Bioengineering Day Poster Addendum: ABET Questions

Project: AI Vision-Based Structural Variant (SV) Classification Pipeline from IGV Screenshots | Shania Bu, Elliott Ou (BINF) | Mentors: Dr. Xiaohua Wu and Salvatore Loguercio, Scripps Research

1. Desired Needs

- Reduce the manual burden of curating IGV screenshots when validating structural variant (SV) calls, currently a slow, manual step
- Provide a reproducible, scoring-based classifier for consistency
- Support multiple SV types (deletions, inversions, translocations, duplications, insertions) within one pipeline to streamline downstream analyses

2. Major Constraints

- Safety: No physical safety concerns
- Risks: Vision model hallucination, LLM may shift focus between each image/prompt slightly different interpretation every iteration, HPC queue delays up to a day per job.
- Manufacturability: Requires access to LLM API and HPC

3. Standards:

- CLSI MM09 (genomic testing using high-throughput NA sequencing): guides assay validation, QC, and result reporting; supports accuracy and reproducibility for long-read SV detection.
- GA4GH standards for genomic data and workflows: consistent data representation, secure sharing, and reproducible analysis across labs.
- ISO/IEC 27001 (information security management): informs secure data handling, access control, and governance for sensitive human genomic data.
- Good Machine Learning Practice (GMLP) for SaMD: principles for dataset management, validation, transparency, and lifecycle monitoring; applied to inform reproducibility of the pipeline.

4. Ethical, Environmental, Societal Concerns
- All genomic data is obtained through cleared, authorized access to controlled public datasets (TCGA BRCA via dbGaP) and lab-generated knockout cell lines; only team members with approved credentials handle the data, in line with NIH
- Outputs are research-use only, not for clinical use
- Cancer-genomics analyses are compute-heavy, will have carbon footprint

5. Active Teamwork and Leadership
- Weekly syncs with mentors (Dr. Xiaohua Wu, Salvatore Loguercio) and labmates
- The pipeline was split into independent subprojects with each member leading their own piece. Elliott led the SV image classifier. Shania led simple SV and complex SV discovery
- Disagreements resolved through opinions of both mentors.

6. Motivating Factors
- New Knowledge: When the project was originally started, there were not a lot of researchers focused on structural variants. Being able to research a new field helped motivate the project.
- Self Initiating: During our analysis with SVs, manual labeling of images to ensure we only analyze true events took a long time. The SV classifier was created to deal with this step.
- Persistence: While some parts of the project were very frustrating, being able to create something rewarding that will save time for not only the Wu lab, but also other labs was motivating.

7. What are your most innovative and/or entrepreneurial ideas for this project
- Long term, this pipeline can be coded into a package that other researchers can freely use.