Longitudinal structural variant phylogenies define tumor evolution under therapeutic selection pressure in metastatic prostate cancer

Yunzhou Liu[1], Jiaying Lai[1], Mark C. Markowski[2], Emmanuel S. Antonarakis[2,6], Angelo M. De Marzo[2,7], Srinivasan Yegnasubramanian[2,7], Laura D. Wood[2,3,4], Laura A Sena[2*], Rachel Karchin[1,2,5,*]

[1]Department of Biomedical Engineering, Institute of Computational Medicine, Johns Hopkins University, Baltimore, MD, USA; [2]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [3]Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA; [4]Department of Medicine, Division of Gastroenterology and Hepatology, Johns Hopkins University, Baltimore, MD, USA; [5]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA; [6]Department of Medicine, University of Minnesota, Masonic Cancer Center, Minneapolis, MN, USA; [7]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD, USA

Abstract

Therapeutic pressure shapes tumor evolution by selecting for subclones with distinct genomic architectures that drive resistance. In metastatic castration-resistant prostate cancer (mCRPC), structural variants (SVs) are major drivers of this process but are rarely incorporated into clonal reconstruction from bulk DNA sequencing data. Here we describe SVCFit, a computational framework for estimating the cellular fraction of diverse classes of structural variants from whole-genome sequencing and using these estimates to reconstruct clonal evolutionary relationships. SVCFit explicitly models SV-type–specific breakpoint and breakend patterns and accounts for local copy-number context, enabling accurate estimation of SV cellular fraction across heterogeneous tumor genomes. Using simulated datasets and in silico mixtures of metastatic prostate cancer samples, we show that SVCFit achieves consistently lower estimation error than the current state-of-the-art approaches. We then apply SVCFit to longitudinal whole-genome sequencing data from patients with mCRPC treated with bipolar androgen therapy (supraphysiologic testosterone). Structural variant–based clonal phylogenies reveal marked treatment-associated clonal reconfiguration, including contraction of highly rearranged subclones and expansion of resistant populations defined by distinct structural alterations. Together, these results demonstrate that integrating structural variants into clonal evolutionary analysis provides critical insight into tumor evolution under therapy. SVCFit enables reconstruction of SV-defined clonal architecture from routine whole-genome sequencing, expanding the toolkit for studying cancer evolution in precision oncology.

Contact: karchin@jhu.edu, laura.sena@jhmi.edu

Introduction

Intratumor heterogeneity encompasses a broad spectrum of molecular alterations within a single tumor, including single nucleotide variants (SNVs), structural variants (SVs), copy number variants (CNVs), protein expression differences, and epigenetic modifications. Genomic instability and selective pressures can drive the emergence of diverse subpopulations, or subclones, that differ in their aggressiveness, metastatic potential, and resistance to therapy[1,2]. Reconstructing clonal architecture can reveal critical events that shape cellular phenotypes and underlying mechanisms[3]. Early research into cancer evolution focused primarily on aneuploidy[4], and modern bulk-sequencing methods continue to leverage copy-number and allelic-imbalance signals to infer tumor purity, ploidy, and subclonal mixture proportions. The advent of Next Generation Sequencing (NGS) enabled scientists to investigate genetic alterations at the single-base level, using SNVs as markers for cancer progression[5]. More recently, paired-end sequencing has made the detection of structural variants more feasible and cost-effective, allowing their integration into studies of cancer evolution[6,7]. For both SNVs and SVs, estimating the cellular fraction, the proportion of cells that harbor a specific alteration is key to inferring their relative evolutionary ordering during tumor development.

Structural variants contribute significantly to genomic diversity and genome instability[8]. SVs come in various forms, such as insertions, deletions, duplications, inversions, and translocations. Throughout, we use “duplication” to mean a tandem duplication unless otherwise specified. These large-scale changes can disrupt gene function by shifting or removing nucleotide sequences, which can lead to dysfunction in gene regulation or abnormalities in protein structure[9]. Because of their significance in shaping cancer genomes, SVs are critical to studying cancer genomics and metastasis, and they have direct relevance to therapeutic outcomes. Prostate cancer is increasingly recognized as a disease driven not only by small DNA sequence alterations but also by the presence and evolution of SVs. In primary prostate tumors, hallmark events such as the *TMPRSS2–ERG* fusion and losses of tumor suppressors like *PTEN* may arise through chromosomal rearrangements[10–13]. In metastatic castration-resistant prostate cancer (mCRPC), whole-genome sequencing has revealed that these structural alterations become even more pervasive, with complex events and noncoding rearrangements at loci such as *AR*, *MYC*, and *FOXA1* dominating the genomic landscape[14,15]. Under the selective pressure of androgen deprivation and modern *AR*-targeted therapies, clones carrying *AR* amplification, enhancer co-amplification, or other *AR* locus rearrangements are positively selected, while alternative trajectories involving SV-mediated loss of *TP53*/*RB1* and lineage plasticity give rise to *AR*-indifferent phenotypes. Understanding and monitoring this SV evolution is therefore important to predicting treatment response

and prognosis, designing rational therapeutic combinations, and ultimately improving outcomes for men with metastatic prostate cancer.

Somatic alterations like SNVs and SVs in the human genome can be studied through bulk DNA sequencing, which is widely used due to its affordability. Incorporating somatic SVs detected from bulk sequencing into tumor evolution analysis can be done by first estimating variant allele frequency (VAF) at a specific locus. VAF is often used in clonal reconstruction, but the process is noisy, due to confounding factors like aneuploidy and tumor impurity, both common in cancer samples[16,17]. CNA-focused methods such as HATCHet2[18] infer clone-specific copy-number states and mixture proportions from bulk sequencing, but they do not directly estimate cellular fractions for individual SVs across SV classes and copy-number contexts. A complementary, variant-centric strategy is to convert read support into cellular fraction, explicitly accounting for tumor purity and local copy-number state. While this method has been widely applied to SNVs[19–21], to our knowledge, only one published method, SVclone[22], estimates the cellular fraction of SVs from bulk sequencing data. VAF is calculated as the ratio of altered reads to total reads at a genomic position, which works well for SNVs but does not transfer well to SVs. SNVs are detected at a single genomic position, while SVs are detected across multiple genomic positions. SVs also encompass various types, each affecting the genome differently, leading to distinct patterns of altered and reference reads.

Despite their biological importance, SVs are typically excluded from tumor phylogenies inferred from bulk sequencing, largely due to challenges in estimating their cellular prevalence across diverse SV classes and copy-number contexts. Addressing this gap is critical for reconstructing tumor evolution under therapy. Here, we introduce SVCFit, a computational framework for estimating the cellular fraction of structural variants and use these estimates to reconstruct clonal evolutionary relationships from longitudinal whole-genome sequenced samples[23].

