SVCFit: Inferring structural variant cellular fraction in tumors

Yunzhou Liu[1], Jiaying Lai[1], Laura D. Wood[2,3,4] and 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.

Corresponding Author

Rachel Karchin

Johns Hopkins University

217A Hackerman Hall, 3400 North Charles Street, Baltimore, MD 21218

Phone: 410-516-5578

Fax: 410-516-5294

karchin@jhu.edu

Abstract

Many tumors evolve through cellular mutation and selection, where subpopulations of cells (subclones) with shared ancestry compete for dominance. Introduction of next generation sequencing enables subclone identification using small somatic. However, there are several advantages to marking subclones with structural variants: they have greater functional impact, play a crucial role in late-stage tumors, and provide a more complete view of genomic instability driving tumor evolution. Here, we present SVCFit, a scalable method to estimate the cellular prevalence of somatic deletions, duplications and inversions. We demonstrate that cellular prevalence estimation can be improved by incorporating distinct read patterns for each structural variant type. Additionally, this improvement is achieved without prior knowledge of tumor purity, which is often inaccurate. Using a combination of simulated data and patient-derived metastatic samples with known mixture proportions, we show that our algorithm achieves significantly greater accuracy than state-of-the-art in estimating the structural variants cellular prevalence ($p<0.05$). The speed of SVCFit estimation from cost-effective bulk whole-genome sequencing (WGS) makes it well-suited for analyzing large cohorts of sequenced tumor samples, enhancing the accessibility of SV-based clonal reconstruction.

Introduction

Intratumor heterogeneity encompasses a broad spectrum of molecular alterations within a single tumor, including small somatic mutations (SSMs), structural variants (SVs), copy number variants (CNVs), protein expression changes, 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(Negrini, Gorgoulis and Halazonetis 2010, Venkatesan et al. 2017). Reconstructing the clonal evolution of a tumor can reveal critical events that shape its various phenotypes and underlying mechanisms. Early research into cancer evolution focused primarily on aneuploidy(Nowell 1976), but the advent of Next Generation Sequencing (NGS) enabled scientists to investigate genetic alterations at the single-base level, using SSMs as markers for cancer progression(Mardis 2019). 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(Li et al. 2020, Hamdan and Ewing 2022). For both SSMs and SVs, estimating the cellular fraction—the proportion of cells that harbor a specific alteration—is key to mapping the sequence of these changes throughout a tumor's development. Alterations with the highest cellular fraction likely occurred earlier, and under the infinite sites model (which assumes no mutation loss or parallel evolution), events in the first transformed cell should be present in all tumor cells. These early, clonal events have a cellular fraction of 1.0 with respect to tumor cells, while later, subclonal events occur in smaller cell populations and have lower cellular fractions.

Compared to SSMs, SVs are less frequent, and they are large-scale alterations spanning more than 50 base pairs of DNA. Despite their relative rarity, SVs contribute significantly to genomic diversity, genome instability, and disease susceptibility(Hurles, Dermitzakis and Tyler-Smith 2008). While SSMs affect a single base in a gene or regulatory region, SVs can span thousands of base pairs, impacting multiple genetic components. SVs come in various forms, such as insertions, deletions, tandem duplications, inversions, and translocations. 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(Scott, Chiang and Hall 2021). Additionally, SVs are key drivers of species-level genome evolution, providing the raw material for natural selection and driving evolutionary innovation by creating new genes or altering existing ones(Berdan et al. 2024). This role isn't limited to speciation—SVs can also accelerate tumor

evolution, often leading to poorer patient outcomes(Cosenza, Rodriguez-Martin and Korbel 2022). 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.

Somatic alterations like SSMs and SVs in the human genome can be studied through bulk sequencing, which is widely used due to its affordability. Incorporating somatic SVs detected from bulk sequencing into clonal evolution analysis can be done through variant allele frequency (VAF), which estimates the frequency of an altered allele at a specific locus. VAF is often used to reconstruct cancer evolution, but the process is noisy, due to confounding factors like aneuploidy and tumor purity, both common in cancer samples(Aran, Sirota and Butte 2015, Taylor et al. 2018). A more accurate approach involves using VAF to derive the cellular fraction, which helps account for these confounders. While this method has been widely applied to SSMs(Yuan et al. 2018, Satas et al. 2021, Zheng et al. 2022, Roth et al. 2014), only one published method—SVclone(Cmero et al. 2020)—estimates cellular fraction of SVs from bulk sequencing data. The equation to calculate VAF, as the ratio of mutated reads to total reads at a genomic position, works well for SSMs but does not transfer well to SVs, because it fails to account for the differences between SSMs and SVs. SSMs are detected at a single genomic position, while SVs are detected across multiple genomic positions due to their larger size. SVs also encompass various types, each affecting the genome differently, leading to distinct detection patterns. Therefore, it's essential to study the specific characteristics of SVs and integrate them into VAF and cellular fraction estimation. Furthermore, most current methods require an estimation of tumor purity, which can be assessed either through visual inspection by a pathologist—prone to human error—or computational approaches, which can also be inaccurate.

