as hypermethylation of *GSTP1* and observations that methylation of *MYC* correlate with Gleason scores[40,117]. To expand on these findings, recent work has been done to build classification algorithms to stratify PCa tumors based on methylation data[118]. Notably, most studies have relied on bulk sequencing using Illumina 450k methylation arrays applied to heterogeneous cohorts. Given the highly heterogeneous nature of PCa, it is possible that such approaches may lack in sensitivity. Indeed, many of these studies separate samples based on total summed Gleason score, such that a sample rated as 5+3 would be grouped with a 4+4 sample, masking the present heterogeneity within each tumor. Furthermore, as 450k methylation arrays only measure a pre-specified region of the genome, it is possible that whole genome sequencing could reveal new information associated with the transition from indolent to aggressive disease.

To begin to address these limitations, our group has acquired matched Gleason grade 3 and grade 4/5 samples captured using laser-capture micro-dissection from the same tumors, allowing for intra-tumor comparisons while minimizing the inherent heterogeneity of bulk analyses, as part of an exploratory study. Whole genome analysis was then performed using Enzymatic Methylation Sequencing (EM-Seq) to help gain insights into regions not detectable by methylation arrays[119,120]. Using this approach, we were able to perform an initial, exploratory assessment of the differential methylation in indolent and aggressive disease and develop an initial epigenetic signature to aid in tumor classification. This work serves as a proof of concept for the application of whole genome methylation sequencing in this setting and highlights the benefits of the approach. Future studies are required to validate the findings presented here.

4.2 Results

A total of 23 aggressive and 11 indolent samples, including 6 matched pairs, were obtained for this exploratory study. Samples were deeply sequenced using EM-Seq resulting in an average of 615M reads per sample.

4.2.1 EM-Seq results are consistent with published data

For our initial analysis, we sought to ensure that the results achieved by EM-Seq were comparable to published 450K microarray data when examining CpG sites present in both datasets. Indolent samples (Gleason 3) sequenced with EM-Seq were compared to PCa samples

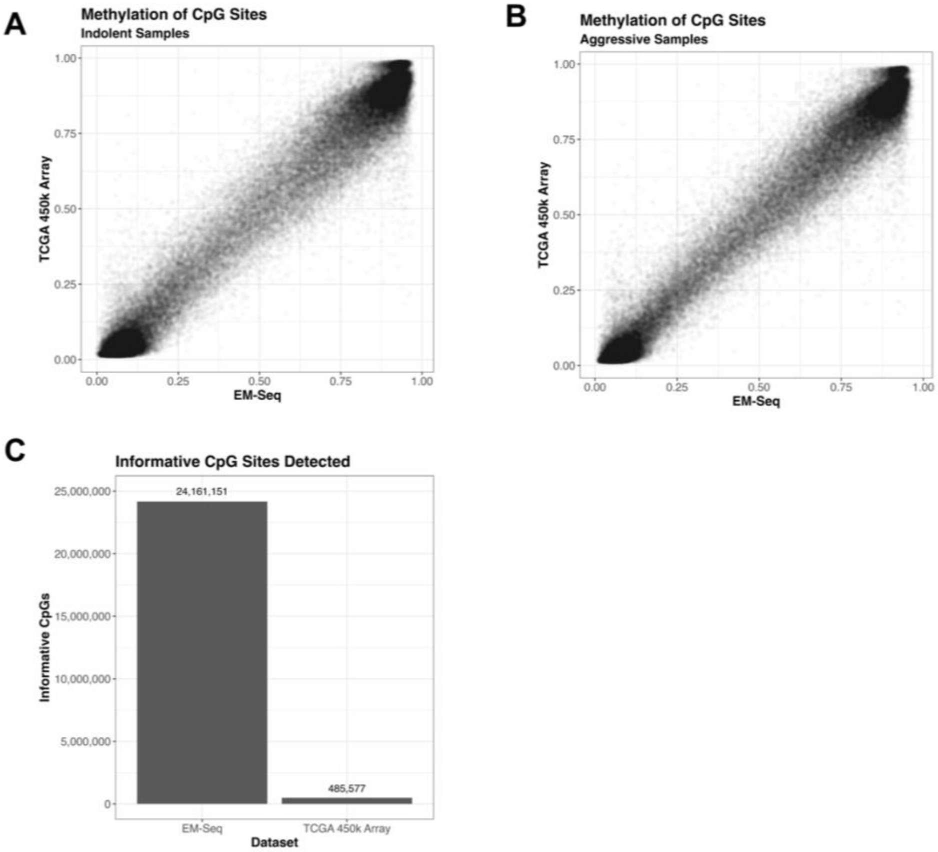


Figure 4-1. Comparison of EM-Seq and 450k array data. Indolent (A) and aggressive (B) samples from our cohort sequenced using EM-Seq and the TCGA cohort processed using the Illumina 450k array. C) Number of CpG sites with methylation data after filtering.

with a summed Gleason score of 6 in the TCGA cohort analyzed with the Illumina 450k array (Figure 4-1A). Comparisons were also made between our aggressive samples (Gleason 4-5) sequenced with EM-Seq and TCGA samples reported as having a summed Gleason score of 9 (Figure 4-1B). Although comparisons could only be made using CpGs found using both approaches, the overlapping CpGs were found to be highly concordant, with both comparisons yielding Pearson correlations >0.97. The need to rely on the summed total Gleason score of samples in the TCGA cohort highlights the ambiguity that is highly prevalent in most studies in this space, which hide the heterogeneity that may be present in tumors. Importantly, we also found that even after filtering CpGs with less than 10x coverage in our cohort, the EM-Seq samples contained approximately 50x more informative CpG sites than what was present in the array data (Figure 4-1C). This difference is clinically important, as many previous studies that have built classifiers for stratifying PCa patients use individual CpG sites as input, but they have been limited to those sites available on the 450k array. Based on these results, we concluded that while our EM-Seq data is comparable to published array-based studies, it contains a large amount of information that has been systematically ignored in most studies.

4.2.2 Unique methylation differences distinguish indolent from aggressive PCa

Having confirmed that our EM-Seq data was consistent with other methodologies, we next sought to further evaluate differences between indolent and aggressive PCa. Methylation patterns near transcript start sites (TSSs) were found to be tightly regulated, as expected, in both indolent and aggressive disease, with a general tendency toward hypomethylation in aggressive samples that becomes more prevalent in regions that are further removed from TSSs (Figure 4-2A). This is consistent with the general observation of global hypomethylation in cancer outside of tightly regulated regions[37]. While this appeared to be the more common global trend, gene-specific patterns were also identified. A total of 263,362 differentially methylated regions (DMRs) were identified based on 1kb bins tiled throughout the genome, with the overwhelming majority of statistically significant DMRs located near TSSs (Figure 4-2B). A representative example of these DMRs is shown in Figure 4-2C, representing the DMR located in the promoter region of *GSTP1*. *GSTP1* methylation has been well-studied in the context of localized PCa and our detection of a DMR in this promoter region served as a form of positive control.