Results

SVCFit improves SV cellular fraction estimation compared to state-of-the-art methods

Evaluation on simulated data: While the state-of-the-art method SVclone[22] provides a widely used framework for estimating SV cellular fraction (SVCF) from breakpoint read evidence, SVCFit extends this idea by leveraging SV-type–specific breakpoint/breakend

patterns (and copy-number context) to enable consistent estimation across diverse SV classes. We benchmarked SVCFit against SVclone, utilizing simulated genomes to assess the performance of SVCF estimation over a range of SV types, zygosities, temporal orderings of overlapping events, clonal architectures, and tumor purities. Accuracy was quantified as the absolute difference between estimated and true SV cellular fractions (SVCFs). Overall, SVCFit demonstrated consistently lower absolute error, achieving consistently lower mean estimation errors than SVclone across all SV types (translocation, deletion, duplication, inversion) and clonality states (Fig. 1A). Translocations refer to inter-chromosomal rearrangements unless otherwise specified. SVCFit maintained robust performance at SV tumor purities of 80% and 40% for deletions, inversions, and breakends; however, duplications exhibited a purity-dependent effect, with increased mean estimation error at lower purity (40%). Furthermore, we analyzed the impact of zygosity on estimation precision. While both methods exhibited comparable error distributions for heterozygous SVs, SVCFit significantly outperformed SVclone in resolving homozygous SVs. Fig. 1B shows that SVclone[22] displayed high error variance for homozygous variants, whereas SVCFit maintained a tight error distribution near zero. A comprehensive breakdown of these results, stratified by specific temporal orderings zygosity, SV type and clonal configurations, is provided in Supplementary Fig. S4.

Evaluation on in-silico mixtures: Next, we applied an in silico mixing benchmark developed by the SVclone authors, using two previously sequenced metastatic prostate tumor samples with known purities (49% and 46%). Both metastatic samples contain a single clone with shared ancestry[23]. Mixing the two source samples at varying ratios produced nine three-clone mixtures with controlled SV cellular fractions and clonal architecture. Additional complexity was achieved by generating four- and five-clone tumors through differential subsampling of odd and even chromosomes, as specified in the SVclone framework[22]. The mean error was calculated by the difference between the ground truth and the estimated cellular fraction. SVCFit achieved a substantially lower cumulative mean error across all 11 mixtures compared to SVclone (Total mean error: 0.33 vs. 1.07). Fig. 1C illustrates that SVclone exhibited a systematic positive bias (overestimation), which remained constant regardless of mixing proportion. In contrast, SVCFit estimation consistently aligned closer to ground truth. While SVCFit displayed minor sensitivity to sample imbalance, with mean error slightly elevated at extreme mixing ratios (e.g., 10:90), its estimation errors remained lower than SVclone across every tested configuration.

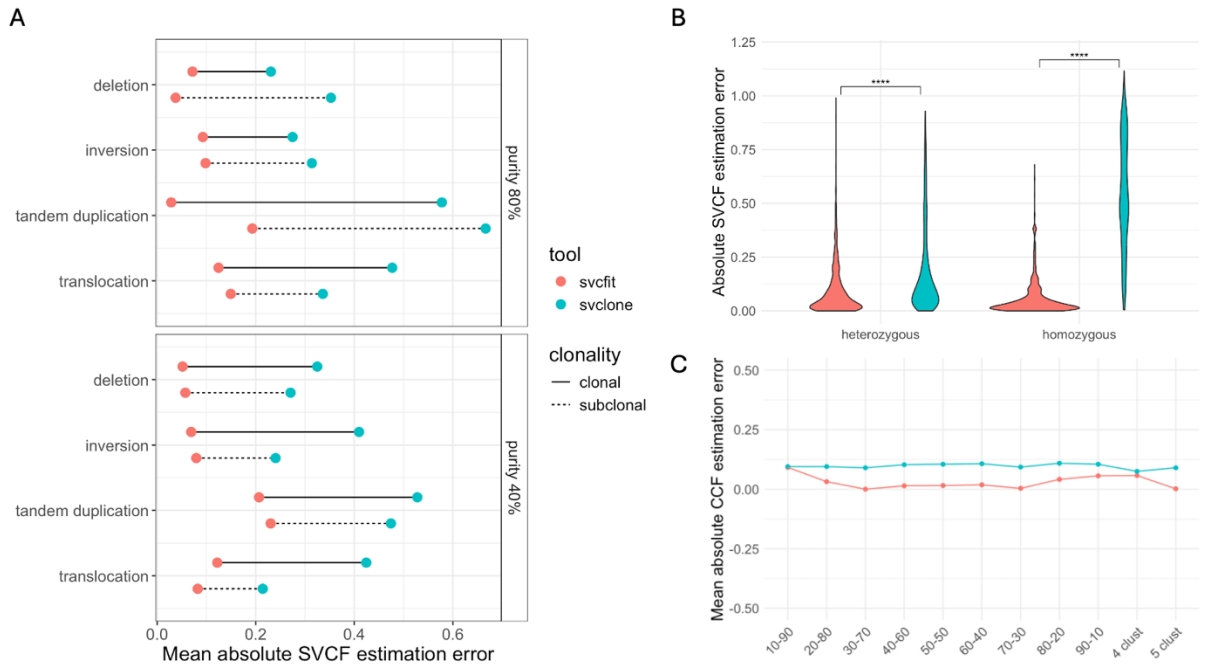

Figure 1. Benchmarking of SVCFit versus SVclone on simulated and prostate cancer mixture datasets. (A) Mean absolute structural variant cellular fraction (SVCF) estimation error derived from VISOR simulations, stratified by structural variant type (translocation, deletion, duplication, inversion) and SV tumor purity (80% top, 40% bottom). Performance is compared between SVCFit (red) and SVclone (blue) for both clonal (solid lines) and subclonal (dashed lines) variants. Error is defined as the absolute difference between estimated and true SVCF. (B) Distribution of absolute SVCF estimation error in VISOR simulations, stratified by variant zygosity (heterozygous vs. homozygous). SVCFit exhibits significantly lower error than SVclone in both settings (paired Wilcoxon signed-rank test, two-sided, $p < 0.0001$). Statistical significance is indicated as follows: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***), $p < 0.0001$ (****). (C) Mean absolute cancer cell fraction (CCF) estimation error evaluated on prostate cancer synthetic mixture data across varying mixture proportions and cluster configurations.

Longitudinal analysis of SV clones in mCRPC subjects receiving bipolar androgen therapy

To demonstrate the clinical utility of SVCFit in resolving tumor evolution, we analyzed longitudinal whole-genome sequencing (WGS) data from two subjects with metastatic castration-resistant prostate cancer (mCRPC) treated with bipolar androgen therapy (BAT)[24,25], a form of supraphysiologic testosterone treatment, on the COMBAT trial[26]. BAT is the cyclical use of high dose testosterone as treatment for CRPC, which is associated with a 30% response rate, but fairly rapid acquired resistance with a median progression-free survival of 6 months[27]. Subjects on the COMBAT study were required to undergo paired sequential metastatic tumor biopsies at baseline (pre-treatment) and

after 3 months of BAT treatment. Two subjects, 10 and 31, were notable because they initially demonstrated a biochemical response (assessed by serum prostate specific antigen [PSA]) levels) to BAT, followed by a subsequent PSA rise at the time of the second biopsy, suggesting early development of acquired therapy resistance under selection pressures (Fig 2). Therefore, we applied SVCFit to estimate the cellular fractions of somatic SVs, followed by Dirichlet Process Gaussian Mixture Models (DP-GMM) to cluster SVs and a modified Gabow-Myers algorithm to reconstruct the evolutionary tree topologies and relative ordering of SV-defined clones of these cancers to assess how tumor evolution may have contributed to acquired therapy resistance (Methods).

