Copy number variation profiling using multi-individual single-nucleus data in prostate cancer tumors

THEA TRAW[1], Terence Li[2], Cuining Liu[2], Yi Zhang[3], Kevin Abuhanna[3], Rong Rong Huang[4], Paul Boutros[5], Huihui Ye[4], Chongyuan Luo[3]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA
2 Bioinformatics Interdepartmental PhD Program, UCLA
3 Department of Human Genetics, David Geffen School of Medicine, UCLA
4 Department of Pathology and Laboratory Medicine, UCLA
5 Department of Urologic Oncology, UCLA

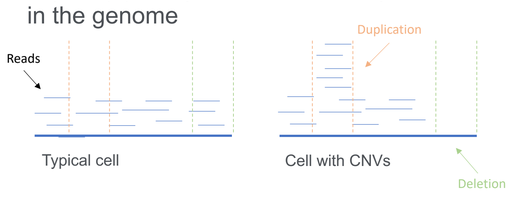
Abstract

Cancer is a highly heterogeneous disease and the molecular basis of its progression is not well understood. A common hallmark in cancerous cells is the development of copy number variations (CNVs) across the genome, which then propagate as the tumor grows. To determine the extent to which CNVs vary at an individual level, we analyzed a 5-patient prostate cancer dataset generated from single nucleus methyl-3C sequencing (sn-m3C-seq). After applying a single-cell based CNV caller, we investigated the differences between CNV profiles across individuals, cell types, and tumor spatial locations. For instance, we found that donor BS13497 expressed significantly more CNVs, particularly in luminal cells. Also, CNV counts varied widely by spatial location, with benign cells containing the least. This project highlights the variability of CNV profiles in prostate cancer across individuals, cell types, and spatial locations, with the further potential to identify the biological significance of the most salient CNVs.

Background/Motivation

What is a copy number variation?

A deviation from the typical number of copies (2) of a particular DNA sequence in the genome



Why look at CNVs?

- Associated with cancer
- Heterogeneous profiles of CNVs in tumor cells

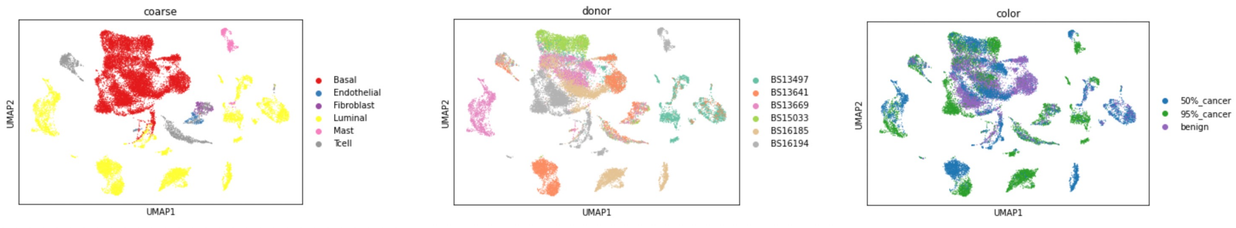
The dataset

- sn-m3C-seq sequencing (joint methylation and chromatin capture)
- Cell type inferred from cell-specific profile
- Spatial location in the tumor
  - Batched in plates of 384 cells

Project objectives

- To use single-cell CNV calling pipeline (Ginkgo) to identify differences across cell type, donor, and spatial location
- To validate those CNVs

UMAPs identifying clusters by cell type, donor, and spatial location



Methodology

Ginkgo pipeline

Sequencing data (23040 cells) → Alignment (hg38)
Coverage → Binning (1Mb non-overlapping windows)
QC (> 900K reads) (19613 cells) → Ginkgo
CNV calls → Cell profiles (19592 cells)

CNV profiles per cell

- Re-bin each cell by median CNV length in chromosome
- A bin contains a CNV if more than 25% of it is covered

Validation

- QC to control for technical artifacts
- Calculate similarity within spatial locations with Jaccard index
  - Single-cell to pseudobulk
  - Single-cell to single-cell

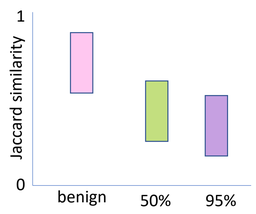
What is a "pseudobulk"?