To better understand the biological implications of the detected DMRs, differentially expressed genes (DEGs) were identified using the publicly available TCGA PRAD cohort. A total of 223 genes were found to be both differentially expressed and differentially methylated. Pathway enrichment analysis revealed that the most highly enriched InterPro terms for this gene set included Homeodomain, Homeodomain-like, and Homeobox, which have previously been implicated in PCa[29,121].

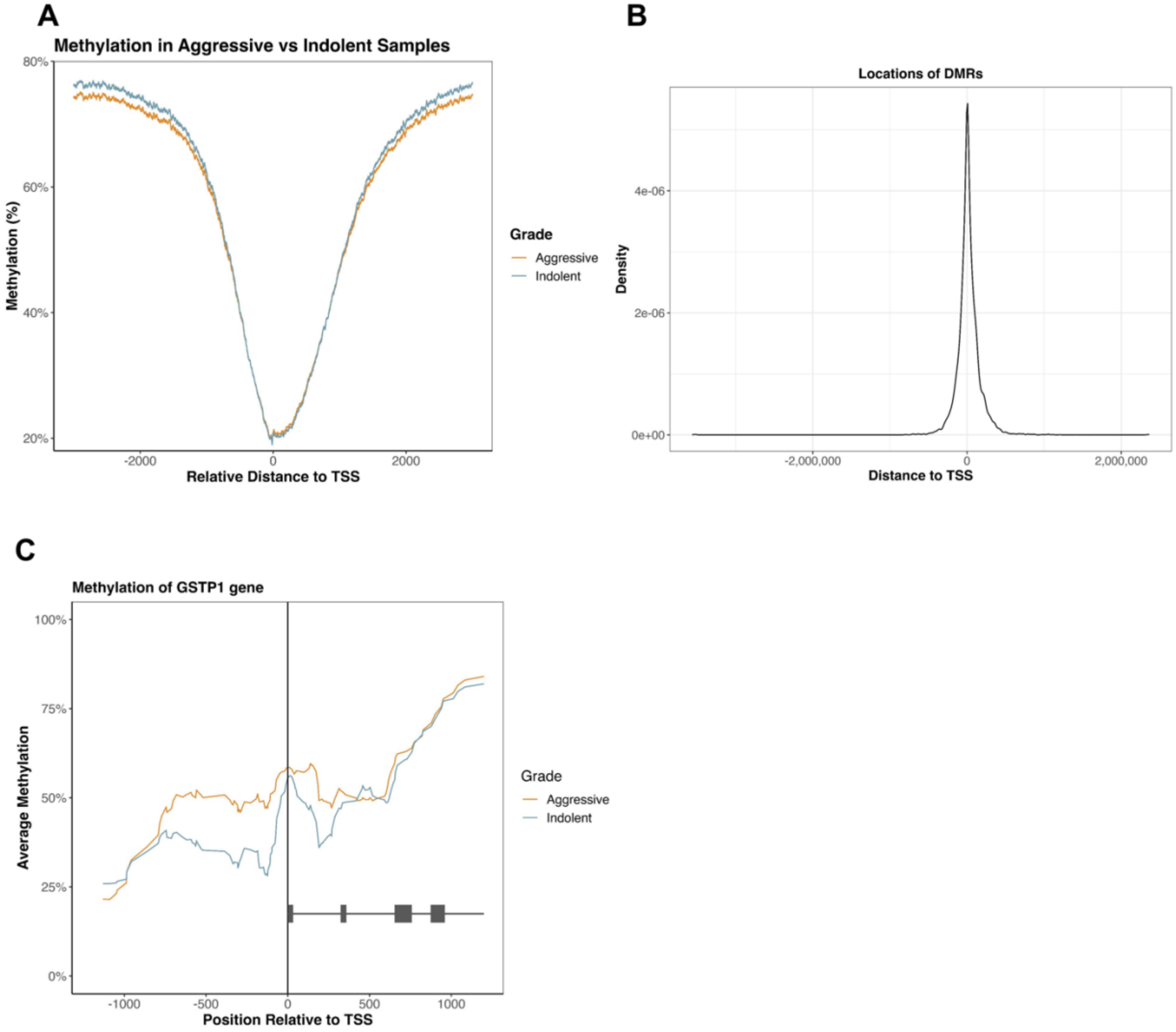


Figure 4-2. Methylation profiles of indolent and aggressive PCa. A) Rolling window of average methylation among indolent and aggressive samples near transcription start sites. B) Location of 1kb tiles found to be differentially methylated, with respect to transcription start sites. C) Rolling window of average methylation near the *GSTP1* gene in indolent and aggressive samples.

4.2.3 Classification of PCa severity using CpG methylation

Having confirmed that distinct methylation profiles distinguish between indolent and aggressive disease, we sought to develop a specific methylation signature for classifying tumor samples. Two different approaches for identifying differentially methylated CpG sites (DMSs) were attempted, both of which were limited to tumors with matched indolent and aggressive foci available to serve as a smaller, discovery cohort. The first approach used methylKit applied to all matched samples grouped together (n=5 patients), while the second approach used the union of DMSs identified by comparing each individual set of matched samples (n=5 patients).We found that 99% of DMSs found using the union approach were also found using the grouped pair approach (Figure 4-3A). Approximately 43% of DMSs found using the grouped approach were located in promoter regions, whereas 58% of sites using the union approach were located in promoter regions (Figure 4-3B). This suggests that the union approach, though yielding far fewer DMSs, was enriched for sites that may be more biologically important. We therefore focused on DMSs called using the union approach. A machine learning algorithm was then applied to identify 22 DMSs for use as an epigenetic signature. This signature, derived only from the CpGs identified using the matched samples, was then applied to all available samples. We found that it accurately grouped indolent and aggressive samples in all but two cases (Figure 4-3C). A full list of the identified CpGs can be found in Table 4-S1. Importantly, ~95% (21/22) of the identified

CpGs in our epigenetic signature are not found on the Illumina 450k methylation array, highlighting the importance of a whole genome sequencing approach.