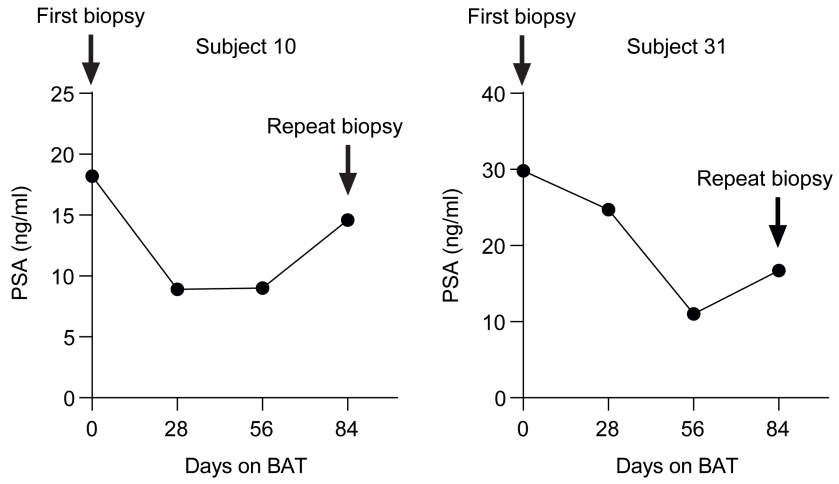

Figure 2. PSA dynamics during bipolar androgen therapy (BAT) and timing of repeat biopsy. Serum PSA (ng/mL) is plotted over time (days on BAT) for two representative subjects (Subject 10 and Subject 31) at baseline and during treatment (days 0, 28, 56, and 84). Both subjects show an initial decline in PSA (therapeutic response) followed by a subsequent PSA rise by week 12 (secondary resistance). Arrows indicate the time point at which the repeat biopsy was performed (day 84).

In Subject 10, SVCFit resolved three distinct clonal populations, consistent with a branched evolutionary architecture (Fig. 3A). Throughout, edge annotations indicate clusters of alterations with similar SVCF, whereas node labels/colors indicate clones defined by the cumulative clusters along each root-to-node path; pie charts report clone fractions per sample. Unless otherwise noted, annotated “deletions” and “duplications” are defined from SV breakpoint evidence rather than read-depth copy-number calls; hemizygous deletion is annotated only when supported by copy-number evidence (CN=1), and intragenic duplication/deletion only when breakpoints fall within the gene body.

A truncal clone (Clone 1), present in all tumor cells at both time points, harbored canonical prostate cancer alterations, including *TP53* deletion, *CHEK2* intragenic

deletion, *MDM2* duplication, and the *TMPRSS2–ERG* fusion. Following BAT, we observed a marked clonal reconfiguration, characterized by the disappearance of Clone 2 and the expansion of Clone 3 (Fig. 3B, 3C).

Clone 2 was notable for extensive genome-wide rearrangement, with SV cluster 2 containing the most translocations (n=21) and duplications (n=269), together with major structural alterations in DNA repair genes, including hemizygous deletions of *RAD51B*, and a deletion of *PALB2*. In contrast, Clone 3 was defined by an upstream duplication involving *EZH2*, consistent with increased dosage of this driver oncogene. The duplication mapped to a promoter-proximal regulatory region containing MYC and E2F1 binding sites. The expansion of Clone 3 under therapy illustrates how SVCFit can nominate candidate resistance-associated drivers from structural events that may be overlooked by SNV-centric analyses. The phylogeny reflects relative ordering of SV-defined clones constrained by pre-and on-treatment sampling, enabling direct comparison of clonal composition before and during therapy.

Longitudinal analysis on Subject 10 further revealed a striking remodeling of the genomic landscape following BAT. The pre-BAT (baseline) biopsy exhibited a hyper-rearranged phenotype indicative of pronounced genomic instability, with 30 translocations and 297 duplications (Fig. 3D). In contrast, the on-BAT sample showed a 7.5-fold reduction in translocations and a 5-fold reduction in duplications (4 translocations, 56 duplications; Fig. 3E). To determine whether these differences could be explained by tumor purity in the on-BAT sample (0.68 pre-BAT, 0.36 on-BAT), we performed *in silico* downsampling of the pre-BAT DNA sequencing data to match the effective tumor depth of the on-BAT biopsy. Here, tumor purity refers to conventional tumor cellularity estimated by FACETS. This analysis recapitulated the observed fold reductions: 5.5-fold for translocations and 2.5-fold for duplications (pre-BAT: 22 translocations, 143 duplications, on-BAT: 2 translocations, 56 duplications) (Supplementary Fig. S1), supporting that the genomic shifts represent a biological response to therapy, likely driven by the elimination of highly rearranged subclones.

Remarkably, a similar pattern of branched evolution and clonal replacement was observed in Subject 31 (Fig. 3F). This tumor was founded by a truncal clone (Clone 2) characterized by complex genomic duplications involving *MYC*, *FOXA1*, *NCOA2*, and *RAD21*, together with *RB1* duplication and *CHD1* hemizygous deletion. BAT treatment coincided with the divergence of two subclonal populations. Clone 3, which harbored *BRCA2*, *RB1,* and *ATM* deletion, contracted following therapy (Fig. 3G, 3H). Conversely, Clone 1 expanded to become the dominant population and was defined by a characteristic set of structural alterations, including *NCOA2* duplication, and multiple *ATM* inversions predicted to be inactivating. In each inversion, the right breakpoint mapped within the *ATM* gene body, truncating the coding sequence. Together, these observations demonstrate that mCRPC tumors can evolve through the selective

expansion or contraction of SV-defined subclones. By estimating the cellular prevalence of these events, SVCFit enables inference of clonal evolutionary relationships, distinguishing truncal drivers (e.g., *TP53* deletion, *MYC* duplication) from later-arising alterations associated with therapeutic sensitivity (e.g., *BRCA2* deletion) or resistance (e.g., *ATM* inactivating inversion).

Consistent with these findings, the on-BAT sample from Subject 31 exhibited an eight-fold reduction in translocations, all of which were confined to Clone 3, alongside a modest increase in duplications (pre-BAT: 8 translocations, 16 duplications; on-BAT: 1 translocation, 18 duplications; Fig. 3I, 3J). As both samples had comparable tumor purity (~0.6), no downsampling was required. Collectively, these cases suggest that BAT can selectively eliminate highly rearranged subclones, while leaving the dominant genomic architecture of resistant disease largely intact.