In this work, we introduce a new approach, SVCFit, to estimate the cellular fraction of structural variants (SVCF). Our method addresses issues with VAF calculation and purity estimation. We demonstrate how to incorporate the unique characteristics of different SV types into VAF calculation and derive the SVCF metric, which can be used for clonality inference, without needing to estimate sample purity. SVCFit outperforms SVclone in both a simulated dataset and a mixture of prostate metastases samples(Hong et al. 2015), and it is 412 times faster on larger dataset.

Method

VISOR benchmark set

Five single-clone and five multiple-clone tumor samples with known SVs and SVCFs were simulated using VISOR (version 1.1.2)(Bolognini et al. 2020). VISOR requires a Browser Extensible Data (BED) file(Kent et al. 2002) containing the desired SVs, along with SV tumor purity and clone proportion parameters, and generates a Binary Alignment Map (BAM) file of paired-end DNA sequencing reads aligned to a reference genome. Each clone within a sample is simulated independently. Simulations were conducted on human Chromosome 1 for both single-clone and multiple-clone samples. For single-clone simulations, 1,000 somatic SVs of varying sizes (100 to 10000 base pairs, non-overlapping) were used as input across five SV tumor purity levels: 20%, 40%, 60%, 80%, and 100%, with an average sequencing coverage of 40x. In multiple-clone simulations, the same SV parameters were applied. First, a tree with three clones was generated using the ape algorithm (version 5.8)(Paradis and Schliep 2019). Then, 1000 somatic SVs were randomly divided into three files, where each file represents a clone. The clones were simulated with given clone proportions: 100%, 60%, and 30%. A matched normal sample with 100% SV tumor purity and 40x coverage was also simulated to enable somatic SV detection, which was performed using Manta (version 1.6.0) (Chen et al. 2016) with default parameters.

Prostate metastasis mixtures benchmark set

Eleven multiple-clone samples were generated by mixing two prostate metastatic samples, downloaded from European Genome-phenome Archive (EGAD00001001343)(Hong et al. 2015), following the protocol presented in SVclone(Cmero et al. 2020). Briefly, whole-genome sequenced samples from the same patient were mixed at known subclonal proportions. Three-cluster mixtures were subsampled at 10% clonal frequency increments, and four- and five-cluster mixtures subsampled by chromosomes. Samples were downloaded from the European Genome-phenome Archive (EGAD00001001343)(Hong et al. 2015). Raw sequences were aligned to the hs37d5 reference genome with BWA (version 0.7.18)(Li and Durbin 2009). SVs were called using Manta (version 1.6.0)(Chen et al. 2016) and filtered as described in SVclone(Cmero et al. 2020) (normal reads < 1, tumor reads >2 in at least one tumor sample, length > 300bp, cancer cell fraction (*CCF*) capped at 1.0, "PASS" only). For direct comparison to SVclone, we converted *SVCF* to *CCF* as

$$CCF = \text{SV_tumor_purity} \times \text{SVCF} \quad (1)$$

SVCFit and SVclone comparison

Benchmark on data simulated with VISOR

The start and end genomic positions of each SV, the SV type, and reference and alternative read counts from normal and tumor BAMs simulated with VISOR were called by Manta. SVclone (v1.1.2) was downloaded from https://github.com/mcmero/SVclone on March 10, 2024.

