Command line to pipeLine: Cross-biobank analyses with Nextflow

Anurag Verma
Department of Pathology and Laboratory Medicine,
University of Pennsylvania
Philadelphia, PA 19104, USA
Email: anurag.verma@pennmedicine.upenn.edu

Zachary Rodriguez
Department of Pathology and Laboratory Medicine,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: zachary.rodriguez@pennmedicine.upenn.edu

Lindsay Guare
Department of Pathology and Laboratory Medicine,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: lindsay.guare@pennmedicine.upenn.edu

Katie Cardone
Department of Genetics,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: katie.cardone@pennmedicine.upenn.edu

Christopher Carson
Department of Pathology and Laboratory Medicine,
University of Pennsylvania, Philadelphia, PA 19104, USA
Email: christopher.carson@pennmedicine.upenn.edu

Biobanks hold immense potential for genomic research, but fragmented data and incompatible tools slow progress. This workshop equipped participants with Nextflow, a powerful workflow language to streamline bioinformatic analyses across biobanks. We taught participants to write code in their preferred language and demonstrated how Nextflow handles the complexities, ensuring consistent, reproducible results across different platforms. This interactive session was ideal for beginner-to-intermediate researchers who want to (1) Leverage biobank data for genomic discoveries, (2) Build portable and scalable analysis pipelines, (3) Ensure reproducibility in their findings, (4) Gain hands-on experience through presentations, demonstrations, tutorials, and discussions with bioinformatics experts.

Keywords: bioinformatics, genomics, phenome, biobanks

1. Introduction, Background, and Motivation

The field of genomics has entered a transformative era fueled by the rapid expansion of biobanks. These repositories, including public entities like the UK Biobank (Sudlow et al., 2015) and the NIH's All of Us Research Program (Ramirez et al., 2022), along with numerous institutional biobanks such as Million Veteran Program (Gaziano et al., 2016) and Penn Medicine BioBank (Penn Medicine BioBank, n.d.), have been instrumental in accelerating genomic discovery at an unprecedented pace. By bringing together extensive collections of biological samples and rich clinical data, biobanks have been a goldmine for medical research. We can leverage biobanks to pinpoint genetic variations linked to diseases and unravel the complexities of various phenotypes.

Despite the move towards cloud computing to share data, biobanks face significant technical hurdles that slow down their potential. Data is often kept in isolated pockets, and researchers have to navigate a technical maze to use different platforms and tools. This not only hinders the speed of research but also leads to the same work being repeated and a mix of data analysis practices that can cast doubt on findings and make it challenging to scale up genomic studies.

As bioinformatic analyses grow in scale and popularity, the methodology and best practices are becoming more standardized. Rather than introducing redundant code by copying commands between projects, pipeline managers offer a way to re-configure the code while recycling it from the same source. Avoiding redundancy is important because each new copy-paste of a section of code results in propagating errors which will take additional time to track and fix (Leitão, 2004). Furthermore, highly parallel computational work on university computing clusters has often looked like manually watching the queue to wait for all jobs of a particular step to finish. Pipeline managers have automated interfaces which work with multiple platforms, allowing them to track jobs and submit dependent ones as they finish.

Workflow languages like Nextflow play a pivotal role in the development of scalable and reproducible genomic pipelines by offering a platform-agnostic framework for seamless data analysis across diverse computing environments (Figure 1). By abstracting the complexities of platform-specific hardware/software configurations, Nextflow enables researchers to focus on the scientific logic of their analyses and interpretation of results. This abstraction allows researchers to create workflows from their pre-existing code written in any language that can be easily deployed on local servers, high-performance computing clusters, or cloud-based platforms without modification. Further, Nextflow’s containerization support through technologies like Docker and Singularity ensures analyses can be deployed and parallelized across different computing architectures without risk of data conflicts, dependency issues, or concurrent data access and processing.