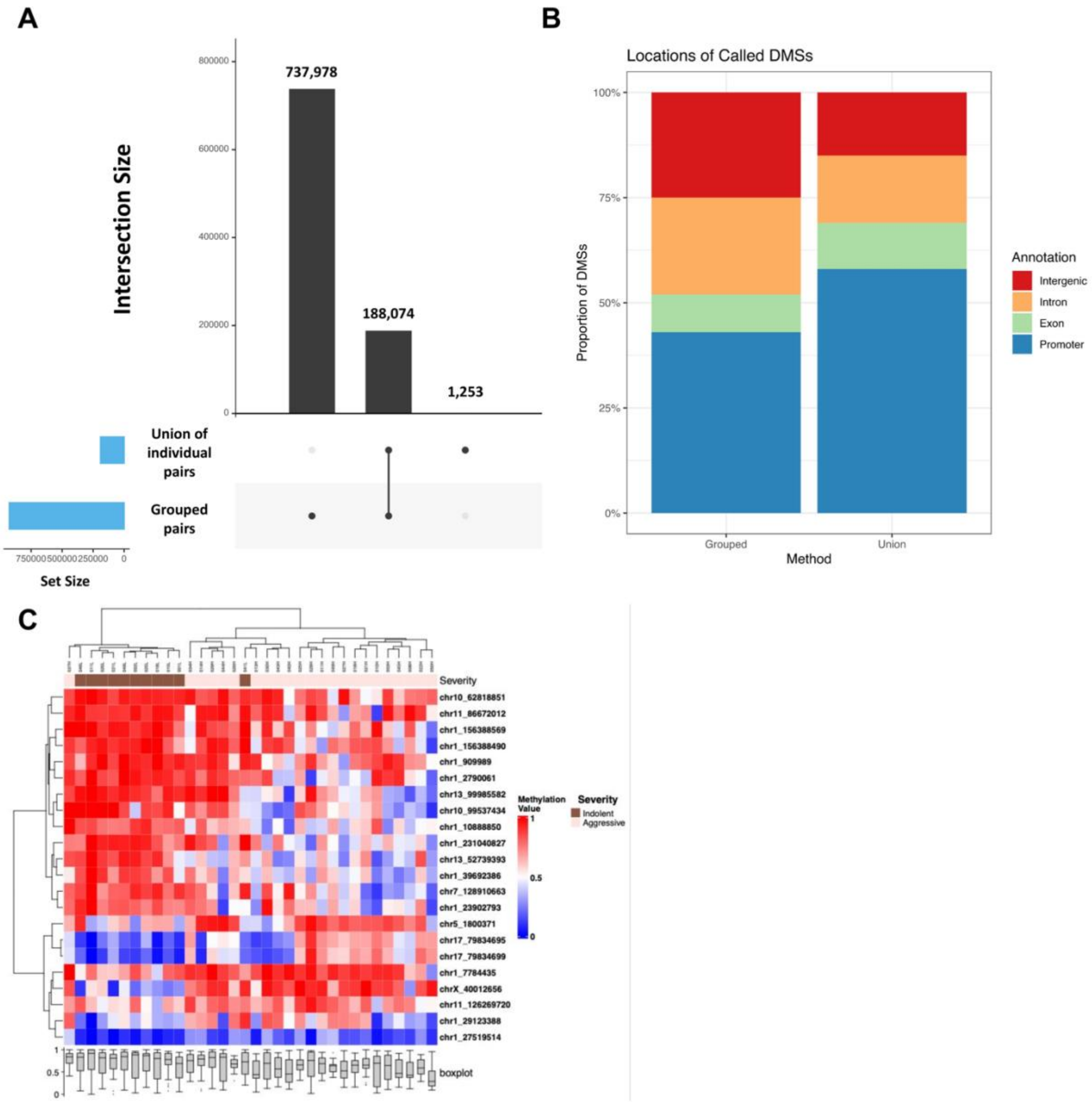


Figure 4-3. Epigenetic signature for PCa stratification. A) Upset plot showing the overlap between DMSs called using two different approaches. B) Annotation of DMSs called using two different approaches. C) Heatmap showing how indolent and aggressive samples cluster based on methylation observed at 22 CpGs.

4.2.4 Unique copy number alterations in aggressive disease

Copy number alteration (CNA) analysis also revealed distinct profiles in indolent and aggressive samples, with the average amount of the genome altered in aggressive samples being 2.4x greater than the amount of the genome altered in indolent samples (Figure 4-4A). PCa is known to have a large number of SVs and we were able to confirm that some of these may emerge during the transition between indolent and aggressive disease[27]. Interestingly, copy number gains were particularly prominent on chromosome 8 in aggressive samples, which includes the *MYC* oncogene, while being entirely absent in the indolent samples (Figure 4-4B). While tools exist for copy number analysis using 450k arrays, they are thought to have poor

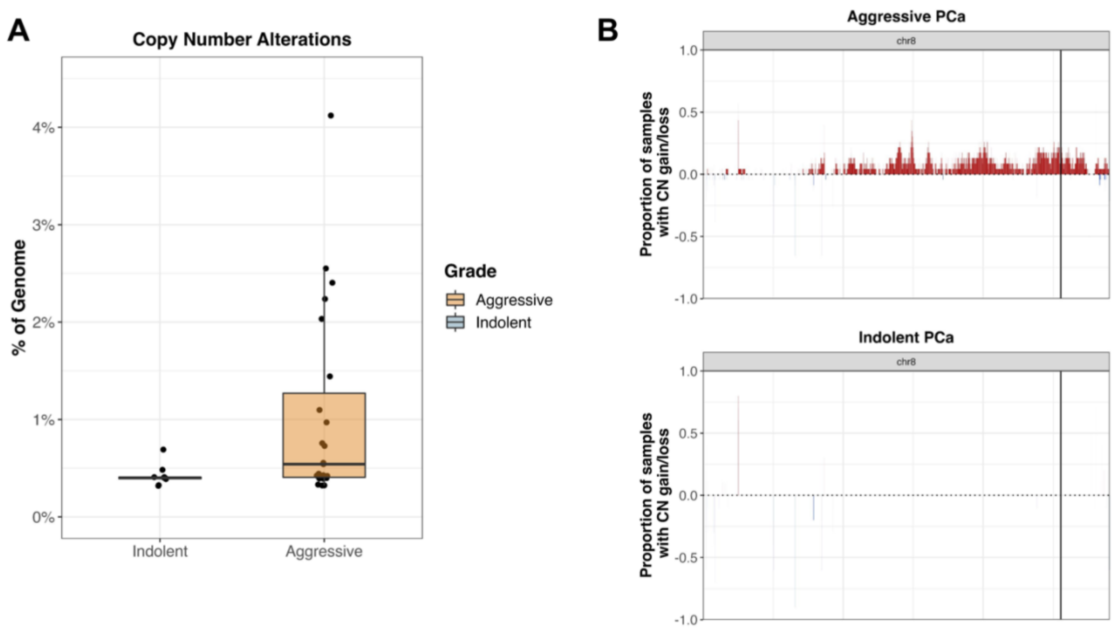


Figure 4-4. Copy number alterations in localized PCa. A) Total percentage of the genome affected by CNAs in each sample. B) Frequency of CNAs across chromosome 8 in the aggressive and indolent samples.

reliability compared to whole genome sequencing approaches, further highlighting the benefits of our whole genome sequencing approach[122].

4.3 Discussion

There is a clinical need for improved methods for accurate and non-invasive classification of indolent and aggressive PCa. Methylation profiling has been identified as a promising approach in this space, in part because of the few recurrent genetic mutations that exist in early PCa and because of the potential for future non-invasive assays. However, the majority of current studies comparing indolent and aggressive disease are limited in that they do not fully address the highly heterogeneous nature of PCa. Furthermore, most classification methods rely on specific differentially methylated CpGs found using the Illumina 450k methylation array, suggesting that there are large portions of the genome that are being systematically ignored.