For SVclone, genomic positions, VISOR-defined tumor purity and tumor ploidy, and the tumor BAM file were provided. For SVCFit, genomic positions, SV type, and reference and alternative counts from the normal and tumor BAMs were provided. SVclone was run with default parameters except for mean coverage, which was set to 40. The *SVCF* for each SV was calculated by taking the *CCF* (Eq 1) estimate from SVclone at each reported breakpoint and/or breakend. For SVCFit, the *SVCF* for each SV was calculated with Eq S1 (Supplementary Material). Error was defined as the difference between a method's estimated *SVCF* and the true *SVCF* from VISOR. Performance was evaluated as a significant difference in median error between SVclone and SVCFit (Wilcoxon Ranked Sum Test, $p<0.05$).

Benchmark on prostate metastasis mixtures

Manta was used to call start and end genomic positions of each SV, the SV type, and reference and alternative read counts from normal and tumor mixture BAMs. SVs were filtered based on the protocol described in SVclone(Cmero et al. 2020). *SVCF* was calculated by SVCFit as described previously. *SVCF* was converted to *CCF* (Eq 1), using the tumor purity provided in the source publication(Hong et al. 2015). Error was defined as the difference between the estimated *CCF* and the true *CCF* calculated from the known mixture proportions. Mean SVclone error for the eleven mixtures was taken from Fig. 2C in SVclone(Cmero et al. 2020). Raw error data was not available in the publication materials and the authors did not respond to our requests.

Timing comparison

Fifty single-clone tumor samples were simulated with VISOR. To balance computational feasibility and provide representative results for the genome-wide performance of our approach, the timing test was limited to three chromosomes (Chr 1, Chr 2, Chr3). Samples were generated with structural variations (SVs) ranging from 50 to 50,050, incrementing by 1,000, with 100% tumor purity and 50x coverage. The BAM files generated by VISOR were processed by Manta to detect somatic SVs. Manta's VCF outputs were preprocessed to extract key information: SV start and end genomic positions, SV type, and reference and alternative read counts for SVCFit. VISOR-defined tumor purity and ploidy were also processed for SVclone analysis only.

