Bioengineering Day Poster Addendum: ABET Summary

Shania Bu

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

1. Desired Needs

- Automation of SV Validation: Replace the slow, subjective, and inconsistent process of manual IGV screenshot curation with an automated system.
- Complex Rearrangement Detection: Develop methods to identify complex events like templated insertion chains (TICs) and rearranged duplications (rDups) often missed by standard callers.
- Standardized Workflow: Create a reproducible, end-to-end pipeline for both lab samples and large-scale analysis of public datasets to enable consistent comparisons across studies.

2. Major Constraints

- Risks: Technical risks include classifier failure on small structural variants and storage limits (10TB) preventing the processing of entire large-scale cohorts.
- Global Impact: Providing a scalable, automated framework allows researchers worldwide to conduct mechanistic DNA repair studies without the bottleneck of manual validation.
- Quality Control/Marketability: Aiming for >90% accuracy and reducing manual review time to ensure the tool is scientifically trusted and adopted by the research community.

3. Engineering Standards

- CLSI MM09: Guideline for human genetic and genomic testing using high-throughput sequencing for assay validation and results reporting.
- GA4GH Standards: Implementation of consistent genomic data formats (BAM, VCF) and secure data sharing protocols.

4. Ethical, Environmental, or Societal Concerns

- Ethical/Societal: Protecting sensitive biological information within human genomic datasets through secure data governance and access control.
- Environmental: Reducing carbon emissions associated with high-performance GPU infrastructure by optimizing LLM prompts and minimizing unnecessary API calls.

5. Active Teamwork and Leadership

- Collaboration: Integrated molecular knowledge from wet-lab assays with computational results through close cooperation with lab members.
- Delegation: I led the simple and complex SV discovery, repair pathway analysis, and long-read processing. Elliot led the automated SV validation tool development and TCGA parsing.
- Goals and Deadlines: Used a 10-week Gantt chart and weekly internal updates to manage HPC queue delays and development milestones.
- Constructive Feedback: Received technical mentorship from Dr. Salvatore Loguercio on pipeline scalability and biological guidance from Dr. Xiaohua Wu.

6. Motivating Factors

- New Knowledge: The high discordance (40-60%) between existing SV callers motivated the development of an ensemble-based discovery engine.
- Self-Initiating: Transitioned from reliance on external TCGA metadata to in-house engineered RPE cell lines to bypass data availability bottlenecks.
- Persistence: Iteratively refined the classifier, eventually switching to Claude Opus 4.6 when previous models failed to accurately resolve small deletion signals .

7. Innovative and Entrepreneurial Ideas

- Automated AI Validation: Replacing subjective human curation with a vision-based AI system that interprets genomic coverage plots with reasoning comparable to experts.
- Integrated Discovery Ecosystem: A "flag-and-verify" framework that identifies candidate loci via short-reads and automatically validates them with targeted long-read sequencing.