In our exploratory analysis, we have found that whole genome methylation sequencing of specific portions of localized tumors is able to give important insights into the methylation profile of aggressive disease while observing nearly 50x more CpG sites than traditional array-based approaches, even after accounting for filtering of uninformative CpGs. Furthermore, our use of whole genome methylation sequencing allowed for reliable copy number analysis, as opposed to the limitations present when using array-based techniques. Using this approach, we identified 223 genes with associated DMRs that appear to influence gene expression and identified recurrent copy number gains near the *MYC* locus found only in aggressive tumor samples.

The development of our epigenetic signature, though effective in our small cohort, is limited. Calculated beta values for methylation at specific CpG sites is prone to noise, which is only exacerbated by our small cohort size. Validation through the use of larger cohorts is important. Alternatively, the development of a classifier using bulk data from a larger cohort and then applying that algorithm to our unique cohort may also yield interesting results. This approach would allow for evidence that an epigenetic signature developed using a highly heterogeneous cohort could accurately distinguish between indolent and aggressive sub-clones from within a single tumor. In either case, further work is required before further application of the epigenetic signature described here.

4.4 Methods

4.4.1 Sample collection and processing

Tumor samples were collected by clinicians using an IRB-approved protocol. Pathologists reviewed the collected samples and identified specific regions within each sample that corresponded to Gleason grade 3 or grade 4/5 morphology. Specified regions were collected using laser capture microdissection. All samples were processed using the NEBNext Enzymatic Methyl-seq Kit using the manufacturer's instructions and 20ng of material.

4.4.2 Copy number analysis

Copy number calls were calculated using default settings using CNVnator[59]. This was done using the CNVnator workflow described in McDonnell Genome Institute's publicly available pipelines found at https://github.com/genome/analysis-workflows. CNVnator was run using a bin size of 100bp and outputs were filtered requiring an e-value < 0.05 as reported by

CNVnator. Genome alteration percentage was calculated based on a genome size of 3 billion bp. Identification of recurrent alterations on chromosome 8 was performed using the GenVisR R library[123].

4.4.3 Methylation analysis

Quality control, sequence alignment and initial calculations for methylation values were performed using a public workflow made available by the McDonnell Genome Institute at https://github.com/genome/analysis-workflows. Briefly, raw reads were trimmed using flexbar and resulting reads were aligned using Biscuit[124,125]. Methylation data and quality control metrics were determined with the Biscuit pileup and Biscuit qc commands, respectively. Conversion efficiency was evaluated by repeating the pipeline and using pUC19 and Lambda reference genomes as input.

Differential methylation was then calculated using methylKit as previously described[126]. Filtering was performed using methylKit::filterByCoverage(low.count=10, lo.perc=NULL, hi.count=NULL, hi.perc=99.9) and subsequently normalized using methylKit::normalizeCoverage(method='median'). DMRs were identified by breaking the genome into 1kb tiles and then filtered, requiring an absolute methylation difference of 10% and q-value < 0.05.

4.4.4 Differential expression analysis

TCGA PRAD RNA-Seq expression data were accessed using the ExperimentHub R package and analyzed using DESeq2[127,128]. Samples with a summed Gleason score of 6 were labeled as "indolent" and samples with a summed Gleason score >= 8 were labeled as “aggressive”. DEGs were required to have an absolute log2() fold change >= 1 and an adjusted p-value < 0.05.

4.4.5 Pathway enrichment analysis

Genes that were both differentially expressed and differentially methylated were analyzed using DAVID’s functional annotation tool[129]. Comparisons were made against a background of all human genes. Pathway identification was performed using gene labels from InterPro[130].

4.4.6 CpG site selection for epigenetic signature

DMSs were identified using methylKit by comparing each individual set of matched samples (n=5 patients) as described in 4.4.3 and then finding the union of the results. Remaining DMSs with a q-value < 0.05 and located within a DMR (see 4.4.3). Elastic net regression was then performed 70 times, each time keeping the 50 top hits. After 70 iterations, the 500 most frequently identified CpGs were evaluated using xgboost to create an importance summary. A mixed effects logistic regression model was then applied to identify the optimal subset of remaining CpGs, resulting in 22 CpG sites.

4.5 Supplemental Materials

Table 4-S1. CpG sites included in epigenetic signature. All positions are based on hg19.

chr1:909989
chr1:2790061
chr1:7784435
chr1:10888850
chr1:23902793
chr1:27519514
chr1:29123388
chr1:39692386
chr1:156388569
chr1:156388490
chr1:231040827
chr5:1800371
chr7:128910663
chr10:62818851
chr10:99537434
chr11:86672012
chr11:126269720
chr13:52739393
chr13:99985582
chr17:79834695

chr17:79834699
chrX:40012656

Chapter 5: Summary, future directions and conclusions

5.1 Summary

This work sought to address computational needs associated with the detection of biomarkers in prostate cancer. We identified points in early and late-stage disease progression that might benefit the most from improved use of predictive biomarkers and developed tools that leveraged different biological aspects that were most relevant to the different disease states. Specifically, we used methylation profiling in early-stage disease due to the low number of recurrent genetic mutations at this stage, while using somatic mutation detection in late-stage cancer because of specific known mutations that associate with treatment resistance. Additionally, we determined that although recent evidence suggested that fcircRNAs may associate with cancer progression and a number of fcircRNAs had been detected in early prostate cancer, no systematic method existed for detecting these isoforms. Development of INTEGRATE-Circ and INTEGRATE-Vis provides the community with the tools needed to gain a better understanding of fcircRNAs and their potential as biomarkers. By developing software tools to address needs, we have provided tools to improve future work not only in prostate cancer, but in oncology as a whole.

5.2 Future Directions

While software development and benchmarking are critical for performing high quality and reproducible research, it is only in the application of such tools that the full potential of the software can be realized. The following areas of research have been identified as possible future directions for the utilization of the methods described here.

5.2.1 Validation of prognostic biomarkers in mCRPC

PACT, our pipeline for the analysis of circulating tumor DNA, presents the field with a reproducible way to identify somatic mutations from liquid biopsies. Our lab, in collaboration with Drs. Pachynski and Chaudhuri, previously published a prognostic liquid biopsy assay for detection of key variants in mCRPC[22]. PACT provides an effective way for additional studies to validate these previous findings and for others to confirm those findings. Indeed, ongoing collaborations with the labs of Drs. Pachynski and Chaudhuri has already resulted in the use of PACT to further investigate the association between *AR* alterations and clinical progression, including in pre-treatment patients[131]. As PACT is disease-agnostic, this tool has potential utility in other diseases as well. We anticipate that the adoption of PACT will improve the reproducibility and utility of future liquid biopsies in a variety of contexts.