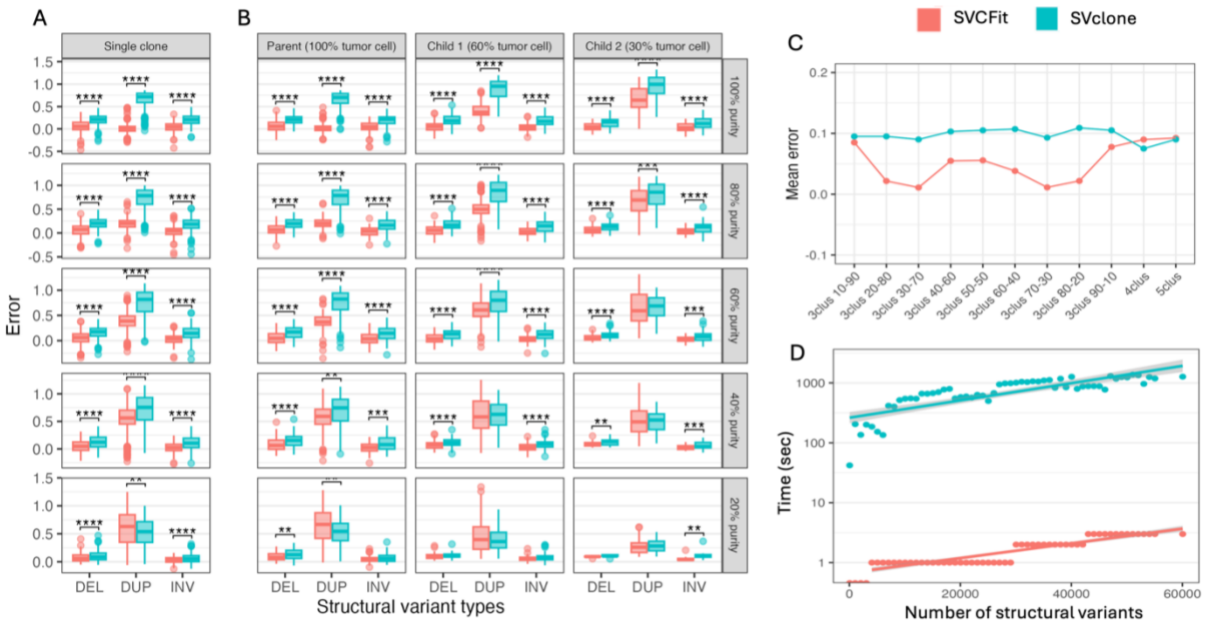

Fig 1. Benchmark results for SVCF estimation accuracy and runtime comparison. All plots share the same legend (SVCFit: red, SVclone: blue). A) Benchmark result on VISOR-simulated single-clone samples summarized in a boxplot. The x-axis are three types of structural variants (deletion, tandem duplication, and inversion) included in this study, and y-axis is the difference of estimated *SVCF* from VISOR defined true *SVCF*. Wilcoxon sum rank test is applied to assess if two groups are significantly different (* p<0.05, ** p<0.01, *** p<0.001, **** p< 0.0001) B) Benchmark result on VISOR-simulated multiple-clone samples summarized in a boxplot. Each column is a (sub)clone, and each row is structural variants purity level. This boxplot has the same axis and annotation as described in A. C) Mean error comparison on Prostate metastasis mixtures. The x-axis are different mixtures created from metastasis samples and the y-axis is the mean of the difference between estimated *CCF* and known *CCF* from mixture proportion. D) Algorithm runtime summarized in a dot plot with fitted linear trendline. The x-axis is the number of structural variants included in a tumor sample simulated by VISOR, and y-axis is how many seconds an algorithm takes to calculate cellular fraction presented in log scale.

Performance evaluation

SVCFit improves estimation of the fraction of tumor cells harboring a structural variant

Evaluation on simulated data

Accurate estimation of the cellular fraction of cells harboring a structural variant (SV) is essential to characterize the clonal architecture of a tumor. We compared the estimates of SVCFit and that of the state-of-the-art algorithm SVclone on two benchmark data sets. First, we generated artificial tumor and normal BAMs of Chromosome 1 with “spiked-in” SVs, using the Haplotype-Aware Structural Variants Simulator (VISOR)(Bolognini et al. 2020) for short, long and linked reads. Each tumor clone was assigned a cellular fraction, and each “spiked-in” SVs was assigned to a tumor clone, yielding a known SVCF, over a range of structural variant tumor purities (SV_tumor_purity = 20%, 40%, 60%, 80%, 100%).

$$SV_tumor_purity = \frac{N_{sv}}{N} \tag{2}$$

where N_{sv} is the number of cells in a tumor sample that contain any somatic structural variant and N is the total number of cells in the sample.

For the VISOR single-clone simulations, the SVCFit error ($SVCF_{true} - SVCF_{est}$) was significantly lower than the SVclone error for all tested structural variant tumor purities (Wilcoxon Ranked Sum Test, p<1e-4), except for tandem duplications at 20% tumor purity where SVclone outputs lower error (Figure 1A). For the multiple-clone simulations, the SVCFit error was significantly lower than SVclone for all (sub)clones, at most structural variant tumor purities (Wilcoxon Ranked Sum Test *: p < 0.05; **: p < 0.01; ***: p < 0.001; ****: p < 0.0001) (Figure 1B).

Evaluation on in-silico mixtures

Next, we applied the in silico mixing approach developed by the SVclone authors, using two previously sequenced metastatic prostate tumors, each with known purity (49% and 46%) and harboring a single clone that diverged from the same clone in the primary tumor(Hong et al. 2015). The two tumors were mixed with varying proportions to create heterogeneous samples with known SV cellular fractions and clonal composition. Nine mixtures were created to generate tumors with three clones, in which the first clone harbors SVs shared between the two tumors, the second harbors shared SVs and those private to the first tumor and the third harbors shared SVs and those private to the second tumor. In addition, four-clone and five-clone mixtures were created using different subsampling rates for odd and even chromosome sets, as

described in SVclone(Cmero et al. 2020). Because the tumor purity, mixture proportions and clone overlaps are known, the true *SVCF* for each SV is also known. Next, we calculated the average difference between the true and estimated *SVCF* for three types of SVs (inversion, tandem duplication, deletion) with SVCFit and SVclone. SVCFit had lower total mean error than SVclone for the 11 mixtures (Total mean error: SVclone=1.067, SVCFit=0.56). For all mixtures with 3 clones, SVCFit had lower or comparable mean error than SVclone. For the 4-clone and 5-clone mixtures, SVclone had lower error (Fig 1C).