Aggregate of all reads across a plate to simulate bulk data

- BS16194 benign cells
- Run Ginkgo on each plate-wise pseudobulk
- Provide cumulative picture of non-cancerous CNVs

Strategies for validating

Determine similarity amongst benign cells vs amongst cancerous cells
- More heterogeneity in cancerous cells should lead to lower similarity scores



CNV profile concordance

- Pseudobulk "model vectors" to single-cell
- Single-cell to single-cell

Validation

Pseudobulk to single-cell comparison

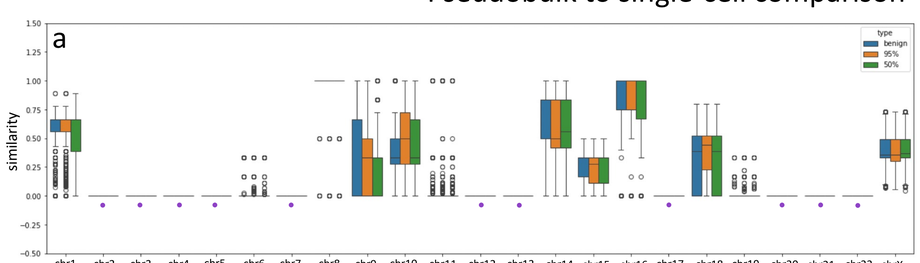


Figure 2a. Jaccard similarity scores per chromosome between benign BS16194 pseudobulk "model vectors" and all BS16194 single-cell profiles, by spatial location. Chromosomes marked with purple dots have no CNVs called in the pseudobulk data. Similar comparison results across locations may suggest that the pseudobulk CNV calls were not sufficiently representative of a "true" benign cell profile.

Single-cell to single-cell comparison

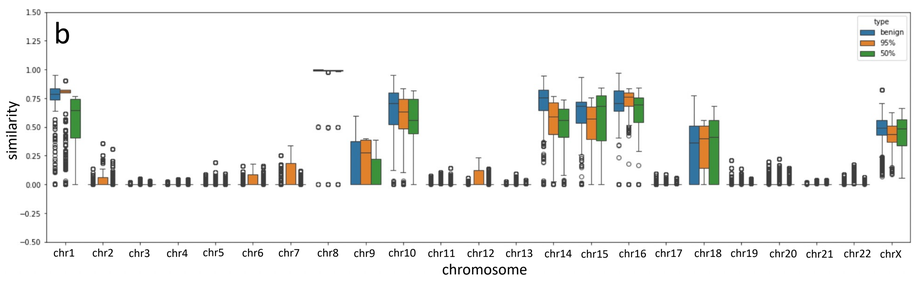


Figure 2b. Jaccard similarity scores per chromosome between all single-cell profiles of each spatial location against themselves. Benign values tend to be higher than that of 95% cancer and 50% cancer, indicating a more homogenous set of CNV profiles.

CNV calling results

Overall distribution of CNV counts

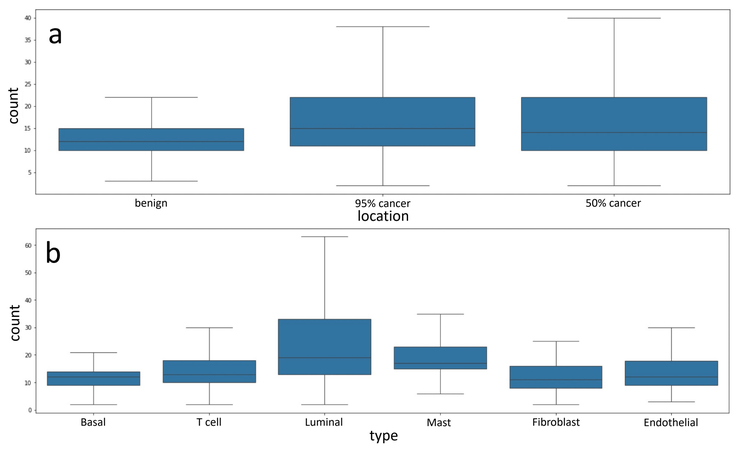


Figure 1a-b. Number of CNVs per cell across spatial location (a) and cell type (b). Cancerous cells have more CNVs than benign, but 95% vs 50% regions do not show a noticeable difference. Luminal cells exhibit the most CNVs.