5.2.2 Identification of fcircRNAs as possible biomarkers

Given the reported oncogenic nature of some fcircRNAs and the relative stability of circRNA in general, fcircRNAs represent an attractive clinical biomarker. However, due to the lack of detection methods, few fcircRNAs have been identified. With the creation of INTEGRATE-Circ, a large number of possible studies across different cancer types are now feasible. Although little is known about fcircRNAs, it is possible that studies that specifically look at cancer types with frequent structural variations (such as PCa) may identify recurrent fcircRNAs that could serve as biomarkers. We expect that the use of INTEGRATE-Circ and INTEGRATE-Vis will empower the field to identify and evaluate fcircRNAs as potential biomarkers in a wide variety of disease states.

5.2.3 Development of localized PCa prognostic assays

Pending the validation of our approach using additional cohorts, it is feasible that a methylation-based assay for detecting aggressive disease using liquid biopsies could be developed based on the biomarkers identified in this work. The use of non-invasive assays for early stage PCa patients would improve the efficiency of disease progression monitoring while reducing the inconvenience of more invasive methods. Future work should therefore seek to use larger cohorts to allow for more in-depth analyses and for validation of the exploratory findings presented here.

5.3 Conclusion

In summary, this work describes the development of bioinformatic tools to aid in the future detection and evaluation of predictive biomarkers. The methods described herein represent a multi-pronged approach using multiple biomarker classes (somatic mutations, methylation changes, RNA circularization and expression) based on the unique biological changes previously observed in specific stages of PCa progression. These tools may empower future work both in early-stage and late-stage PCa. Furthermore, through our development of PACT, INTEGRATE-Circ and INTEGRATE-Vis we have provided the cancer research field at large with novel tools to encourage future clinically relevant work in any cancer type, beyond PCa. As a result of this work, we have helped ensure the accuracy and reproducibility of future studies aiming to identify and evaluate cancer biomarkers.

References

1. Siegel RL, Giaquinto AN, Jemal A. Cancer statistics, 2024. *CA Cancer J Clin*. 2024;74(1):12-49. doi:10.3322/caac.21820

2. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA Cancer J Clin*. 2022;72(1):7-33. doi:10.3322/caac.21708

3. Ryan CJ, Smith MR, de Bono JS, et al. Abiraterone in metastatic prostate cancer without previous chemotherapy. *N Engl J Med*. 2013;368(2):138-148. doi:10.1056/NEJMoa1209096

4. Scher HI, Fizazi K, Saad F, et al. Increased survival with enzalutamide in prostate cancer after chemotherapy. *N Engl J Med*. 2012;367(13):1187-1197. doi:10.1056/NEJMoa1207506

5. African-American Prostate Cancer Disparities | Current Urology Reports. Accessed March 12, 2024. https://link.springer.com/article/10.1007/s11934-017-0724-5

6. Prostate Cancer Prognosis. Published November 6, 2023. Accessed March 12, 2024. https://www.hopkinsmedicine.org/health/conditions-and-diseases/prostate-cancer/prostate-cancer-prognosis

7. Rawla P. Epidemiology of Prostate Cancer. *World J Oncol*. 2019;10(2):63-89. doi:10.14740/wjon1191

8. Modern Active Surveillance in Prostate Cancer: A Narrative Review - ScienceDirect. Accessed March 12, 2024. https://www.sciencedirect.com/science/article/pii/S1558767322001938?via%3Dihub

9. Borza T, Konijeti R, Kibel AS. Early detection, PSA screening, and management of overdiagnosis. *Hematol Oncol Clin North Am*. 2013;27(6):1091-1110, vii. doi:10.1016/j.hoc.2013.08.002

10. Van Poppel H, Albreht T, Basu P, Hogenhout R, Collen S, Roobol M. Serum PSA-based early detection of prostate cancer in Europe and globally: past, present and future. *Nat Rev Urol*. 2022;19(9):562-572. doi:10.1038/s41585-022-00638-6

11. Screening for Prostate Cancer: US Preventive Services Task Force Recommendation Statement | Oncology | JAMA | JAMA Network. Accessed March 5, 2024. https://jamanetwork.com/journals/jama/fullarticle/2680553

12. Mottet N, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer—2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur Urol*. 2021;79(2):243-262. doi:10.1016/j.eururo.2020.09.042

13. Tikkinen KAO, Dahm P, Lytvyn L, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a clinical practice guideline. *BMJ*. 2018;362:k3581. doi:10.1136/bmj.k3581

14. Gleason Score - an overview | ScienceDirect Topics. Accessed April 7, 2022. https://www.sciencedirect.com/topics/medicine-and-dentistry/gleason-score

15. Klein EA, Cooperberg MR, Magi-Galluzzi C, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur Urol*. 2014;66(3):550-560. doi:10.1016/j.eururo.2014.05.004

16. True L, Coleman I, Hawley S, et al. A molecular correlate to the Gleason grading system for prostate adenocarcinoma. *Proc Natl Acad Sci*. 2006;103(29):10991-10996. doi:10.1073/pnas.0603678103

17. Hu JC, Tosoian JJ, Qi J, et al. Clinical Utility of Gene Expression Classifiers in Men With Newly Diagnosed Prostate Cancer. *JCO Precis Oncol*. 2018;(2):1-15. doi:10.1200/PO.18.00163

18. Viswanathan SR, Ha G, Hoff AM, et al. Structural alterations driving castration-resistant prostate cancer revealed by linked-read genome sequencing. *Cell*. 2018;174(2):433-447.e19. doi:10.1016/j.cell.2018.05.036

19. Luo H, Wei W, Ye Z, Zheng J, Xu RH. Liquid Biopsy of Methylation Biomarkers in Cell-Free DNA. *Trends Mol Med*. 2021;27(5):482-500. doi:10.1016/j.molmed.2020.12.011

20. Skvortsova K, Stirzaker C, Taberlay P. The DNA methylation landscape in cancer. *Essays Biochem*. 2019;63(6):797-811. doi:10.1042/EBC20190037

21. Henríquez I, Roach M, Morgan TM, et al. Current and Emerging Therapies for Metastatic Castration-Resistant Prostate Cancer (mCRPC). *Biomedicines*. 2021;9(9):1247. doi:10.3390/biomedicines9091247

22. Dang HX, Chauhan PS, Ellis H, et al. Cell-Free DNA Alterations in the AR Enhancer and Locus Predict Resistance to AR-Directed Therapy in Patients With Metastatic Prostate Cancer. *JCO Precis Oncol*. 2020;(4):680-713. doi:10.1200/PO.20.00047

23. Prostate Cancer, Version 2.2019, NCCN Clinical Practice Guidelines in Oncology in: Journal of the National Comprehensive Cancer Network Volume 17 Issue 5 (2019). Accessed April 7, 2022. https://jnccn.org/view/journals/jnccn/17/5/article-p479.xml?utm_campaign=JNCCN_TrendMD_1&UTM_medium=CPC&UTM_source=Trendmd&utm_source=TrendMD&utm_medium=cpc