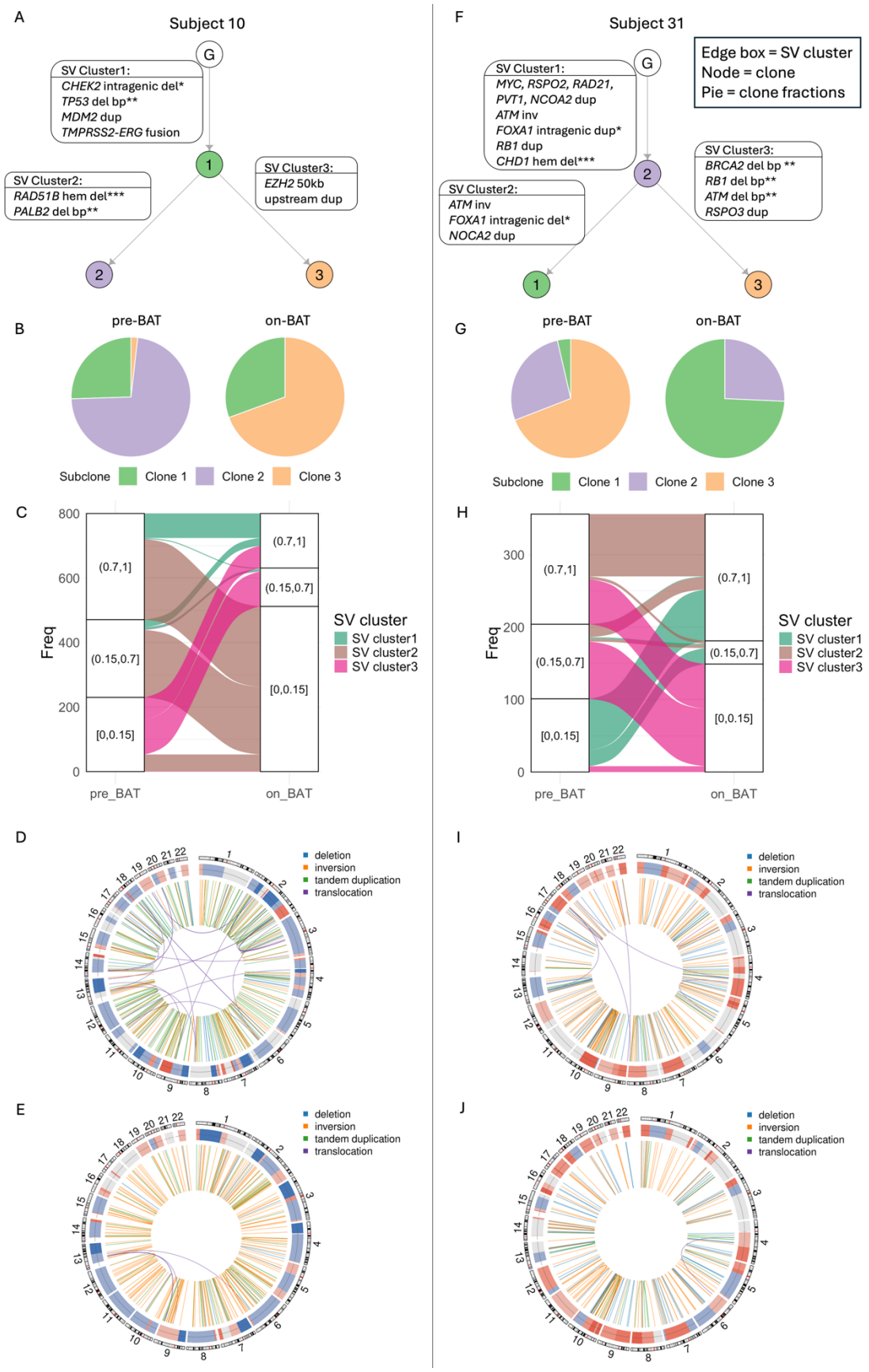
A
Subject 10
G
SV Cluster1:
CHEK2 intragenic del*
TP53 del bp**
MDM2 dup
TMPRSS2-ERG fusion
1
SV Cluster2:
RAD51B hem del***
PALB2 del bp**
SV Cluster3:
EZH2 50kb
upstream dup
2
3
B
pre-BAT
on-BAT
Subclone
Clone 1
Clone 2
Clone 3
C
Freq
800
600
400
200
0
(0.7,1]
(0.15,0.7]
[0,0.15]
pre_BAT
on_BAT
SV cluster
SV cluster1
SV cluster2
SV cluster3
D
deletion
inversion
tandem duplication
translocation
E
F
Subject 31
Edge box = SV cluster
Node = clone
Pie = clone fractions
SV Cluster1:
MYC, RSPO2, RAD21,
PVT1, NCOA2 dup
ATM inv
FOXA1 intragenic dup*
RB1 dup
CHD1 hem del***
SV Cluster2:
ATM inv
FOXA1 intragenic del*
NOCA2 dup
SV Cluster3:
BRCA2 del bp **
RB1 del bp**
ATM del bp**
RSPO3 dup
G
H
300
200
100
I
J

Figure 3. Longitudinal genomic landscape and tumor evolution in COMBAT mCRPC subjects treated with bipolar androgen therapy (BAT). Two representative subjects are shown (left: Subject 10; right: Subject 31), illustrating subclonal dynamics between the pre-BAT and on-BAT biopsies. (A,F) Phylogenetic trees inferred by SVCFit. Boxes annotate event clusters (groups of alterations with similar inferred structural variant cellular fraction (SVCF)) mapped to tree edges; clones correspond to nodes and represent the cumulative set of clusters along the path from the germline root (G) to each node. Key driver events are annotated on the corresponding edges. Unless otherwise noted, annotated “deletions” and “duplications” are defined from SV breakpoint evidence rather than read-depth copy-number calls; hemizygous deletion is annotated only when supported by copy-number evidence (CN=1), and intragenic deletion only when breakpoints fall within the gene body. Only SVs that overlap with known prostate cancer driver events are annotated on the tree. Edge lengths indicate inferred evolutionary relationships and ordering rather than elapsed time. (B,G) Pie charts show the estimated clone proportions in each sample (not per-cluster prevalence), with colors matching node labels. (C,H) Alluvial plots show changes in SVCF for each event cluster between pre-BAT and on-BAT, highlighting expansion or contraction of the corresponding lineages under therapy; colors correspond to the cluster associated with each lineage. (D,E,I,J) Circos plots summarize genome-wide structural variation (Subject 10: D pre-BAT, E on-BAT; Subject 31: I pre-BAT, J on-BAT). The outer ring shows chromosomes, the inner track shows copy-number state (gain, loss, neutral), and links denote rearrangements colored by type: deletions, duplications, inversions, and translocations. *Intragenic del/dup indicates an SV breakpoint-defined event affecting only part of the gene; **del bp indicates a SV breakpoint–defined deletion (copy-number state not implied); ***hem del indicates a hemizygous deletion supported by copy-number evidence (CN=1).

Discussion

While clonal reconstruction in cancer genomics has historically relied on single nucleotide variants for clonal reconstruction[28], this study demonstrates that genomic structural variants are not merely "co-passengers" but may be primary drivers of clonal architecture and evolution. By explicitly modeling the SV breakpoint patterns and the copy-number context, SVCFit provides a level of precision in cellular fraction estimation that has been difficult to achieve with bulk sequencing-based approaches. Benchmarking against SVclone shows that failure to account for these structural features can bias clonality estimates, potentially misclassifying truncal drivers as subclonal or obscuring emerging resistant populations.