SVCFit run faster with large dataset

Fifty single-clone samples with SVs of 1000 increment ranging from 50 to 50050 were simulated by VISOR with 100% tumor purity and 50x coverage. Structural variants from these simulations were detected by Manta and the data was processed by SVCFit and SVclone to compare algorithm runtime. For all fifty simulated samples with the number of SVs ranging from 50 to 50050, SVCFit run significantly faster than SVclone, even on a logarithmic scale (Fig 1D). For 50 SVs simulation, SVCFit took less than 1 seconds and SVclone took 42 seconds; for 50050 SVs simulation, SVCFit took 3 seconds and SVclone took 1236 seconds.

Discussion

SVCFit was developed to estimate the cellular prevalence of somatic structural variants in cancers, using bulk whole genome sequencing data. SVs with similar cellular prevalence are expected to have originated in a common ancestral cell and can be used as markers of clonal and subclonal populations within a heterogeneous tumor. This new method advances the state-of-the-art, achieving significantly lower error with respect to the previously published SVclone tool in single and multiple-clone simulated data. Out of 60 comparisons across simulated deletions, tandem duplications, and inversions of known cellular prevalence (measured as either *SVCF* or *CCF*), SVCFit had significantly lower error in 46 comparisons, SVclone in two, and 12 comparisons were not significantly different between the two methods. On a set of 11 artificially mixed metastatic prostate samples that were used as a benchmark in the SVclone paper, SVCFit had lower mean error on seven, SVclone on two, and two were tied. SVCFit was significantly faster than SVclone,

and it was able to estimate the *SVCF* of 50050 SVs in 3 seconds. SVclone required 1236 seconds to jointly cluster and estimate the *CCF* of 50050 SVs. An exact head-to-head speed comparison was not possible because SVCFit does not do clustering. Finally, SVCFit uniquely estimates *SVCF* differently for each SV type.

Accurate, high-throughput estimation of structural variant (SV) cellular prevalences from cost-effective bulk whole-genome sequencing (WGS) could greatly enhance the accessibility of SV-based clonal reconstruction. While single-cell sequencing has been instrumental in reconstructing clonal evolution using SSMs(Potter et al. 2013, Bai et al. 2024, Bristy, Fu and Schwartz 2023) , its utility for SVs remains limited due to challenges such as amplification artifacts, dropout, and low read depth(Yi, Nam and Jeong 2023, Macaulay and Voet 2014). Recent advances, such as DLP Plus(Laks et al. 2019), have improved single-cell technologies by providing higher and more uniform coverage, reduced amplification artifacts, and enhanced resolution for SV detection, making it suitable for clonal reconstruction. However, the high cost of DLP Plus remains a significant barrier for many research labs.

The speed of SVCFit estimation makes it well-suited for analyzing large cohorts of sequenced tumor samples. In a timing comparison between SVCFit and SVclone, we demonstrated that SVCFit is significantly faster. However, this comparison has some limitations, as SVclone jointly infers both *CCF* and clonal assignments. Nevertheless, it is clear that SVclone's *CCF* estimation alone is substantially slower than SVCFit, as it requires multiple rounds of parameter estimation using variational inference, while SVCFit relies on interpretable and efficient equations implemented in just a few lines of code.

Due to its increased accuracy and speed, SVCFit can be incorporated into algorithms that reconstruct tumor evolutionary trees at cohort-level scale, using the cellular fractions of both SSMs and SVs. This will be an important future extension of the current method. While there are methods(Ricketts et al. 2020, Fu et al. 2022) that integrate both SSMs and SVs, they don’t estimate cellular fractions and they may not scale to larger data sets(Fu et al. 2022).

For many tumor types, reconstructing tumor evolution exclusively from SSMs fails to account for driver alterations that could explain the emergence of new clones(Fu et al. 2022). We are now investigating SVs using deep whole genome sequencing data of pancreatic cancer precursor lesions, to explore the hypothesis that clones in these lesions may be initiated by SVs.

As long-read sequencing continues to improve in accuracy and affordability, we anticipate that SVCFit can be adapted to sequencing modalities beyond paired-end sequencing. Long-read technologies will enhance the detection of structural variants (SVs), particularly very large SVs and subclonal SVs present at low cellular fractions. Additionally, incorporating phasing information to infer zygosity is expected to further enhance the accuracy of our approach.

Acknowledgements

We thank Yi Yang, Dr. Winston Timp, and Dr. Ben Langmead for their valuable feedback.

Data and code availability

The open access data underlying this article are available in Mendeley at https://dx.doi.org/10.17632/bwzb6n3xbc.1. The prostate metastasis raw data can be acquired upon request from European Genome-phenome Archive (EGAD00001001343). All code is open source and is available under an MIT license at https://github.com/KarchinLab/SVCFit.