24. Wan JCM, Massie C, Garcia-Corbacho J, et al. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nat Rev Cancer*. 2017;17(4):223-238. doi:10.1038/nrc.2017.7

25. AR-V7 and Resistance to Enzalutamide and Abiraterone in Prostate Cancer | NEJM. Accessed April 7, 2022. https://www.nejm.org/doi/full/10.1056/nejmoa1315815

26. Scher HI, Lu D, Schreiber NA, et al. Association of AR-V7 on Circulating Tumor Cells as a Treatment-Specific Biomarker With Outcomes and Survival in Castration-Resistant Prostate Cancer. *JAMA Oncol*. 2016;2(11):1441-1449. doi:10.1001/jamaoncol.2016.1828

27. Fraser M, Sabelnykova VY, Yamaguchi TN, et al. Genomic hallmarks of localized, non-indolent prostate cancer. *Nature*. 2017;541(7637):359-364. doi:10.1038/nature20788

28. van Dessel LF, van Riet J, Smits M, et al. The genomic landscape of metastatic castration-resistant prostate cancers reveals multiple distinct genotypes with potential clinical impact. *Nat Commun*. 2019;10(1):5251. doi:10.1038/s41467-019-13084-7

29. Nyberg T, Govindasami K, Leslie G, et al. Homeobox B13 G84E Mutation and Prostate Cancer Risk. *Eur Urol*. 2019;75(5):834-845. doi:10.1016/j.eururo.2018.11.015

30. Shah S, Rachmat R, Enyioma S, Ghose A, Revythis A, Boussios S. BRCA Mutations in Prostate Cancer: Assessment, Implications and Treatment Considerations. *Int J Mol Sci*. 2021;22(23):12628. doi:10.3390/ijms222312628

31. Spans L, Clinckemalie L, Helsen C, et al. The Genomic Landscape of Prostate Cancer. *Int J Mol Sci*. 2013;14(6):10822-10851. doi:10.3390/ijms140610822

32. García-Perdomo HA, Chaves MJ, Osorio JC, Sanchez A. Association between TMPRSS2:ERG fusion gene and the prostate cancer: systematic review and meta-analysis. *Cent Eur J Urol*. 2018;71(4):410-419. doi:10.5173/ceju.2018.1752

33. Huggins C, Hodges CV. Studies on prostatic cancer: I. The effect of castration, of estrogen and of androgen injection on serum phosphatases in metastatic carcinoma of the prostate. *CA Cancer J Clin*. 1972;22(4):232-240. doi:10.3322/canjclin.22.4.232

34. Takeda DY, Spisák S, Seo JH, et al. A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. *Cell*. 2018;174(2):422-432.e13. doi:10.1016/j.cell.2018.05.037

35. Quigley DA, Dang HX, Zhao SG, et al. Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell*. 2018;174(3):758-769.e9. doi:10.1016/j.cell.2018.06.039

36. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. *Nucleic Acids Res*. 2013;41(16):e155. doi:10.1093/nar/gkt599

37. Ehrlich M. DNA methylation in cancer: too much, but also too little. *Oncogene*. 2002;21(35):5400-5413. doi:10.1038/sj.onc.1205651

38. Hong J, Rhee JK. Genomic Effect of DNA Methylation on Gene Expression in Colorectal Cancer. *Biology*. 2022;11(10):1388. doi:10.3390/biology11101388

39. Lee WH, Morton RA, Epstein JI, et al. Cytidine methylation of regulatory sequences near the pi-class glutathione S-transferase gene accompanies human prostatic carcinogenesis. *Proc Natl Acad Sci*. 1994;91(24):11733-11737. doi:10.1073/pnas.91.24.11733

40. Barry KH, Mohanty K, Erickson PA, et al. MYC DNA Methylation in Prostate Tumor Tissue Is Associated with Gleason Score. *Genes*. 2020;12(1):E12. doi:10.3390/genes12010012

41. Annala M, Vandekerkhove G, Khalaf D, et al. Circulating Tumor DNA Genomics Correlate with Resistance to Abiraterone and Enzalutamide in Prostate Cancer. *Cancer Discov*. 2018;8(4):444-457. doi:10.1158/2159-8290.CD-17-0937

42. Jiang N, Meng X, Mi H, et al. Circulating lncRNA XLOC_009167 serves as a diagnostic biomarker to predict lung cancer. *Clin Chim Acta Int J Clin Chem*. 2018;486:26-33. doi:10.1016/j.cca.2018.07.026

43. Wang C, Liu H. Factors influencing degradation kinetics of mRNAs and half-lives of microRNAs, circRNAs, lncRNAs in blood in vitro using quantitative PCR. *Sci Rep*. 2022;12(1):7259. doi:10.1038/s41598-022-11339-w

44. Zhang H da, Jiang LH, Sun DW, Hou JC, Ji ZL. CircRNA: a novel type of biomarker for cancer. *Breast Cancer Tokyo Jpn*. 2018;25(1):1-7. doi:10.1007/s12282-017-0793-9

45. Tan S, Gou Q, Pu W, et al. Circular RNA F-circEA produced from EML4-ALK fusion gene as a novel liquid biopsy biomarker for non-small cell lung cancer. *Cell Res*. 2018;28(6):693-695. doi:10.1038/s41422-018-0033-7

46. El Messaoudi S, Mouliere F, Du Manoir S, et al. Circulating DNA as a Strong Multimarker Prognostic Tool for Metastatic Colorectal Cancer Patient Management Care. *Clin Cancer Res*. 2016;22(12):3067-3077. doi:10.1158/1078-0432.CCR-15-0297

47. Rich TA, Reckamp KL, Chae YK, et al. Analysis of Cell-Free DNA from 32,989 Advanced Cancers Reveals Novel Co-occurring Activating RET Alterations and Oncogenic Signaling Pathway Aberrations. *Clin Cancer Res Off J Am Assoc Cancer Res*. 2019;25(19):5832-5842. doi:10.1158/1078-0432.CCR-18-4049

48. De Laere B, van Dam PJ, Whitington T, et al. Comprehensive Profiling of the Androgen Receptor in Liquid Biopsies from Castration-resistant Prostate Cancer Reveals Novel Intra-AR Structural Variation and Splice Variant Expression Patterns. *Eur Urol*. 2017;72(2):192-200. doi:10.1016/j.eururo.2017.01.011

49. Vandekerkhove G, Struss WJ, Annala M, et al. Circulating Tumor DNA Abundance and Potential Utility in De Novo Metastatic Prostate Cancer. *Eur Urol*. 2019;75(4):667-675. doi:10.1016/j.eururo.2018.12.042

50. Ho SS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nat Rev Genet*. 2020;21(3):171-189. doi:10.1038/s41576-019-0180-9