The application of SVCFit to the COMBAT trial patients offers a high-resolution view of how metastatic castration-resistant prostate cancer (mCRPC) navigates therapeutic bottlenecks. In Subject 31, *FOXA1* showed a truncal focal 5′ amplification (CN ~3 pre-BAT; ~5 on-BAT), while a dominant resistant subclone acquired a small intragenic deletion affecting exon 2. This suggests that therapeutic pressure during BAT may

select for *FOXA1* locus remodeling with potential consequences for androgen receptor cistrome regulation. Routine gene-level copy-number reporting would likely summarize *FOXA1* as a single dosage state, masking subclonal gene-body alterations that emerge in resistant lineages. Furthermore, the expansion of Clone 3 in Subject 10, characterized by an *EZH2* upstream duplication, illustrates how SVCFit can pinpoint non-coding rearrangements that drive resistance through enhancer hijacking, events often missed by SNV-focused pipelines. Interestingly, in both of these cases of early acquired resistance to BAT, we observed either eradication of clones with SV of *BRCA2* (Subject 31) or SV of *PALB2* and *RAD51B* (Subject 10). While homologous recombination repair deficiency (HRD) has been proposed to be a predictive biomarker of response to BAT[29,30], we previously reported that it was not associated with durable response in the COMBAT study[31]. Application of SVCFit to these cases suggests that HRD may in fact be a valid predictive biomarker of response to BAT, but these responses can be limited by tumor HR status heterogeneity enabling outgrowth of resistant HR-proficient clones. Notably, subjects enrolled on the COMBAT trial had progressed on more prior therapies than those on previous trials of BAT and therefore may have had a higher degree of intra- and intertumoral heterogeneity and plasticity enabling acquired therapy resistance by clonal selection to be more pervasive.

We acknowledge the rising utility of single-cell sequencing in resolving tumor heterogeneity, however, technical limitations such as allelic dropout and high costs currently restrict its use in routine clinical practice[32,33]. SVCFit could bridge this gap by extracting "single-cell-like" insights from more affordable bulk WGS. This scalability makes SVCFit a viable candidate for integration into molecular tumor boards, where it might help clinicians distinguish stable truncal alterations from emerging resistant subclones.

SVCFit depends on the sensitivity of upstream SV callers. While the framework performs robustly for simple and moderately complex rearrangements, highly complex processes such as chromoplexy, chromothripsis, and breakage–fusion–bridge cycles generate multi-breakpoint architectures that are not yet explicitly represented in the current model, even when the breakpoints that make up these events are detectable with short-read sequencing. Future extensions will address this by incorporating structure-aware representations of complex SVs and by integrating SNVs, CNVs, and SVs within a unified evolutionary model, with long-read sequencing providing additional breakpoint resolution where available.

In conclusion, we demonstrate that SVCFit enables robust reconstruction of SV-defined clonal architecture from genomic data. Although shown here in metastatic prostate cancer, SVCFit is broadly applicable to other malignancies in which structural variation plays a central role. By enabling reconstruction of SV-defined clonal phylogenies from routine bulk sequencing, this framework supports more comprehensive analysis of

tumor evolution and reduces the risk that clinically consequential genomic alterations are overlooked in precision oncology.

Methods

Structural variant cellular fraction estimation

We estimate the cellular fraction of each structural variant (SV) using SVCFit, which derives SV cellular fraction (SVCF) from breakpoint and breakend read counts obtained from whole-genome sequencing. SVCFit accounts for SV type, local copy-number state, and allele-specific copy number, and integrates phasing information to adjust read counts in non-diploid regions. When SVs overlap copy-number alterations (CNVs), SVCFit applies an iterative optimization procedure that converges to closed-form solutions under all modeled configurations. These solutions encode the relative ordering of overlapping SV and CNV events. Formal derivations, robustness analyses, and convergence guarantees under all modeled SV–CNV configurations are provided in the Supplementary Information.

Datasets

Synthetic benchmark sets (VISOR). We used the Haplotype-Aware Structural Variants Simulator VISOR[34] (v1.1.3) to generate multi-clonal synthetic tumor genomes (with one matched normal). Somatic structural variants (SVs) with known structural-variant cellular fractions (SVCFs) and copy-number variants (CNVs) were spiked into tumor genomes to explicitly model SV–CNV overlaps in both cis and trans configurations. To optimize compute, simulations were restricted to chromosomes 1 and 2, spanning a range of SV/CNV overlap contexts, tumor purities, and subclonal compositions.

For each scenario, we specified a phylogeny with a germline root and a clonal tumor ancestor that bifurcates into two terminal subclones. SVs were simulated on chromosome 1, ordered from the p- to q-arm, and partitioned into three sets assigned to the truncal clone and two subclones (33/33/34 events, respectively). Chromosome 2 was used only as a partner for inter-chromosomal translocations. Restricting the simulation to these chromosomes simplified the controlled design of SV–CNV overlap patterns across tumor purities and subclonal mixtures.

Next, we simulated five SV–CNV configurations: (i) SV without CNV; (ii) SV preceding cis amplification; (iii) SV preceding trans amplification; (iv) SV following cis amplification;

and (v) SV following trans deletion. Each configuration was crossed with four SV tumor-purity levels (20%, 40%, 60%, 80%; Eq. 1) and three subclonal mixture settings (10:90, 30:70, 50:50), yielding 60 scenarios at 50x mean coverage (one BAM per scenario). For each spiked-in SV we retained start/end positions, SV type, and reference/alternate read counts in both tumor and normal BAMs, enabling direct comparison with SVCFit and SVclone.

$$SVTP = \frac{N_{sv}}{N} \ (Eq. 1)$$

where $SVTP$ is SV tumor purity, N_{sv} is the number of cells containing any somatic SV, and N is the total number of cells in the sample.

Tumor purity terminology. In this study, we distinguish between two related but distinct concepts of purity. *SV tumor purity (SVTP)* refers to the fraction of cells in a sample harboring somatic structural variants and is the primary quantity modeled and estimated by SVCFit. In contrast, *tumor purity* (or *sample purity*) refers to the fraction of tumor cells in a sample, as estimated by copy-number–based methods such as FACETS. Unless otherwise specified, references to “SV tumor purity” explicitly denote SVTP, whereas references to “tumor purity” denote conventional tumor cellularity.

Prostate metastasis mixtures (benchmark set). For direct comparison with SVclone, we generated eleven multi-clonal tumors by mixing BAM files from two single-clone metastatic prostate cancer samples. These samples were originally sequenced in a landmark study of subclonal structure in prostate cancer metastasis[23] and were subsequently used as a key benchmark in the SVclone paper[22]. Raw tumor and matched-normal FASTQs were obtained from EGA (EGAD00001001343)[23] and aligned to hs37d5 with BWA mem[35] (v0.7.19) to generate a single-clone tumor BAM. Then, each single-clone tumor BAM was downsampled in 10% increments of clonal frequency and mixed to create BAMs of heterogeneous samples with known SVCFs and known clonal composition. Nine mixtures yielded three-clone tumors (clone 1: SVs shared by both tumors; clone 2: shared + private SVs from tumor A; clone 3: shared + private SVs from tumor B). Four- and five-clone mixtures were generated by subsampling odd and even chromosomes from each metastasis at different rates prior to mixing, producing multiple subclonal populations with distinct cellular fractions, as described in Cmero et al[22]. Because SV tumor purity, mixture proportions, and per-SV clonality are known, the true SVCF can be calculated for every SV. For clonal SVs, true SVCF equals the mixture’s SV tumor purity (SVTP); for subclonal SVs, true SVCF equals the source tumor’s SVTP times its mixing proportion. SVs were called with Manta[36] (v1.6.0) and genotyped with SVtyper[37] (v0.7.1). For direct comparison with SVclone, SVCF was converted to cancer cell fraction (CCF) via Eq. 2.

$$CCF = \frac{SVCF}{SVTP} \ (Eq. 2)$$

We regenerated the mixture BAM files using the same script described in the SVclone paper (available at: https://github.com/mcmero/SVclone_Rmarkdown/make_insilico_mixtures.sh). We also contacted the first and senior authors to request the original mixture BAM files, but did not receive a response.

