Fast and Cost-effective Cloud-based Pipelines to Analyze Cancer Sequencing Data

David I Heiman[1], Scott Ritterbush[1], **Sam Wiseman**[1], **Samantha Van Seters**[1], Saveliy Belkin[1], David Lehotzky[1], Chunyang Bao[3], Antonia Kowalewski[1,2], Ron Solan[1], Chip Stewart[1], Brian P Danysh[1], Andrew D Cherniack[1], Luis Antonio Corchete Sanchez[1,4], Haruna Tomono[1], Gengchao Wang[1,4], Vasuki Narasimha Swamy[1], Xavi Loinaz[1], Zachary Everton[1], Gang-Hee Lee[3], Won-Chul Lee[3], Hansol Park[3], Ryul Kim[3], Young Seok Ju[3,5], Julian Hess[1], Esther Rheinbay[1,2,4], Gad Getz[1,2,4]

[1]Broad Institute of MIT and Harvard; [2]Harvard Medical School; [3]Inocras Inc.; [4]Massachusetts General Hospital; [5]Korea Advanced Institute of Science and Technology

Background

- Over the course of many years, our lab has developed and rigorously tested standardized computational pipelines for comprehensive analysis of cancer genome data, including whole-exomes and whole-genomes from tumor–normal pairs.
- Our lab has built and maintains **wolF**, a custom workflow manager and cloud orchestration tool leveraging the Google Cloud Platform (**GCP**)
- With decades of experience, our lab offers optimized, cost-effective, and cutting-edge solutions and is open to collaboration

TCGA Characterization

- Our custom pipelines are capable of analyzing various types of data to detect single-nucleotide variants (**SNVs**), allele-specific copy number variants (**CNVs**), and structural variants (**SVs**), while also performing quality control (**QC**), purity/poidy estimation, and other analyses (**Figure 1**)
- After years of development, our characterization pipeline faced its greatest scalability challenge in the **TCGA** dataset, which contains more than 8,000 whole-genome sequenced (**WGS**) tumor-normal pairs
- In collaboration with **Inocras, Inc.**, we are currently validating our findings from our TCGA analysis and performing cohort-level analysis to discover novel mutations associated with 31 cancer types

Characterization Pipeline Overview

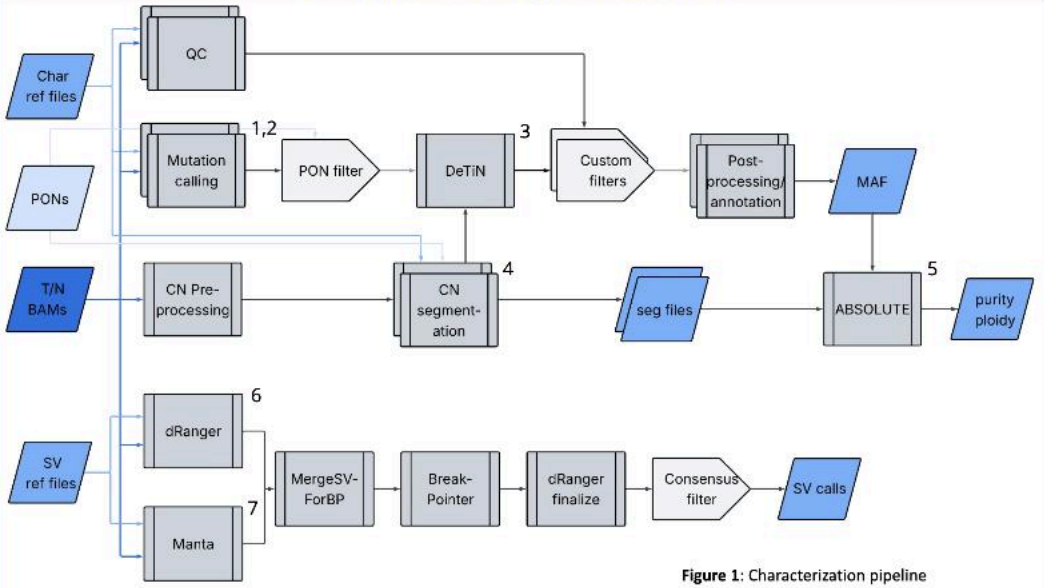

Figure 1: Characterization pipeline

Optimization

- Our optimization dramatically decreased the time and cost of the analysis of the TCGA runs (**Figure 2**)
- Significant improvements came from updates to wolF parameters
- GCP optimizations suggested by expert Google engineers further reduced costs
- wolF's profiling tools provide details about cloud resource consumption, allowing us to identify expensive pipeline steps and plan future optimizations (**Figure 3**)
- Thanks to several months of optimization work, we were able to fully characterize more than 8,000 pairs in less than four weeks in our production run (**Figure 4**)

