Bioinformatics Summer School Report

Name: Raghad Ahmed Alkharouby
Internship Organization: The Broad Institute of MIT and Harvard
Supervisor's Name: Ricky Magner
Project title: Quantifying the Effect of Joint Calling

1. Introduction

Humans exhibit a high degree of genetic similarity, with subtle variations that make each of us unique. These genetic variations play a crucial role in shaping diverse phenotypes and influencing susceptibility to various diseases. Variant calling is a fundamental process used to identify these genetic variants from sequence data.

Joint variant calling improves the accuracy of variant detection by analyzing genetic data across multiple samples simultaneously. This method revisits individual variant calls within the context of a larger cohort, providing a more comprehensive view and reducing the risk of errors.

Joint calling leverages population statistics and biological assumptions about inheritance to enhance variant detection accuracy. Genetic variants are primarily passed down from parents, so analyzing variants across a cohort can leverage this underlying biology to better distinguish true variants from sequencing/alignment errors.

Studying genetic variants in a population context is essential for understanding their effects on health and disease. Statistical analysis of cohort data can uncover associations between genetic variants and phenotypic traits, such as disease presence. In contrast, single-sample data may miss critical variants or incorrectly identify artifacts.

A previous experiment (Gauthier et al., 2016 (ASHG))measured the benefits of joint calling and demonstrated that increasing the size of the cohort improves variant sensitivity, with benefits plateauing around 600 samples. [1]

Our goal was to reproduce this experiment and measure benefits of joint calling using modern data and to identify the minimum sample size required to achieve optimal performance in joint variant calling. And to explore how varying population structures within the cohort affect the accuracy of joint variant calling.

2. Materials and Methodology

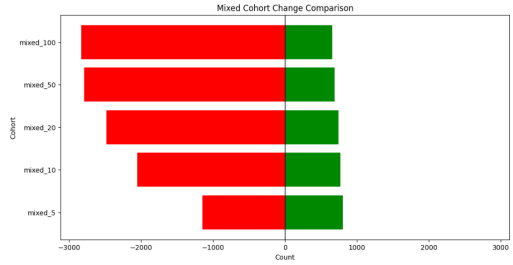
We analyzed genotypic changes by creating cohorts of increasing size from the 1000 Genomes dataset, ensuring that the HG002 sample was included in each cohort. We performed joint calling on these cohorts using GenotypeGVCFs, without applying any filters. This approach enabled us to observe how joint calling affects genotype data.

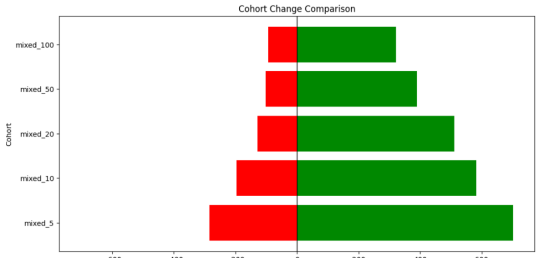
From the joint VCF files, we extracted the HG002 sample and used VCFeval for benchmarking. This process allowed us to categorize variants into true positives (TP), false positives (FP), and false negatives (FN), using the well-established GIAB sample as our reference.

To evaluate the impact of joint calling, we compared the HG002 sample's VCF from single-sample calling against the VCF obtained from the joint calling process. We assessed the changes in variants—whether genotypes were altered or dropped—and how these changes affected accuracy relative to the benchmark. Variants were classified based on their change status: "bad" (where TP became FN) or "good" (where FP became TN).

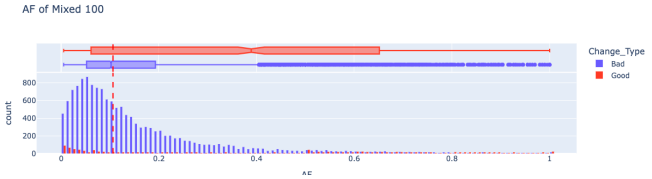
3. Results and Discussion

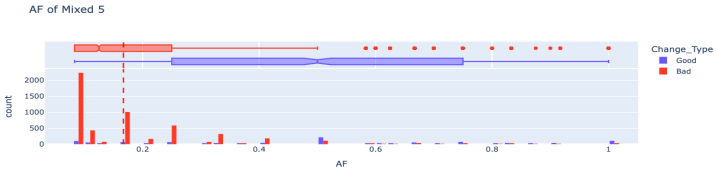
Initially, we observed that larger cohort sizes unexpectedly had a more detrimental effect than beneficial. The left plot, which describes variants that had their genotype changed, shows that as the cohort size increases, the count of bad changes increased. This finding challenges our initial expectations and warrants closer examination. However, the right plot, which describes variants that were dropped, reveals the opposite trend. It shows that as the cohort size increases, the count of good changes increases.



