Scripps Research
Scripps Research Translational Institute
UC San Diego
JACOBS SCHOOL OF ENGINEERING
Shu Chien-Gene Lay Department of Bioengineering

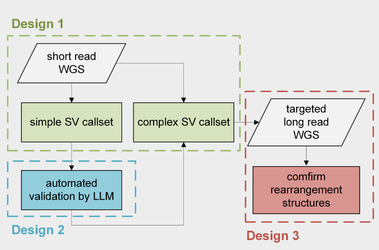
Team 22: Deciphering DNA Repair Pathways through Structural Variant Analysis in Cancer Genomes

Shania Bu, Elliott Ou
Mentors: Xiaohua Wu, Salvatore Loguercio
Wu Lab, Department of Molecular & Cellular Biology, Scripps Research

INTRODUCTION

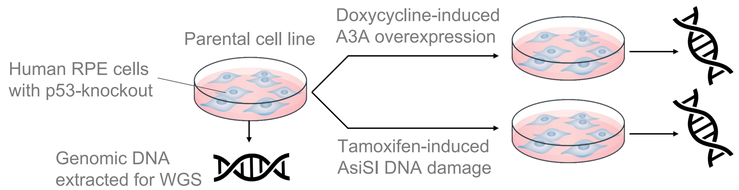
Structural variants are large-scale genomic alterations that fundamentally reshape the architecture of cancer DNA, yet their accurate detection remains a significant hurdle. Current computational tools often diverge by as much as 40-60% because different algorithms prioritize distinct alignment signals, while the traditional bottleneck of manual visual verification is simply not scalable for modern research. To bridge this gap, our project establishes a robust, high-confidence discovery pipeline designed to sensitively identify both simple and complex rearrangements. Our approach integrates three complementary strategies to ensure maximum accuracy and scale. First, we merge the outputs of five specialized callers to capture a comprehensive union of variants. Second, we eliminate manual curation bottlenecks by implementing an LLM-based vision system that automates the classification of genomic coverage plots into true or false events. Finally, we utilize targeted PacBio long-read sequencing to resolve ambiguous complex sites and verify long-range repair outcomes.

By leveraging these structural patterns, our computational workflow maps genomic instability back to specific DNA repair pathway deficiencies, creating a foundation for moving into clinical cancer genomics.

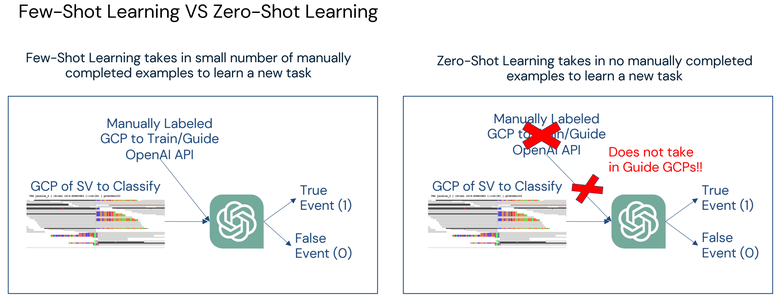


METHODS

Experimental Design

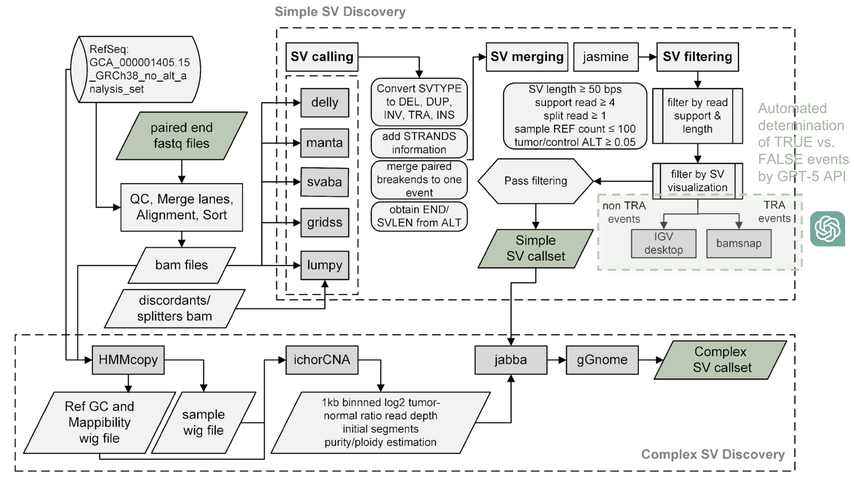


Classifier Design



Computational pipeline for simple and complex structural variant discovery

Detecting structural variants (SVs) with short-read data is hindered by caller discordance and the subjectivity of manual review. To address this, we developed an integrated framework that merges SV calls from DELLY, LUMPY, Manta, svABA, and GRIDSS, applying stringent filtering to the union set. We automate validation via GPT-5-mediated IGV plot assessment and resolve complex rearrangements using JaBbA and gGnome for breakpoint graph reconstruction.