mCRPC subject samples (COMBAT). The COMBAT trial was a phase 2, multi-center, single arm clinical trial of BAT (testosterone cypionate 400 mg intramuscularly every 28 days) for 3 months followed by BAT in combination with nivolumab (480 mg intravenously every 28 days) until radiographic or clinical progression for patients with metastatic castration-resistant prostate cancer[26]. Participants must have received at least one prior androgen receptor pathway inhibitor and were permitted to have received up to one taxane chemotherapy agent prior to enrollment. All participants provided written informed consent prior to enrollment. Response to treatment was assessed by PSA measurement and clinical assessment every 28 days and computed tomography and technitium-99m bone scan every 3 months. The study was approved by the institutional review board at Johns Hopkins University, was registered on ClinicalTrials.gov (NCT03554317), and was conducted in accordance with the criteria set by the Declaration of Helsinki. Clinical results of this trial were previously published[26]. Participants on this trial underwent a biopsy of a soft tissue metastasis during screening (baseline sample), and the same site was re-biopsied following 3 months of BAT monotherapy (prior to the addition of nivolumab) for correlative studies. Flash-frozen tissue biopsies underwent laser-capture microdissection to isolate regions of cancer from normal tissue. Paired-end WGS was then performed on a NovaSeq 6000 S4 (2×150 bp; ~110 Gb per sample). For each biopsy and matched PBMCs normal sample, FASTQ files were trimmed with trim_galore[38] (v0.6.1) to remove adapters, aligned to GRCh38 with BWA mem[35] (v0.7.19), and processed with GATK4[39] (MarkDuplicates, BaseRecalibrator, ApplyBQSR; v4.6.2.0).

Data processing

Somatic structural variant calling. For both benchmark datasets, somatic SVs were called with Manta[36] (v1.6.0) using tumor–normal pairs with default parameters. For COMBAT, somatic SVs were called using Manta (v1.6.0), Delly[40] (v1.5.0), and GRIDSS[41] (v2.13.2) with default parameters, each provided with the tumor BAM and its matched normal. SVtyper[37] (v0.7.1) was then used to genotype SV breakpoints and collect reference/alternate supporting reads from the output VCF files of SV callers. Because SVtyper expects a LUMPY-formatted VCF with INFO fields MATEID, END,

CIPOS, CIEND, SVTYPE, we manually appended CIPOS and CIEND to VCF outputs from GRIDSS, Delly, and Manta to ensure compatibility. Finally, for COMBAT, all genotyped VCF files were harmonized using SURVIVOR[42] (v1.0.7) to produce a consensus VCF prior to cellular fraction analysis.

Rescue of a key driver event. Using default parameters, SVtyper detected the canonical prostate cancer driver fusion *TMPRSS2–ERG* in the pre-BAT biopsy of COMBAT Subject 10 (tumor purity 68%; breakpoint coverage 95), but not in the on-BAT biopsy, which had substantially lower tumor purity and reduced read depth at the breakpoint (purity 36.6%; breakpoint coverage 31). Given that *TMPRSS2–ERG* is a well-established early event in prostate tumorigenesis and is rarely lost, we manually inspected the on-BAT BAM file and identified a single read supporting the TMPRSS2–ERG fusion (deletion at chr21:38,499,931–41,501,273). While this read alone is insufficient to confidently call the fusion, ERG protein expression was positive by immunohistochemistry in both the pre- and on-BAT biopsies, a pattern strongly suggestive of an underlying *ERG* rearrangement, most commonly *TMPRSS2–ERG*. Together with its robust detection by WGS in the pre-BAT sample, these orthogonal data support *TMPRSS2–ERG* as a clonal event, and we therefore placed it on the trunk of the inferred phylogeny.

Germline heterozygous SNP detection and phasing candidates. Germline SNPs were identified from matched normal BAMs using GATK4[39] HaplotypeCaller (v4.6.2.0; default parameters) and filtered to heterozygous sites using bcftools[43] (v1.20). For SV phasing, we retained heterozygous SNPs within 500 bp of SV breakpoints. In SVCFit, these phased hetSNPs are used to assign breakpoint-supporting reads to the altered versus reference haplotype in non-diploid regions, improving SVCF estimation when allele-specific copy number is imbalanced (Supplementary Note S2). To test overlap with SV breakpoint reads, tumor BAMs were filtered to breakpoint-supporting reads using samtools[44] (v1.21; -f 1 -F 2), and bcftools[43] was used to generate SNP pileups from the filtered BAMs.

Somatic copy-number variant detection. CNVs overlapping SVs were called with FACETS[45] (v0.6.2). FACETS first generates a SNP pileup file using tumor and matched-normal BAMs and SNPs detected from GATK4[39] HaplotypeCaller (v4.6.2.0; default parameters); then coverage and the germline B-allele frequencies (BAFs) were computed from the SNPs pileup to detect CNV. FACETS SNP pileups were generated with: -g -q15 -Q20 -P100 -r25,0.

SVCF comparison and benchmarking

SVCFit requires inputs including: SV positions and types, SV read counts (SVtyper), CNV data (FACETS), and germline heterozygous SNP pileups (GATK4 and bcftools). SVCF per SV was computed using Eq. S1 (Supplementary Notes).

SVclone (v1.1.2) was obtained from https://github.com/mcmero/SVclone on March 10, 2024. SVclone was run with default parameters using as input the SV genomic coordinates, VISOR-specified tumor purity and ploidy from the simulations, and the tumor BAM file.

Benchmarking on VISOR simulations. For each of the 60 VISOR simulation scenarios, structural variants (SVs) were identified using Manta and subsequently genotyped with SVtyper to obtain reference and alternate supporting read counts. Heterozygous single-nucleotide variants (SNVs) were detected using GATK4 HaplotypeCaller and further processed with bcftools. Copy-number variants (CNVs) were inferred with FACETS. Additionally, tumor purity and ploidy estimates from FACETS, along with the simulated mean sequencing depth (50×), were supplied as input to SVclone. The ground-truth SVCF values derived from the simulation were compared to estimates produced by SVCFit and SVclone. Error was defined as:

$$Error = SVCF_{est} - SVCF_{true} \ (Eq. 3)$$

We tested for a difference in median error between methods using a two-sided paired Wilcoxon signed-rank test. A p-value < 0.05 was considered statistically significant.

Benchmarking on prostate metastasis mixtures. SVs were called with Manta and genotyped with SVtyper on the mixture BAMs. SVs were filtered following Cmero et al. (2020)[22]: fewer than one supporting read in the matched normal, more than two supporting reads in at least one tumor sample, SV length >300 bp, CCF capped at 1.0, and PASS calls only. SVCF estimates from SVCFit were converted to cancer cell fraction (CCF) using Eq. 2 (Methods), with tumor purity taken from Hong et al. (2015)[23], which reported purity estimates for the source metastases. Error was defined as the difference between the true CCF (computed from the known mixture proportions) and the estimated CCF. For SVclone, we report the mean error across the eleven mixtures as published in Fig. 2C of Cmero et al. (2020), because per-variant error values were not available.