51. Overview of Structural Variation. Accessed April 7, 2022. https://www.ncbi.nlm.nih.gov/dbvar/content/overview/

52. Li Y, Roberts ND, Wala JA, et al. Patterns of somatic structural variation in human cancer genomes. *Nature*. 2020;578(7793):112-121. doi:10.1038/s41586-019-1913-9

53. Nickless A, Zhang J, Othoum G, et al. Pan-Cancer Analysis Reveals Recurrent BCAR4 Gene Fusions across Solid Tumors. *Mol Cancer Res MCR*. 2022;20(10):1481-1488. doi:10.1158/1541-7786.MCR-21-0775

54. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol*. 2014;15(6):R84. doi:10.1186/gb-2014-15-6-r84

55. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinforma Oxf Engl*. 2012;28(18):i333-i339. doi:10.1093/bioinformatics/bts378

56. Chen X, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinforma Oxf Engl*. 2016;32(8):1220-1222. doi:10.1093/bioinformatics/btv710

57. van Belzen IAEM, Schönhuth A, Kemmeren P, Hehir-Kwa JY. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *Npj Precis Oncol*. 2021;5(1):1-11. doi:10.1038/s41698-021-00155-6

58. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016;12(4):e1004873. doi:10.1371/journal.pcbi.1004873

59. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res*. 2011;21(6):974-984. doi:10.1101/gr.114876.110

60. Doebley AL, Ko M, Liao H, et al. A framework for clinical cancer subtyping from nucleosome profiling of cell-free DNA. *Nat Commun*. 2022;13(1):7475. doi:10.1038/s41467-022-35076-w

61. Gawroński AR, Lin YY, McConeghy B, et al. Structural variation and fusion detection using targeted sequencing data from circulating cell free DNA. *Nucleic Acids Res*. 2019;47(7):e38. doi:10.1093/nar/gkz067

62. Newman AM, Bratman SV, Stehr H, et al. FACTERA: a practical method for the discovery of genomic rearrangements at breakpoint resolution. *Bioinforma Oxf Engl*. 2014;30(23):3390-3393. doi:10.1093/bioinformatics/btu549

63. Liu H, Yin H, Li G, Li J, Wang X. Aperture: alignment-free detection of structural variations and viral integrations in circulating tumor DNA. *Brief Bioinform*. 2021;22(6):bbab290. doi:10.1093/bib/bbab290

64. Wang Z, Wang Y, Zhang J, et al. Significance of the TMPRSS2:ERG gene fusion in prostate cancer. *Mol Med Rep*. 2017;16(4):5450-5458. doi:10.3892/mmr.2017.7281

65. Wang E, Aifantis I. RNA Splicing and Cancer. *Trends Cancer*. 2020;6(8):631-644. doi:10.1016/j.trecan.2020.04.011

66. Chen LL, Yang L. Regulation of circRNA biogenesis. *RNA Biol*. 2015;12(4):381-388. doi:10.1080/15476286.2015.1020271

67. Holdt LM, Kohlmaier A, Teupser D. Circular RNAs as Therapeutic Agents and Targets. *Front Physiol*. 2018;9:1262. doi:10.3389/fphys.2018.01262

68. Li Z, Huang C, Bao C, et al. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol*. 2015;22(3):256-264. doi:10.1038/nsmb.2959

69. Hansen TB, Jensen TI, Clausen BH, et al. Natural RNA circles function as efficient microRNA sponges. *Nature*. 2013;495(7441):384-388. doi:10.1038/nature11993

70. Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature*. 2013;495(7441):333-338. doi:10.1038/nature11928

71. Huang A, Zheng H, Wu Z, Chen M, Huang Y. Circular RNA-protein interactions: functions, mechanisms, and identification. *Theranostics*. 2020;10(8):3503-3517. doi:10.7150/thno.42174

72. Othoum G, Coonrod E, Zhao S, Dang HX, Maher CA. Pan-cancer proteogenomic analysis reveals long and circular noncoding RNAs encoding peptides. *NAR Cancer*. 2020;2(3):zcaa015. doi:10.1093/narcan/zcaa015

73. Visci G, Tolomeo D, Agostini A, Traversa D, Macchia G, Storlazzi CT. CircRNAs and Fusion-circRNAs in cancer: New players in an old game. *Cell Signal*. 2020;75:109747. doi:10.1016/j.cellsig.2020.109747

74. Vo JN, Cieslik M, Zhang Y, et al. The Landscape of Circular RNA in Cancer. *Cell*. 2019;176(4):869-881.e13. doi:10.1016/j.cell.2018.12.021

75. Tan S, Sun D, Pu W, et al. Circular RNA F-circEA-2a derived from EML4-ALK fusion gene promotes cell migration and invasion in non-small cell lung cancer. *Mol Cancer*. 2018;17(1):138. doi:10.1186/s12943-018-0887-9

76. Tan Y, Huang Z, Wang X, Dai H, Jiang G, Feng W. A novel fusion circular RNA F-circBA1 derived from the *BCR-ABL* fusion gene displayed an oncogenic role in chronic myeloid leukemia cells. *Bioengineered*. 2021;12(1):4816-4827. doi:10.1080/21655979.2021.1957749

77. Pan Y, Lou J, Wang H, et al. CircBA9.3 supports the survival of leukaemic cells by up-regulating c-ABL1 or BCR-ABL1 protein levels. *Blood Cells Mol Dis*. 2018;73:38-44. doi:10.1016/j.bcmd.2018.09.002

78. Wang J, Ma HL, Liu WR, Peng Y, Zhou JK, Yang JL. CircBA1 derived from BCR-ABL fusion gene inhibits cell proliferation in chronic myeloid leukemia. *Cancer Commun Lond Engl*. 2021;41(1):79-82. doi:10.1002/cac2.12120

79. Chen S, Huang V, Xu X, et al. Widespread and Functional RNA Circularization in Localized Prostate Cancer. *Cell*. 2019;176(4):831-843.e22. doi:10.1016/j.cell.2019.01.025

80. You X, Conrad TO. Acfs: accurate circRNA identification and quantification from RNA-Seq data. *Sci Rep*. 2016;6:38820. doi:10.1038/srep38820

81. Cai Z, Xue H, Xu Y, et al. Fcirc: A comprehensive pipeline for the exploration of fusion linear and circular RNAs. *GigaScience*. 2020;9(6):giaa054. doi:10.1093/gigascience/giaa054

82. Dal Molin A, Tretti Parenzan C, Gaffo E, et al. Discovery of fusion circular RNAs in leukemia with *KMT2A::AFF1* rearrangements by the new software CircFusion. *Brief Bioinform*. 2023;24(1):bbac589. doi:10.1093/bib/bbac589

83. Corcoran RB, Chabner BA. Application of Cell-free DNA Analysis to Cancer Treatment. *N Engl J Med*. 2018;379(18):1754-1765. doi:10.1056/NEJMra1706174