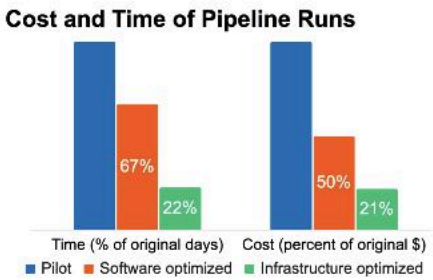

Figure 2: Pipeline optimizations

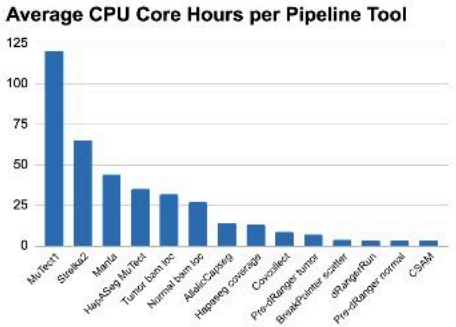

Figure 3: Pipeline task profiling

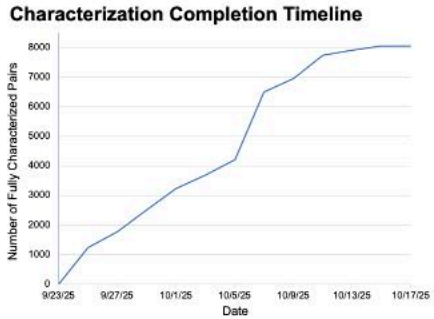

Figure 4: Characterization timeline

Collaborations

- We are seeking scientific collaborations to run the existing pipelines and to refine new ones in large-scale projects
- Visit our website at the below QR code and reach out to bdanysh@broadinstitute.org to learn more about working with us

Future plans

- Neoantigen, fragmentomics, single-cell, and ppmSeq pipelines
- Developing custom pipelines for large-scale collaborations
- Building out tumor-only capabilities
- Automated initial analysis results (significance, signatures, clinical correlations, and more)

Contact

Sam Wiseman & Samantha Van Seters
The Broad Institute of MIT and Harvard
swiseman@broadinstitute.org
svansete@broadinstitute.org
getzlab.org

References

1. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, Gabriel S, Meyerson M, Lander ES, Getz G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Mar;31(3):213-9. doi: 10.1038/nbt.2514. Epub 2013 Feb 10. PMID: 23396013; PMCID: PMC3833702.
2. Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, Chen X, Kim Y, Beyter D, Krusche P, Saunders CT. Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods. 2018 Aug;15(8):591-594. doi: 10.1038/s41592-018-0051-x. Epub 2018 Jul 16. PMID: 30013048.
3. Taylor-Weiner A, Stewart C, Giordano T, Miller M, Rosenberg M, Meerbrey K, Lennon N, Rheinbay E, Livitz D, Wu CJ, Getz G. DeTiN: overcoming tumor-in-normal contamination. Nat Methods. 2018 Jul;15(7):531-534. doi: 10.1038/s41592-018-0036-9. Epub 2018 Jun 25. PMID: 29941870; PMCID: PMC6265093.
4. Prieto O, Solan R, Mouser C, Chu C, Hsu J, Getz G. Haplotype-aware segmentation with **HapASeg** increases accuracy of detecting homolog-specific somatic copy number alterations. Genome Biol. 2026 Feb 14;27(1):88. doi: 10.1186/s13059-026-03971-w. PMID: 41688384; PMCID: PMC12983728.
5. Carter, S., Cibulskis, K., Helman, E. et al. **Absolute quantification of somatic DNA alterations in human cancer**. Nat Biotechnol 30, 413–421 (2012). https://doi.org/10.1038/nbt.2203
6. Erlich Y*, Lawrence MS, Carter SL, Stewart C, Getz G, Lander ES, Meyerson M, Beroukhim R, Getz G. **Somatic recombinogenesis across cancer reveals classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability**. Genome Research 23 (2): 228-35 (2012). (**dRanger**).
7. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. **Manta**: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016 Apr 15;32(8):1220-2. doi: 10.1093/bioinformatics/btv710. Epub 2015 Dec 8. PMID: 26647377.

Acknowledgements

- Kiara Pontious
- Jialin Ma
- Qing Zhang
- Chris Etienne
- Graydon Moorhead
- Oliver Priebe
- Jean-Baptiste Alberge
- David Merrell
- Claudia Chu
- François Aguet

Do Not Post

#6881