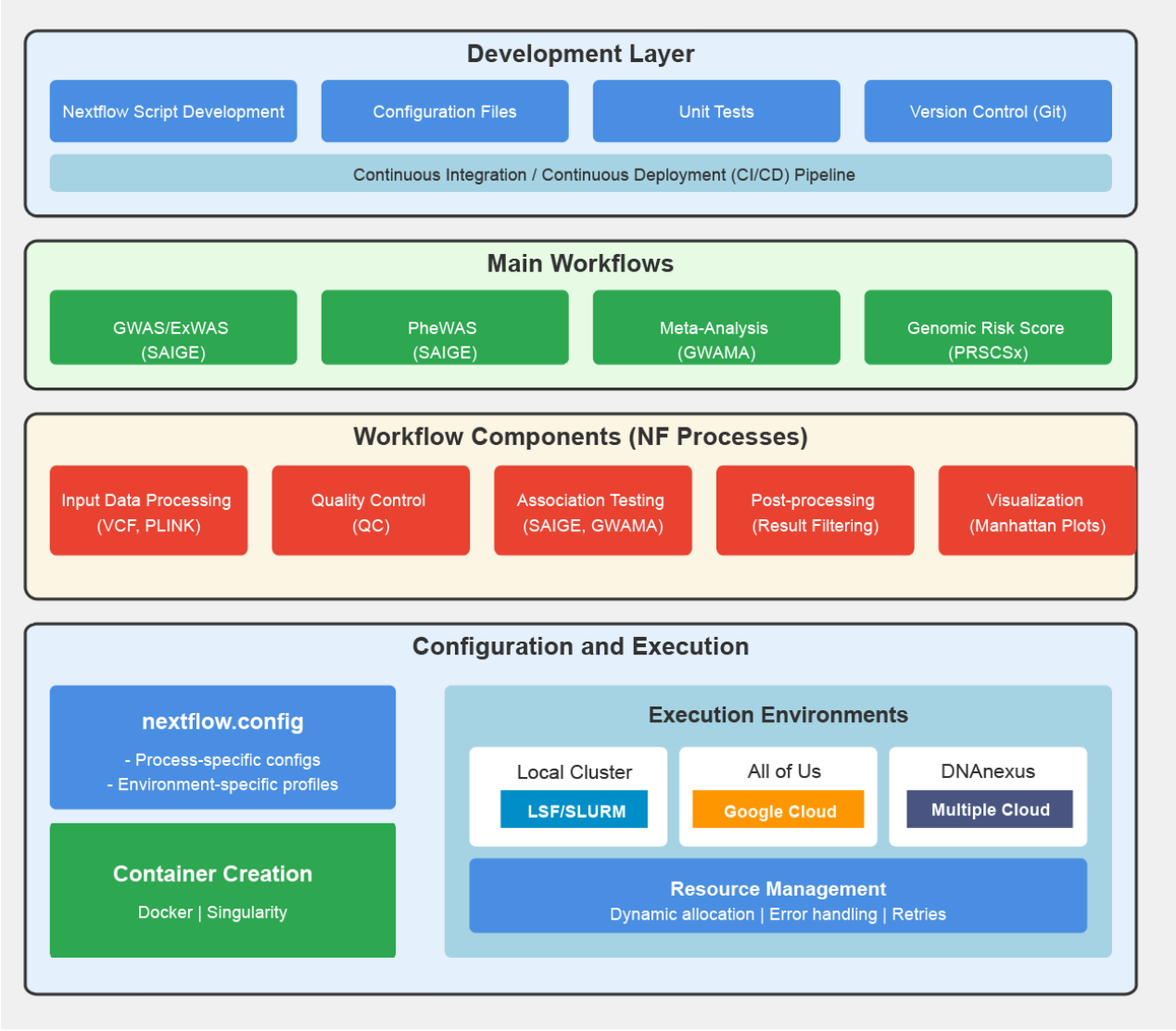

Figure 1. This figure illustrates the multi-layered architecture of genomic workflow development and execution using Nextflow. It encompasses the development layer (including script development, configuration, testing, and version control), main workflows (such as GWAS/ExWAS, PheWAS, Meta-Analysis, and Genomic Risk Score), workflow components (NF processes for various stages of analysis), and the configuration and execution layer (including environment-specific configurations and diverse execution environments). This architecture demonstrates the scalability and flexibility of the workflows across different computing infrastructures, from local clusters to cloud platforms.

Through this workshop, our goal is to address a critical need within the genetic research and bioinformatics community. The rapid expansion of biobank data availability marks a significant milestone in human genetics research, offering unparalleled opportunities to study the genetic predisposition of complex diseases. Although there are platforms and tools for effectively utilizing these datasets for complex, multimodal analysis, there remains an unmet need to develop educational workshops. These workshops are essential to equip participants with the necessary skills and knowledge to fully exploit biobank resources, effectively bridging the gap between the abundance of available data and the capacity for research innovation.