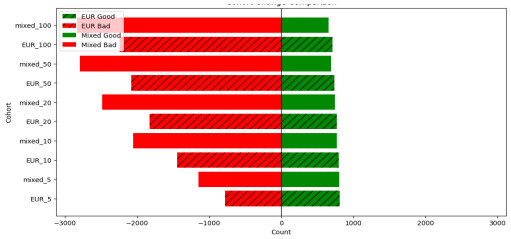


To further investigate, we analyzed the allele frequencies (AF) of the variants that had changed to determine if they were rare variants. Bad changes were enriched for lower AF. One explanation for this phenomenon is that, when working with smaller cohorts, we may be more inclined to retain rare variants. Even if the evidence supporting their existence is limited, there is a chance they will be preserved. However, as we shift to mid-sized cohorts, the landscape changes: rare variants previously retained may now be at risk of being filtered out as potential sequencing artifacts rather than reflecting genuine biology.





We also considered the impact of cohort homogeneity. While similar patterns were observed, the effects were generally less pronounced with lower overall counts. This suggests that a more homogeneous cohort can partially mitigate the negative impact associated with genotyping changes.



Overall, our analysis suggests that re-genotyping often had a net negative rather than positive impact. Specifically, low-allele frequency variants were more frequently associated with adverse changes. However, a more homogeneous cohort appears to alleviate these undesirable effects to some extent.

4. Skills and Knowledge Gained

Through this project, I gained valuable technical skills and knowledge across a range of bioinformatics tools and frameworks. I became proficient in using the Terra platform and writing WDL workflows to automate and scale my analyses. I also expanded my programming expertise, working extensively with Python, pandas, and data visualization libraries like Plotly and Metaplot. I also developed expertise in employing GATK tools, BCFtools, and RTG tools, particularly vcfeval to handle variant calling, genotyping, and benchmarking tasks.

Beyond just the technical competencies, this internship also helped me develop critical soft skills that will serve me well moving forward. Dealing with diverse data types, file formats, and a variety of tools taught me how to quickly adapt and integrate information from various sources. This has significantly improved my ability to look up information and troubleshoot issues.

The most significant takeaway, however, is learning how to learn about new tools, not just how to use specific ones. I have developed a robust framework for independently acquiring knowledge about unfamiliar technologies, which is a crucial skill in the ever-evolving field of

bioinformatics. This adaptability is my new superpower that transcends the specifics of any individual tool, empowering me to approach future obstacles knowing that I no longer just know how to use a specific set of tools - I have developed a versatile skillset for rapidly onboarding and leveraging whichever technologies are necessary.
I also honed my skills in managing the computational and storage requirements of large-scale genomics analyses, learning how to balance performance, cost, and efficiency. Throughout the project, I further developed my communication and collaboration skills, facing and overcoming challenges that enhanced these abilities.

5. Challenges and Solutions

One of the primary challenges we faced was the complexity of managing multiple cohorts and numerous analysis steps throughout the project. Coordinating inputs, outputs, and processing for each cohort while ensuring consistency and reliability proved to be a formidable organizational task.
To overcome this challenge, we utilized the scalable WDL workflows on the Terra platform. Terra's automation capabilities were instrumental in managing inputs and outputs for multiple cohorts simultaneously, significantly reducing the time and effort required.
Despite this, maintaining organization amidst the complexity remained a challenge. Keeping all these moving parts organized and aligned can be an overwhelming task. To address this, we created a comprehensive tracking table that served as a centralized hub. This table listed each cohort (rows) and detailed the various analysis steps (columns), allowing us to quickly assess the status of each cohort and identify which steps were completed and which were still pending. The tracking table also provided a clear record of all inputs and outputs, facilitating the troubleshooting of any issues or discrepancies. The synergy between Terra's automation and our custom tracking table proved to be an effective solution, enabling us to efficiently manage the project's complexity and keep everything aligned.

6. Conclusion

During my time in medical school, I've developed a deep fascination with the field of genetics and how it can unlock new understandings of human health and disease. The opportunity to acquire bioinformatics skills has been particularly valuable, as I recognize their crucial role in modern biomedical research. I'm eager to combine my medical expertise with advanced genomic analysis to uncover new insights that can improve diagnosis, treatment, and prevention of genetic disorders.
As I approach graduation, I am enthusiastic about the prospect of utilizing my bioinformatics expertise and genetic knowledge in my future career as a physician-scientist. My goal is to contribute meaningfully to the advancement of healthcare, leveraging these skills to drive innovation and improve patient outcomes.

7. Acknowledgments

- Ricky Magner
- Kylee Degatano (DSP Variants)
- Aaron Hatcher (DSP Variants)
- Rori Cremer (DSP Variants)
- Megan Shand (DSP Methods)
- Laura Gauthier (DSP Methods)

Special thanks to:

- Sam Lee (DSP Methods)
- Eyad Alqaysi
- Hydro.gen team members
- KAUST Academy
- All DSP interns
- Shadi Zaheri

8. References **(if applicable)**

1. Scaling variant calling up to hundreds of thousands of samples with GATK, Gauthier et al., 2016 (ASHG)