Funding

This work was funded by a grant from the Lustgarten Foundation (YL LDW, RK), CDMRP/DOD grant W81XWH-21-PCARP-IDA (YL LDW, RK), and NIH-NCI grant U01CA271273 (JL, LDW, RK).

Conflict of Interest: none declared.

References

Aran, Dvir, Sirota, Marina, and Butte, Atul J., 'Systematic Pan-Cancer Analysis of Tumour Purity', *Nature Communications*, 6/1 (2015), 8971

Bai, Xiangqi, Duren, Zhana, Wan, Lin, and Xia, Li C, 'Joint Inference of Clonal Structure Using Single-Cell Genome and Transcriptome Sequencing Data', *NAR Genomics and Bioinformatics*, 6/1 (2024), lqae017

Berdan, Emma L., Aubier, Thomas G., Cozzolino, Salvatore, Faria, Rui, Feder, Jeffrey L., Giménez, Mabel D., et al., 'Structural Variants and Speciation: Multiple Processes at Play', *Cold Spring Harbor Perspectives in Biology*, 16/3 (2024), a041446

Bolognini, Davide, Sanders, Ashley, Korbel, Jan O, Magi, Alberto, Benes, Vladimir, and Rausch, Tobias, 'VISOR: A Versatile Haplotype-Aware Structural Variant Simulator for Short- and Long-Read Sequencing', *Bioinformatics*, 36/4 (2020), 1267–69

Bristy, Nishat Anjum, Fu, Xuecong, and Schwartz, Russell, 'Sc-TUSV-Ext: Single-Cell Clonal Lineage Inference from Single Nucleotide Variants (SNV), Copy Number Alterations (CNA) and Structural Variants (SV)', *bioRxiv: The Preprint Server for Biology*, 2023, 2023.12.07.570724

Chen, Xiaoyu, Schulz-Trieglaff, Ole, Shaw, Richard, Barnes, Bret, Schlesinger, Felix, Källberg, Morten, et al., 'Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications', *Bioinformatics (Oxford, England)*, 32/8 (2016), 1220–22

Cmero, Marek, Yuan, Ke, Ong, Cheng Soon, Schröder, Jan, Corcoran, Niall M., Papenfuss, Tony, et al., 'Inferring Structural Variant Cancer Cell Fraction', *Nature Communications*, 11/1 (2020), 730

Cosenza, Marco Raffaele, Rodriguez-Martin, Bernardo, and Korbel, Jan O., 'Structural Variation in Cancer: Role, Prevalence, and Mechanisms', *Annual Review of Genomics and Human Genetics*, 23/Volume 23, 2022 (2022), 123–52

Fu, Xuecong, Lei, Haoyun, Tao, Yifeng, and Schwartz, Russell, 'Reconstructing Tumor Clonal Lineage Trees Incorporating Single-Nucleotide Variants, Copy Number Alterations and Structural Variations', *Bioinformatics (Oxford, England)*, 38/Suppl 1 (2022), i125–33

Hamdan, Alhafidz, and Ewing, Ailith, 'Unravelling the Tumour Genome: The Evolutionary and Clinical Impacts of Structural Variants in Tumourigenesis', *The Journal of Pathology*, 257/4 (2022), 479–93

Hong, Matthew K. H., Macintyre, Geoff, Wedge, David C., Van Loo, Peter, Patel, Keval, Lunke, Sebastian, et al., 'Tracking the Origins and Drivers of Subclonal Metastatic Expansion in Prostate Cancer', *Nature Communications*, 6/1 (2015), 6605

Hurles, Matthew E., Dermitzakis, Emmanouil T., and Tyler-Smith, Chris, 'The Functional Impact of Structural Variation in Humans', *Trends in Genetics : TIG*, 24/5 (2008), 238–45

Kent, W. James, Sugnet, Charles W., Furey, Terrence S., Roskin, Krishna M., Pringle, Tom H., Zahler, Alan M., et al., 'The Human Genome Browser at UCSC', *Genome Research*, 12/6 (2002), 996–1006

Laks, Emma, McPherson, Andrew, Zahn, Hans, Lai, Daniel, Steif, Adi, Brimhall, Jazmine, et al., 'Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing', *Cell*, 179/5 (2019), 1207-1221.e22