We provided attendees with hands-on workflows to develop and deploy existing tools from institutional biobanks to cloud-based platforms such as the UK Biobank and All of Us. We recognize a strong demand for proficiency in integrating omics data with genetic findings and a growing interest in conducting cross-biobank analyses for more extensive and robust research applications. By focusing on these areas, our workshop directly addresses these educational needs, offering content that builds on past experiences while also anticipating future research trends.

2. Workshop Presentations and Tutorials

To enable communication and discussion between experimental scientists and our expert developers, each module in this workshop included presentations that provided brief introductions to key topics before the demonstrations and hands-on exercises. The goal of these presentations was to educate participants on the foundational principles of developing genomic workflows using existing tools and resources. The workshop format featured demonstrations, hands-on tutorials, exercises, and discussions led by our five Bioinformatics experts. Demonstrations included pre-recorded vignettes showing how to configure and run large-scale genomic pipelines, with step-by-step explanations and Q&A sessions. Hands-on tutorials offered guided introductions to Nextflow workflows, while exercises allowed attendees to practice independently and in group settings, with on-demand assistance from our team. Throughout the workshop, we highlighted our: Case study of analysis on local and cloud platforms such as UK Biobank and All of Us.

- **Genomic Pipelines for Biobanks: Development and Deployment. (Speaker: Anurag Verma)**: Overview of current biobank landscapes; Challenges in developing scalable genomic pipelines;
- **PMBB Toolkit: GWAS and PRS (Speaker: Chris):** Demonstration on how to utilize and understand genome-wide association study and polygenic score pipelines built in our PMBB Genomic Toolkit.
- **Command Line to Pipeline (Speaker: Lindsay and Zach):** Introduction to cloud-agnostic workflow languages with a focus on demystifying Nextflow pipeline management concept so participants can write their own Nextflow pipeline with the help of our experts.
- **Overcoming Limitations of Working Across Biobanks & Cloud Platforms (Speaker: Katie):** Deploying a workflow across cloud environments and coding collaboratively with Google Cloud Shell.

3. Conclusion

Through this workshop, participants gained the essential tools and expertise to harness the full potential of biobank data, ultimately accelerating the pace of genomic research and discovery. By the end of this workshop, participants were equipped with the knowledge and skills to develop and deploy scalable and reproducible genomic workflows, navigate the complexities of cloud-based platforms, and conduct meaningful cross-biobank analyses to advance their research projects. This workshop provided a platform not only as a repository of knowledge but also as a forum for academic exchange. Throughout the workshop, scientists discussed (1) The challenges of conducting bioinformatic analyses across different cloud platforms, (2) Best practices for integrating different biobanks with an emphasis on reproducibility, interpretability, and scalability, and (3) How to use GitHub for transparency, version control, and collaboration.

4. Speakers

Anurag Verma, PhD, University of Pennsylvania. Anurag is an Assistant Professor in the Department of Medicine at the University of Pennsylvania, and he also serves as Associate Director of Clinical Informatics and Genomics for Penn Medicine BioBank. His research has focused on the study of the genetic basis of complex diseases using big data techniques with the main focus on studying the genetic architecture of multimorbidity, the phenotypic architecture of common genetic risk, polygenic risk scores, and phenome-wide association studies to identify the complex phenotypic and genomic interactions that lead to complex disease. In his capacity at PMBB, Anurag leads a team called CodeWorks that develops scalable workflows and harnesses both in-house and cloud computing resources for advancements in genetic research. His team's efforts are in expanding the boundaries of how data informatics can be applied to keep pace with the rapidly changing landscape of large-scale biobanks.

Lindsay Guare, University of Pennsylvania. Lindsay is a second-year PhD student in the Genomics and Computational Biology Program at UPenn with a focus in Biomedical Informatics. She has been involved in many large-scale genetic association study collaborations, but her research will be focused on leveraging innovative computational data science approaches to explore clinical and genetic heterogeneity in endometriosis. Her interdisciplinary background includes computer science, contributing to her leadership in CodeWorks.

Katie Cardone, BS, University of Pennsylvania. Katie is a Research Specialist in the Department of Genetics at the University of Pennsylvania and is a Graduate Student in the University of Pennsylvania's Master of Biomedical Informatics Program. In her role, Katie executes a wide range of bioinformatic analyses, including genome-wide association studies, phenome-wide association studies, exome-wide rare variant association studies, and polygenic scores on large biobanks, including the Penn Medicine BioBank, the eMERGE network, and the All of Us research program. She also develops Nextflow pipelines for polygenic score tools.