SV Longitudinal Clustering and Dynamics

Structural variants (SVs) in the COMBAT dataset were clustered on a per-subject basis using Dirichlet Process Gaussian Mixture Models (DP-GMM). DP-GMM inference was performed using the scikit-learn library[46] in R via the reticulate interface[47]. For clustering, all SVCF was converted to CCF using purity estimates from FACETS (Eq 2). SVs detected at only one time point were assigned a CCF of 0 at the other time point. To avoid bias in DP-GMM clustering, we retained only the unique CCF pair observed across the two time points. Two features were then derived for clustering: the pre-BAT CCF and the on-BAT CCF. DP-GMM models were fitted using a maximum of 10 components (kmax = 10), a Dirichlet process concentration parameter of 1 (weight_concentration_prior = 1), and a covariance regularization term of reg_covar = 1e−6, covariance_type = "full", kmax=10, init_params = "random". Clusters containing fewer than five data points were merged with the nearest cluster based on Euclidean distance. In addition, clusters whose centroids were separated by less than 0.2 in Euclidean space were consolidated into a single cluster. Alluvial plots were constructed using the R package ‘ggalluvial’[48]. Circos plots were constructed using the R package ‘circlize’[49]. Clonal trees and pie plots were created with SVCFit software.

SV Tumor Evolution Tree Construction and Annotation

SV clusters were ordered onto the edges of an evolutionary tree. Candidate directed edges between clusters where SV clusters were ordered onto the edges of an evolutionary tree. Candidate directed edges between clusters were first identified by enforcing lineage precedence, which required that descendant CCFs not exceed ancestral CCFs by more than a tolerance $\epsilon 1$. Valid edges were assembled into candidate trees using a modified Gabow–Myers algorithm[50], following the pseudocode of Popic *et al.* (2015)[51]. Briefly, the original Gabow–Myers procedure enumerated all directed spanning trees of a directed graph rooted at the ancestral node, and the modification constrained enumeration to edges that satisfied the lineage-precedence filter. Tree topologies that violated the Sum Condition[52], where the sum of descendant CCFs exceeded the ancestral CCF by more than $\epsilon 2$, were excluded. We used $\epsilon 1=0.2$ and $\epsilon 2=0.2$. The tree reconstruction approach did not assume clock-like evolution or edge lengths proportional to time.

The remaining trees were ranked using the SCHISM fitness function[53], which assigned penalties for violations of lineage precedence and the Sum Condition. These penalties were combined as a weighted sum within the exponent of a negative exponential scoring function, such that trees with fewer or smaller violations received higher fitness scores. Equal weights were assigned to the two violation terms (weights = 1). Per-

sample clonal proportions were then inferred from the top-ranked tree, where the proportion of a parent clone was computed as its CCF minus the sum of the CCFs of its children:

$$max([min(CCF_n - CCF_{A(n)}) - \sum_{q \in D(n)} CCF_q], 0)\ (Eq. 4)$$

where:CCF_n is the CCF of the cluster associated with node n. $CCF_{A(n)}$ is the CCF of the cluster directly upstream of node n. $\sum_{q \in D(n)} CCF_q$ is the sum of CCFs of clusters directly downstream of node n.

To annotate the edges of the trees, we first identified a set of prostate cancer driver genes and their hg38 coordinates from Hosseini et al (2025)[54] and web search. Custom R scripts were used to identify driver gene overlap with the coordinates of the SVs called in each subject, including nearby upstream regions.

Structural Variant Down-sampling Validation. To control for differences in tumor purity in the pre-BAT and on-BAT biopsies from Subject 10 (0.68 vs 0.36), we performed a down-sampling validation. The pre-BAT BAM file was subsampled using *samtools view -s* to reduce the total read count, ensuring the effective tumor coverage matched that of the on-BAT sample. Structural variant calling was re-run on this down-sampled dataset using the identical pipeline.

IHC analysis

ERG IHC was performed on an automated staining platform (Ventana/Roche) using a validated IHC assay as described previously[55] using formalin-fixed, paraffin-embedded tumor sections were that were taken adjacent to the fresh frozen tissue used for the whole genome sequencing.

Statistical analysis.

Unless stated otherwise, analyses used the paired Wilcoxon signed-rank test (two-sided, α = 0.05). Medians and interquartile ranges were reported where appropriate. The function wilcox_test from the R package rstatix (v0.7.2) [56] was used to apply the statistical test, and the R package ggpubr (v0.6.0)[57] was used to visualize statistical significance.

Data availability

All open access data is accessible through Mendeley (10.17632/2nhhdjx225.1)

Code Availability

https://github.com/KarchinLab/SVCFit/
MIT license

Acknowledgements

This work was supported by The Prostate Cancer Foundation, the Lustgarten Foundation (Y.L. L.D.W., R.K.), CDMRP/DOD grant W81XWH-21-PCARP-IDA (Y.L. L.D.W., R.K.), and NIH-NCI grants U01CA271273 (J.L., L.D.W., R.K.), P30CA006973 (S.Y, A.M.D.M) and P50CA272391 (S.Y., A.M.D.M), and Break Through Cancer, Cambridge, MA (to R.K., L.D.W., J.L.). E.S.A. is partially supported by NCI grant P30 CA077598 and DOD grant W81XWH-22-2-0025. M.C.M is supported by DOD grant W81XWH-19-1-0692.

Special thanks to Dr. Winston Timp and Dr. Benjamin Langmead for their support of this project in its early days.

Author Contributions

Y.L. and R.K. conceived the study. Y.L. and R.K. developed the methodology. Y.L. implemented the software and performed the computational analyses. J.L. contributed to data analysis and figure generation. M.C.M., E.S.A., A.M.D.M., S.Y., L.D.W., and L.A.S. provided clinical interpretation, biological insight, and domain expertise. L.A.S. coordinated access to clinical data and resources. Y.L., L.A.S., and R.K. wrote the initial manuscript. All authors contributed to manuscript review and editing. R.K. supervised the study.

Competing Interests

LAS receives funding and resources to her institution from Panbela Therapeutics and Abbott Lingo for research that is not relevant to this work. E.S.A. reports grants and personal fees from Janssen, Johnson & Johnson, Sanofi, Bayer, Bristol Myers Squibb, Convergent Therapeutics, Curium, MacroGenics, Merck, Pfizer, and AstraZeneca; personal fees from Aadi Bioscience, Abeona Therapeutics, Aikido Pharma, Astellas, Amgen, Blue Earth, Boundless Bio, Corcept Therapeutics, Duality Bio, Exact Sciences, Hookipa Pharma, Invitae, Eli Lilly, Foundation Medicine, Menarini-Silicon Biosystems, Tango Therapeutics, Tempus, Tolmar Scientific, VIR Biotechnology, and Z-alpha; grants from Novartis, Celgene, and Orion; and has a patent for an AR-V7 biomarker technology that has been licensed to Qiagen. S.Y. reports grants from the NIH/NCI and Commonwealth Foundation during the conduct of the study, as well as grants from Janssen and Celgene/Bristol Myers Squibb, other support from Brahm Astra

Therapeutics and Digital Harmonic, and grants and personal fees from Cepheid, all outside the submitted work.