Whole-chromosome aneuploidy events

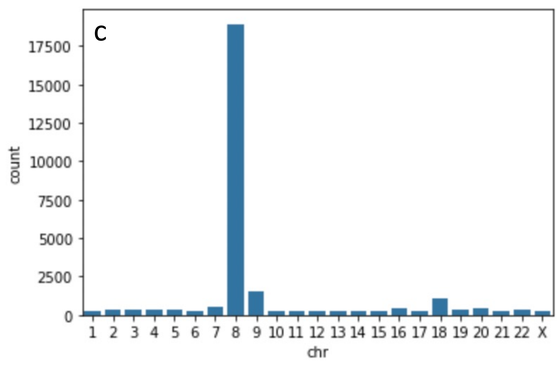
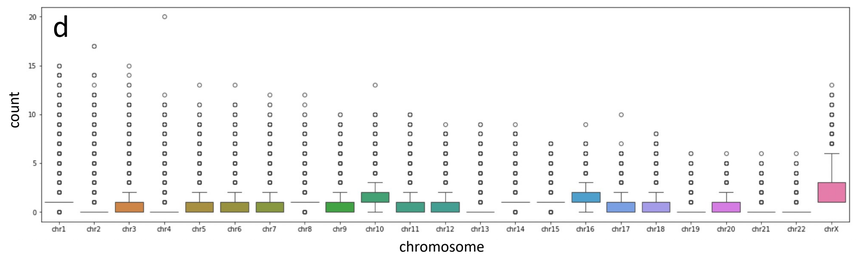


Figure 1c. Number of cells that contain a whole-chromosome (>90%) aneuploidy event. 98% of all cells contain an event at chromosome 8, a common characteristic of prostate cancer. Chromosomes 9 and 18 are the next most frequent, at 8% and 5% respectively, which could be of biological interest.

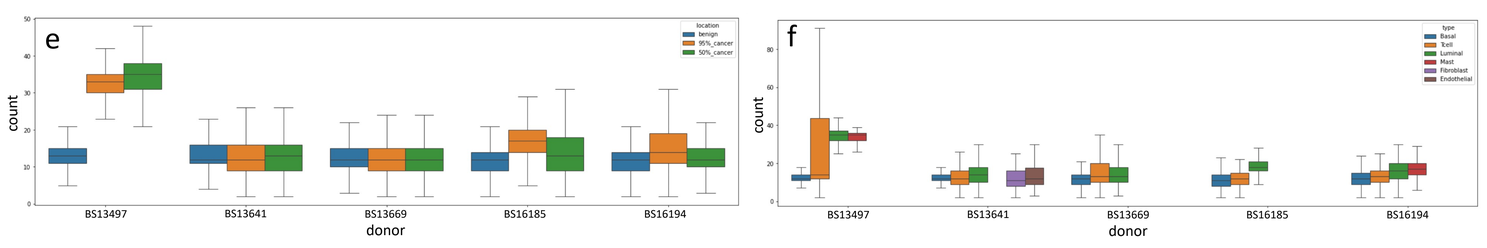
CNV counts at the chromosome level

Figure 1d. CNV counts per chromosome. The X chromosome tends to have more CNVs, as do chromosomes 10 and 16. Chromosomes 2, 4, 13, 19, 21, and 22 are the least commonly affected.



CNV counts for donor BS13497

Figure 1e-f. Number of CNVs by donor, stratified by spatial location (e) and cell type (f). BS13497 has demonstrably higher CNV counts for 95% and 50% cancer cells than those regions in all other donors, with a similar distribution for benign cells that serves as a control. Similarly, BS13497 presents more CNVs in T-, luminal, and mast cells, while having comparable results for basal cells.



Takeaways

- Applied single-cell based CNV caller to multi-omic sequencing prostate cancer data
- Validated calls with a CNV profile-concordance approach based on the expected heterogeneity of tumor cells
- Implemented a non-overlapping windows binning strategy to identify cell-to-cell comparable CNV profiles
- Observed that CNV expression differs across donor, cell type, spatial location, and chromosome, with potential biological significance (donor-unique CNV profiles; varied responses of cell types under prostate cancer; necessity for spatial data; distinct CNV accumulation patterns in chromosomes)

Next steps

- Are the CNVs occurring in biologically interesting locations within the genome?
- Which genes are most frequently impacted by CNVs?
- Validate CNVs using bulk WGS data (establish as "ground truth")
- Can single-cell based CNV callers like Ginkgo effectively handle pseudobulk data?
- Account for the complications of bisulfite conversion in alignment
- The overwhelming majority of presumed benign cells express a whole-chromosome aneuploidy event at chromosome 8. Can these cells fully act as a control against the cancerous cells?

References

Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G. S., Hicks, J., Wigler, M., & Schatz, M. C. (2015). Interactive analysis and assessment of single-cell copy-number variations. *Nature methods*, *12*(11), 1058–1060. https://doi.org/10.1038/nmeth.3578