RESULTS

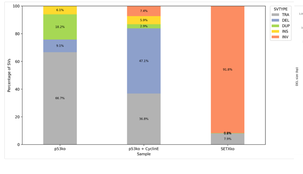
Structural variant discovery results

Simple SVs: Simple structural variants are classified into five types: deletion, duplication, insertion, inversion, and translocation.

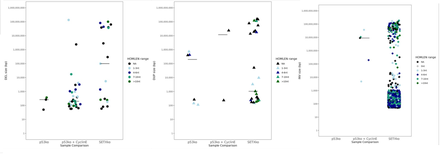
Total SV count

Group	Total Count
p53ko	33
CycE	68
SETXko	18156

SV type distribution

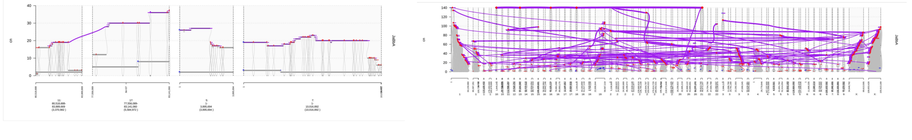


SV length & microhomology length distribution

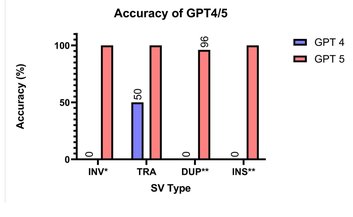


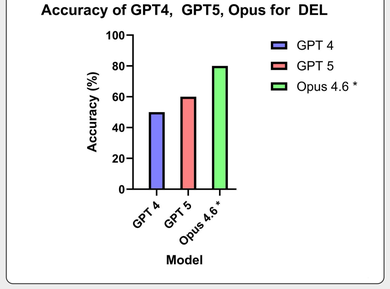
Complex SVs: The different complex SV types are defined by gGnome and inferred from binned read depth data and simple SV breakpoints.

Double minute (DM) events from SETXko



AI Classifier Accuracy





CONCLUSIONS & FUTURE WORK

I. Drivers of SV Formation and Genome Instability

SETX deficiency leads to R-loop accumulation at double-strand breaks, preventing the resolution of RNA/DNA hybrids. This forces a shift from canonical repair toward mutagenic Break-Induced Replication (BIR), which is prone to template switching. Mechanistically, this explains the elevated inversions and interchromosomal junctions observed in our samples, as BIR-like repair generates specific structural "scars" and genome instability. Cyclin E overexpression induces significant replication stress and replication fork collapse, increasing the overall structural variant burden. These collapses generate double-strand breaks that are resolved through error-prone pathways. This instability manifests as a dominance of deletions and translocations resulting from the mis-joining of distant or nearby DNA breaks.

II. Improving Consistency in SV Calling

Our pipeline enhances scalability by merging five specialized callers to resolve the high disagreement seen in individual algorithms. By implementing an automated LLM vision system, we eliminate manual curation bottlenecks and standardize the validation of genomic coverage plots. This reproducible framework reconstructs complex SVs, providing a robust foundation for identifying pathway-specific signatures in cancer.

III. Challenges in Linking Clustered Breakpoints

Short-reads offer only indirect evidence of rearrangements and often fails to resolve the precise architecture of clustered breakpoints. While our framework improves discovery, read length limitations still cause ambiguity in complex regions. Long-read sequencing remains essential to directly link adjacent breakpoints and validate higher-order events like BIR.

ACKNOWLEDGMENT

We sincerely thank Dr. Xiaohua Wu for her invaluable research guidance and mentorship within the Wu Lab. We are grateful to Dr. Salvatore Loguercio for technical expertise on pipeline scalability, and to Yashwin Madakamutil and Dr. Alyssa Taylor for their critical feedback and coordination throughout the BENG 187 senior design sequence.

REFERENCES

Hadi, S., et al. (2020). Distinct classes of complex structural variation uncovered across thousands of cancer genome graphs. Cell, 183(1), 197–210.e32. https://doi.org/10.1016/j.cell.2020.08.006
Li, Y., Roberts, N. D., Wala, J. A., Shapira, O., et al. (2020). Patterns of somatic structural variation in human cancer genomes. Nature, 578(7793), 112–121. https://doi.org/10.1038/s41586-019-1913-9
Wu, T., Li, Y., Zhao, Y., Shah, S. B., Shi, L. Z., & Wu, X. (2025). Break-induced replication is activated to repair R-loop-associated double-strand breaks in SETX-deficient cells. Cell Reports, 44(10), 116386. https://doi.org/10.1016/j.celrep.2025.11638