Li, Heng, and Durbin, Richard, 'Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform', *Bioinformatics (Oxford, England)*, 25/14 (2009), 1754–60

Li, Yilong, Roberts, Nicola D., Wala, Jeremiah A., Shapira, Ofer, Schumacher, Steven E., Kumar, Kiran, et al., 'Patterns of Somatic Structural Variation in Human Cancer Genomes', *Nature*, 578/7793 (2020), 112–21

Macaulay, Iain C., and Voet, Thierry, 'Single Cell Genomics: Advances and Future Perspectives', *PLOS Genetics*, 10/1 (2014), e1004126

Mardis, Elaine R., 'The Impact of Next-Generation Sequencing on Cancer Genomics: From Discovery to Clinic', *Cold Spring Harbor Perspectives in Medicine*, 9/9 (2019), a036269

Negrini, Simona, Gorgoulis, Vassilis G., and Halazonetis, Thanos D., 'Genomic Instability — an Evolving Hallmark of Cancer', *Nature Reviews Molecular Cell Biology*, 11/3 (2010), 220–28

Nowell, P. C., 'The Clonal Evolution of Tumor Cell Populations', *Science (New York, N.Y.)*, 194/4260 (1976), 23–28

Paradis, Emmanuel, and Schliep, Klaus, 'Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R', *Bioinformatics*, 35/3 (2019), 526–28

Potter, Nicola E., Ermini, Luca, Papaemmanuil, Elli, Cazzaniga, Giovanni, Vijayaraghavan, Gowri, Titley, Ian, et al., 'Single-Cell Mutational Profiling and Clonal Phylogeny in Cancer', *Genome Research*, 23/12 (2013), 2115–25

Ricketts, Camir, Seidman, Daniel, Popic, Victoria, Hormozdiari, Fereydoun, Batzoglou, Serafim, and Hajirasouliha, Iman, 'Meltos: Multi-Sample Tumor Phylogeny Reconstruction for Structural Variants', *Bioinformatics (Oxford, England)*, 36/4 (2020), 1082–90

Roth, Andrew, Khattra, Jaswinder, Yap, Damian, Wan, Adrian, Laks, Emma, Biele, Justina, et al., 'PyClone: Statistical Inference of Clonal Population Structure in Cancer', *Nature Methods*, 11/4 (2014), 396–98

Satas, Gryte, Zaccaria, Simone, El-Kebir, Mohammed, and Raphael, Benjamin J., 'DeCiFering the Elusive Cancer Cell Fraction in Tumor Heterogeneity and Evolution', *Cell Systems*, 12/10 (2021), 1004-1018.e10

Scott, Alexandra J., Chiang, Colby, and Hall, Ira M., 'Structural Variants Are a Major Source of Gene Expression Differences in Humans and Often Affect Multiple Nearby Genes', *Genome Research*, 31/12 (2021), 2249–57

Taylor, Alison M., Shih, Juliann, Ha, Gavin, Gao, Galen F., Zhang, Xiaoyang, Berger, Ashton C., et al., 'Genomic and Functional Approaches to Understanding Cancer Aneuploidy', *Cancer Cell*, 33/4 (2018), 676-689.e3

Venkatesan, Subramanian, Swanton, Charles, Taylor, Barry S., and Costello, Joseph F., 'Treatment-Induced Mutagenesis and Selective Pressures Sculpt Cancer Evolution', *Cold Spring Harbor Perspectives in Medicine*, 7/8 (2017), a026617

Yi, Dohun, Nam, Jin-Wu, and Jeong, Hyobin, 'Toward the Functional Interpretation of Somatic Structural Variations: Bulk- and Single-Cell Approaches', *Briefings in Bioinformatics*, 24/5 (2023), bbad297

Yuan, Ke, Macintyre, Geoff, Liu, Wei, Group, PCAWG-11 working, and Markowetz, Florian, 'Ccube: A Fast and Robust Method for Estimating Cancer Cell Fractions' (2018), 484402 <https://www.biorxiv.org/content/10.1101/484402v1> [accessed 30 January 2025]

Zheng, Lily, Niknafs, Noushin, Wood, Laura D., Karchin, Rachel, and Scharpf, Robert B., 'Estimation of Cancer Cell Fractions and Clone Trees from Multi-Region Sequencing of Tumors', *Bioinformatics (Oxford, England)*, 38/15 (2022), 3677–83