84. Dang HX, Krasnick BA, White BS, et al. The clonal evolution of metastatic colorectal cancer. *Sci Adv*. 2020;6(24):eaay9691. doi:10.1126/sciadv.aay9691

85. Jeffares DC, Jolly C, Hoti M, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun*. 2017;8:14061. doi:10.1038/ncomms14061

86. Larson D, abelhj, Chiang C, et al. hall-lab/svtools: svtools v0.5.1. Published online September 12, 2019. doi:10.5281/zenodo.3406745

87. Chiang C, Layer RM, Faust GG, et al. SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods*. 2015;12(10):966-968. doi:10.1038/nmeth.3505

88. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*. 2012;6(2):80-92. doi:10.4161/fly.19695

89. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31(3):213-219. doi:10.1038/nbt.2514

90. Kim S, Scheffler K, Halpern AL, et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat Methods*. 2018;15(8):591-594. doi:10.1038/s41592-018-0051-x

91. Koboldt DC, Chen K, Wylie T, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinforma Oxf Engl*. 2009;25(17):2283-2285. doi:10.1093/bioinformatics/btp373

92. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*. 2009;25(21):2865-2871. doi:10.1093/bioinformatics/btp394

93. Poplin R, Ruano-Rubio V, DePristo MA, et al. *Scaling Accurate Genetic Variant Discovery to Tens of Thousands of Samples*. Genomics; 2017. doi:10.1101/201178

94. Ainscough BJ, Griffith M, Coffman AC, et al. DoCM: a database of curated mutations in cancer. *Nat Methods*. 2016;13(10):806-807. doi:10.1038/nmeth.4000

95. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. *Bioinformatics*. 2015;31(13):2202-2204. doi:10.1093/bioinformatics/btv112

96. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

97. Khanna A, Larson D, Srivatsan S, et al. Bam-readcount - rapid generation of basepair-resolution sequence metrics. *J Open Source Softw*. 2022;7(69):3722. doi:10.21105/joss.03722

98. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Published online 2013. doi:10.48550/ARXIV.1303.3997

99. Broad Institute. Picard Toolkit. Published online 2019. https://github.com/broadinstitute/picard

100. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015;31(12):2032-2034. doi:10.1093/bioinformatics/btv098

101. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008. doi:10.1093/gigascience/giab008

102. Zhang J, White NM, Schmidt HK, et al. INTEGRATE: gene fusion discovery using whole genome and transcriptome data. *Genome Res*. 2016;26(1):108-118. doi:10.1101/gr.186114.114

103. FulcrumGenomics. fgibio. Published online 2022. https://github.com/fulcrumgenomics/fgbio

104. Chen LL, Bindereif A, Bozzoni I, et al. A guide to naming eukaryotic circular RNAs. *Nat Cell Biol*. 2023;25(1):1-5. doi:10.1038/s41556-022-01066-9

105. Soda M, Choi YL, Enomoto M, et al. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature*. 2007;448(7153):561-566. doi:10.1038/nature05945

106. Drilon A. TRK inhibitors in TRK fusion-positive cancers. *Ann Oncol Off J Eur Soc Med Oncol*. 2019;30(Suppl_8):viii23-viii30. doi:10.1093/annonc/mdz282

107. Braun TP, Eide CA, Druker BJ. Response and Resistance to BCR-ABL1-Targeted Therapies. *Cancer Cell*. 2020;37(4):530-542. doi:10.1016/j.ccell.2020.03.006

108. Azatyan A, Zaphiropoulos PG. Circular and Fusion RNAs in Medulloblastoma Development. *Cancers*. 2022;14(13):3134. doi:10.3390/cancers14133134

109. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl*. 2013;29(1):15-21. doi:10.1093/bioinformatics/bts635

110. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015

111. Panda A, Gorospe M. Detection and Analysis of Circular RNAs by RT-PCR. *BIO-Protoc*. 2018;8(6). doi:10.21769/BioProtoc.2775

112. Zhang J, Gao T, Maher CA. INTEGRATE-Vis: a tool for comprehensive gene fusion visualization. *Sci Rep*. 2017;7(1):17808. doi:10.1038/s41598-017-18257-2

113. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinforma Oxf Engl*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033

114. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res*. 2019;47(8):e47. doi:10.1093/nar/gkz114

115. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*. 2010;26(7):873-881. doi:10.1093/bioinformatics/btq057

116. Cancer of the Prostate - Cancer Stat Facts. SEER. Accessed April 14, 2022. https://seer.cancer.gov/statfacts/html/prost.html

117. GSTP1 Methylation and Protein Expression in Prostate Cancer: Diagnostic Implications - PMC. Accessed March 11, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4995330/

118. Toth R. Random forest-based modelling to detect biomarkers for prostate cancer progression. Published online 2019:15.

119. Han Y, Zheleznyakova GY, Marincevic-Zuniga Y, et al. Comparison of EM-seq and PBAT methylome library methods for low-input DNA. *Epigenetics*. 2021;0(0):1-10. doi:10.1080/15592294.2021.1997406

120. Vaisvila R, Ponnaluri VKC, Sun Z, et al. Enzymatic methyl sequencing detects DNA methylation at single-base resolution from picograms of DNA. *Genome Res*. 2021;31(7):1280-1289. doi:10.1101/gr.266551.120

121. Goel S, Bhatia V, Kundu S, et al. Transcriptional network involving ERG and AR orchestrates Distal-less homeobox-1 mediated prostate cancer progression. *Nat Commun*. 2021;12(1):5325. doi:10.1038/s41467-021-25623-2

122. Critical Evaluation of Copy Number Variant Calling Methods Using DNA Methylation - PMC. Accessed March 12, 2024. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7028453/

123. Skidmore ZL, Wagner AH, Lesurf R, et al. GenVisR: Genomic Visualizations in R. *Bioinforma Oxf Engl*. 2016;32(19):3012-3014. doi:10.1093/bioinformatics/btw325

124. Dodt M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR-Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. *Biology*. 2012;1(3):895-905. doi:10.3390/biology1030895

125. Zhou W. biscuit-0.1.3. Published online March 24, 2016. doi:10.5281/zenodo.48262

126. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol*. 2012;13(10):R87. doi:10.1186/gb-2012-13-10-r87

127. ExperimentHub. Bioconductor. Accessed March 12, 2024. http://bioconductor.org/packages/ExperimentHub/

128. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

129. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*. 2022;50(W1):W216-W221. doi:10.1093/nar/gkac194

130. InterPro in 2022 | Nucleic Acids Research | Oxford Academic. Accessed March 12, 2024. https://academic.oup.com/nar/article/51/D1/D418/6814474

131. Chauhan PS, Alahi I, Sinha S, et al. Genomic and epigenomic analysis of plasma cell-free DNA identifies stemness features associated with worse survival in AR-altered lethal prostate cancer. Published online December 1, 2023:2023.12.01.23299215. doi:10.1101/2023.12.01.23299215