Christopher Carson, MS, University of Pennsylvania. Chris is a Bioinformatician at the University of Pennsylvania Institute for Biomedical Informatics. His role in the Verma lab covers an extensive range of workflow pipeline development, conducting genetic analysis requests for the Penn Medicine Biobank (PMBB), and producing bioinformatics software for analyzing large-scale genomic and phenomic datasets. He has experience conducting genome-wide, phenome-wide, and exome-wide association studies using the large-scale datasets retained in the PMBB with the use of SAIGE.

Zachary Rodriguez, PhD, University of Pennsylvania. Zach is a Bioinformatician at the University of Pennsylvania's Perelman School of Medicine. His research has focused on the study of the genetic basis of complex diseases using big data techniques with the focus on studying the genetic architecture of multimorbidity, the phenotypic architecture of common genetic risk, polygenic risk scores, and phenome-wide association studies to identify the complex phenotypic and genomic interactions that lead to complex disease. He has informatics expertise in machine learning, natural language processing, and pipeline development, with extensive experience in analyzing large-scale genomic data, electronic health records (EHR), and biobank datasets, including Penn Medicine BioBank.

5. Acknowledgements

We acknowledge the Penn Medicine BioBank (PMBB) for providing data and thank the patient- participants of Penn Medicine who consented to participate in this research program. We would also like to thank the

Penn Medicine BioBank team and Regeneron Genetics Center for providing genetic variant data for analysis. The PMBB is approved under IRB protocol# 813913 and supported by Perelman School of Medicine at University of Pennsylvania, a gift from the Smilow family, and the National Center for Advancing Translational Sciences of the National Institutes of Health under CTSA award number UL1TR001878. We would like to thank the PMBB leadership team: Daniel J. Rader, M.D., Marylyn D. Ritchie, Ph.D; The PMBB Patient Recruitment and Regulatory Oversight Team: oEllen Weaver, Nawar Naseer, Ph.D., M.P.H., Giorgio Sirugo, M.D., P.h.D., Afiya Poindexter, Yi-An Ko, Ph.D., Kyle P. Nerz; The PMBB Clinifal Informatics Team: Anurag Verma, Ph.D., Colleen Morse Kripke, M.S. DPT, MSA, Marjorie Risman, M.S., Renae Judy, B.S., Colin Wollack, M.S.; The PMBB Genome Informatics Team: Anurag Verma Ph.D., Shefali S. Verma, Ph.D., Scott Damrauer, M.D., Yuki Bradford, M.S., Scott Dudek, M.S., Theodore Drivas, M.D., Ph.D. Lastly, we would like to thank all the bioinformaticians, developers, and testers on our Codeworks team that contributed to the PMBB Geno-Pheno Toolkit Github: Zachary Rodriguez, Lindsay Guare, Chris Carson, Lannawill Caruth, Katie M. Cardone, Aude Ikuzwe, Michael Condiff, Alexis Garofalo, Xueqiong Li, Karl Keat, Rachit Kumar, Trust Odia, Colleen Morse Kripke, Hritvik Gupta, Theodore Drivas, Shefali Setia-Verma, and Anurag Verma.

References

Gaziano, J. M., Concato, J., Brophy, M., Fiore, L., Pyarajan, S., Breeling, J., Whitbourne, S., Deen, J., Shannon, C., Humphries, D., Guarino, P., Aslan, M., Anderson, D., LaFleur, R., Hammond, T., Schaa, K., Moser, J., Huang, G., Muralidhar, S., … O'Leary, T. J. (2016). Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *Journal of Clinical Epidemiology*, *70*, 214–223.

Leitão, A. M. (2004). Detection of Redundant Code Using R 2 D 2. *Software Quality Journal*, *12*(4), 361–382.

Penn Medicine BioBank. (n.d.). Retrieved September 30, 2024, from https://pmbb.med.upenn.edu/

Ramirez, A. H., Sulieman, L., Schlueter, D. J., Halvorson, A., Qian, J., Ratsimbazafy, F., Loperena, R., Mayo, K., Basford, M., Deflaux, N., Muthuraman, K. N., Natarajan, K., Kho, A., Xu, H., Wilkins, C., Anton-Culver, H., Boerwinkle, E., Cicek, M., Clark, C. R., … All of Us Research Program. (2022). The All of Us Research Program: Data quality, utility, and diversity. *Patterns (New York, N.Y.)*, *3*(8), 100570.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779.