References

1. Venkatesan, S., Swanton, C., Taylor, B. S. & Costello, J. F. Treatment-induced Mutagenesis and selective pressures sculpt cancer evolution. *Cold Spring Harb Perspect Med* **7**, a026617 (2017).
2. Negrini, S., Vassilis, G. & Thanos, D. Genomic Instability - Evolving Hallmark Cancer. *Nat. Rev. Mol. Cell Biol.*
3. Rhinehart, D. P. *et al.* Intratumoral heterogeneity drives acquired therapy resistance in a patient with metastatic prostate cancer. *NPJ Precis. Oncol.* **8**, 275 (2024).
4. Nowell, P. C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
5. Mardis, E. R. The impact of next-generation sequencing on cancer genomics: From discovery to clinic. *Cold Spring Harb Perspect Med* **9**, a036269 (2019).
6. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
7. Hamdan, A. & Ewing, A. Unravelling the tumour genome: The evolutionary and clinical impacts of structural variants in tumourigenesis. *J Pathol* **257**, 479–493 (2022).
8. Hurles, M. E., Dermitzakis, E. T. & Tyler-Smith, C. The functional impact of structural variation in humans. *Trends Genet* **24**, 238–245 (2008).
9. Scott, A. J., Chiang, C. & Hall, I. M. Structural variants are a major source of gene expression differences in humans and often affect multiple nearby genes. *Genome Res* **31**, 2249–2257 (2021).
10. Yoshimoto, M. *et al.* Absence of TMPRSS2:ERG fusions and PTEN losses in prostate cancer is associated with a favorable outcome. *Mod. Pathol.* **21**, 1451–1460 (2008).
11. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
12. Tomlins, S. A. *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310**, 644–648 (2005).
13. Weier, C. *et al.* Nucleotide resolution analysis of TMPRSS2 and ERG rearrangements in prostate cancer. *J. Pathol.* **230**, 174–183 (2013).
14. Quigley, D. A. *et al.* Genomic Hallmarks and Structural Variation in Metastatic Prostate Cancer. *Cell* **174**, 758-769.e9 (2018).
15. Viswanathan, S. R. *et al.* Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* **174**, 433-447.e19 (2018).
16. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat Commun* **6**, 8971 (2015).
17. Taylor, A. M. *et al.* Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676-689.e3 (2018).
18. Myers, M. A. *et al.* HATCHet2: clone- and haplotype-specific copy number inference from bulk tumor sequencing data. *bioRxiv* 2023.07.13.548855 (2023) doi:10.1101/2023.07.13.548855.

19. Satas, G., Zaccaria, S., El-Kebir, M. & Raphael, B. J. DeCiFering the elusive cancer cell fraction in tumor heterogeneity and evolution. *Cell Syst* **12**, 1004-1018.e10 (2021).
20. Zheng, L., Niknafs, N., Wood, L. D., Karchin, R. & Scharpf, R. B. Estimation of cancer cell fractions and clone trees from multi-region sequencing of tumors. *Bioinformatics* **38**, 3677–3683 (2022).
21. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396–398 (2014).
22. Cmero, M. *et al.* Inferring structural variant cancer cell fraction. *Nat Commun* **11**, 730 (2020).
23. Hong, M. K. H. *et al.* Tracking the origins and drivers of subclonal metastatic expansion in prostate cancer. *Nat Commun* **6**, 6605 (2015).
24. Schweizer, M. T. *et al.* Effect of bipolar androgen therapy for asymptomatic men with castration-resistant prostate cancer: results from a pilot clinical study. *Sci. Transl. Med.* **7**, 269ra2 (2015).
25. Denmeade, S. R., Sena, L. A., Wang, H., Antonarakis, E. S. & Markowski, M. C. Bipolar Androgen Therapy Followed by Androgen Receptor Inhibition as Sequential Therapy for Prostate Cancer. *The Oncologist* **28**, 465–473 (2023).
26. Markowski, M. C. *et al.* Bipolar androgen therapy plus nivolumab for patients with metastatic castration-resistant prostate cancer: the COMBAT phase II trial. *Nat. Commun.* **15**, 14 (2024).
27. Denmeade, S. R. *et al.* TRANSFORMER: A Randomized Phase II Study Comparing Bipolar Androgen Therapy Versus Enzalutamide in Asymptomatic Men With Castration-Resistant Metastatic Prostate Cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **39**, 1371–1382 (2021).
28. Turajlic, S., Sottoriva, A., Graham, T. & Swanton, C. Resolving genetic heterogeneity in cancer. *Nat. Rev. Genet.* **20**, 404–416 (2019).
29. Molecular and Clinical Characterization of Patients With Metastatic Castration Resistant Prostate Cancer Achieving Deep Responses to Bipolar Androgen Therapy - PubMed. https://pubmed.ncbi.nlm.nih.gov/34538582/.
30. Chatterjee, P. *et al.* Supraphysiological androgens suppress prostate cancer growth through androgen receptor-mediated DNA damage. *J. Clin. Invest.* **129**, 4245–4260 (2019).
31. Sena, L. A. *et al.* Androgen receptor activity in prostate cancer dictates efficacy of bipolar androgen therapy through MYC. *J. Clin. Invest.* **132**, e162396 (2022).
32. Kim, N., Eum, H. H. & Lee, H.-O. Clinical Perspectives of Single-Cell RNA Sequencing. *Biomolecules* **11**, 1161 (2021).
33. Ren, X., Kang, B. & Zhang, Z. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome Biol.* **19**, 211 (2018).
34. Bolognini, D. *et al.* VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing. *Bioinformatics* **36**, 1267–1269 (2020).
35. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
36. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).

37. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**, 966–968 (2015).
38. Krueger, F. TrimGalore: A wrapper around Cutadapt and FastQC to consistently apply adapter and quality trimming to FastQ files, with extra functionality for RRBS data.
39. Van Der Auwera, G. A. & O'Connor, B. D. *Genomics in the Cloud*. (O'Reilly Media, 2020).
40. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
41. Cameron, D. L. *et al.* GRIDSS2: comprehensive characterisation of somatic structural variation using single breakend variants and structural variant phasing. *Genome Biol* **22**, 202 (2021).
42. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.* **8**, 14061 (2017).
43. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
45. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res* **44**, e131–e131 (2016).
46. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
47. Interface to Python. https://rstudio.github.io/reticulate/.
48. Brunson, J. C. ggalluvial: Layered Grammar for Alluvial Plots. *J. Open Source Softw.* **5**, 2017 (2020).
49. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
50. Gabow, H. N. & Myers, E. W. Finding All Spanning Trees of Directed and Undirected Graphs. *SIAM J. Comput.* **7**, 280–287 (1978).
51. Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome Biol.* **16**, 91 (2015).
52. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
53. Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q. & Karchin, R. SubClonal Hierarchy Inference from Somatic Mutations: Automatic Reconstruction of Cancer Evolutionary Trees from Multi-region Next Generation Sequencing. *PLoS Comput. Biol.* **11**, e1004416 (2015).
54. Hosseini, N. *et al.* Divergent Clonal Evolution and Early Dissemination Promote Genetic Heterogeneity of Metastases in Castration-Resistant Prostate Cancer. *Cancer Res.* **85**, 4251–4268 (2025).
55. Morais, C. L. *et al.* PTEN loss and ERG protein expression are infrequent in prostatic ductal adenocarcinomas and concurrent acinar carcinomas. *The Prostate* **75**, 1610–1619 (2015).
56. Kassambara, A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2025).

57. Kassambara, A. ggpubr: ‘ggplot2’ Based Publication Ready Plots. (2025).