Fine Mapping Coronary Artery Disease Loci in UK Biobank's Whole Genome Sequencing Data by Optimising Usage of the Cloud-Based Research Analysis Platform

Letitia M.F. Sng[1*†], Anubhav Kaphle[2*], Mitchell J. O'Brien[1*], Brendan Hosking[1], Roc Reguant[1], Johan Verjans[3,4], Yatish Jain[2,5], Natalie A. Twine[2,5*], Denis C. Bauer[5,6,7*†]

[1]Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Westmead, New South Wales, Australia
[2]Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Melbourne, Victoria, Australia
[3]South Australian Health and Medical Research Institute, Adelaide, South Australia, Australia
[4]Australian Institute for Machine Learning, University of Adelaide, Adelaide, South Australia, Australia
[5]Applied BioSciences, Faculty of Science and Engineering, Macquarie University, Macquarie Park, New South Wales, Australia
[6]Australian e-Health Research Centre, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Adelaide, South Australia, Australia
[7]Department of Biomedical Sciences, Macquarie University, Macquarie Park, New South Wales, Australia.

[*]Authors contributed equally
[†]Corresponding authors

Abstract

We conducted the first comprehensive association analysis of a coronary artery disease (CAD) cohort within the recently released UK Biobank whole genome sequencing dataset. We employed fine mapping tool PolyFun[1] and pinpoint *rs*10757274 as the most likely causal SNP within the 9p21.3 CAD risk locus. Notably, we show that machine-learning (ML) approaches, REGENIE[2] and VariantSpark[3], exhibited greater sensitivity compared to traditional single-SNP logistic regression, uncovering *rs*28451064 in the 21q22.11 risk locus. Our findings underscore the utility of leveraging advanced computational techniques and cloud-based resources for mega-biobank analyses. Aligning with the paradigm shift of bringing compute to data, we demonstrate substantial resource optimisation (44% reduction) and performance gains (40% speedup) through compute architecture optimisation on UK Biobank's Research Analysis Platform and discuss three considerations for researchers implementing novel workflows to such datasets hosted on cloud-platforms. This study paves the way for harnessing mega-biobank-sized data through scalable, cost-effective cloud computing solutions.

Main Text

In November 2023, the UK Biobank (UKB) released the world's largest single set of whole genome sequencing (WGS) data of half a million individuals through their Research Analysis Platform (RAP). In total, 27.5 petabytes of data was created as outlined by UK Biobank WGS consortium[4]. Briefly, the GraphTyper-called[5] WGS data comprised of 1,037,556,156 SNVs and

101,188,713 indels across the 490,640 participants, resulting in 1510 terabytes across 145,207 pVCFs (Supplementary Table 1).

Implementing a new paradigm of data management, the data is stored on the RAP, a cloud-based platform, which allows analysis workflows to be "brought to the data". Registered RAP users with a linked UKB Access Management System (AMS) account can access and analyse datasets through "applets", including the Swiss Army Knife (SAK) app, which holds several bioinformatics tools such as PLINK2[6] and bcftools[7] as well as user-installable tools for custom workflows. Besides the Tier 3 data access fee to UKBB, other costs to the user via DNAnexus include storing uploaded or derived data, egress, and compute using Amazon Web Services (AWS).

Here, we present the first coronary artery disease (CAD) association analysis, leveraging the unprecedented density of the UKB WGS dataset. We discuss how the density of the dataset enables novel insights into CAD genetics and highlight how bringing the compute to the data enables individual-level access for researchers worldwide.

We first investigate chromosome 9 as it includes the well-established CAD locus, 9p21.3 (*CDKN2B-AS1*)[8,9]. Designing purpose-built workflows that are parallelised across pVCFs (Supplementary Material 1; Supplementary Material 2), we confirm the significance of the 9p21.3 CAD risk locus (Supplementary Table 2) with single-SNP logistic regression (sLR) (Figure 1).

The unprecedented density of the UKB WGS dataset enabled fine mapping to identify the most likely causal variant. We applied SuSie[10] and FINEMAP[11] within the PolyFun[1] framework (Supplementary Materials 1) and identify the same 13 SNVs in the first credible sets of both tools (Figure 2; Supplementary Tables 3 and 4). Of these, *rs*10757274 has been linked to *CDKN2B-AS1* expression[12,13], is part of the core risk haplotype region[14], and has been associated with CAD in several populations[15,16].

We also identify the same region and likely causative SNP (*rs*10757274) using the ML GWAS platforms VariantSpark[3] (Figure 1; Supplementary Table 5) and REGENIE[2] (Supplementary Table 6), confirming the increased power of ML on these data sources over sLR.

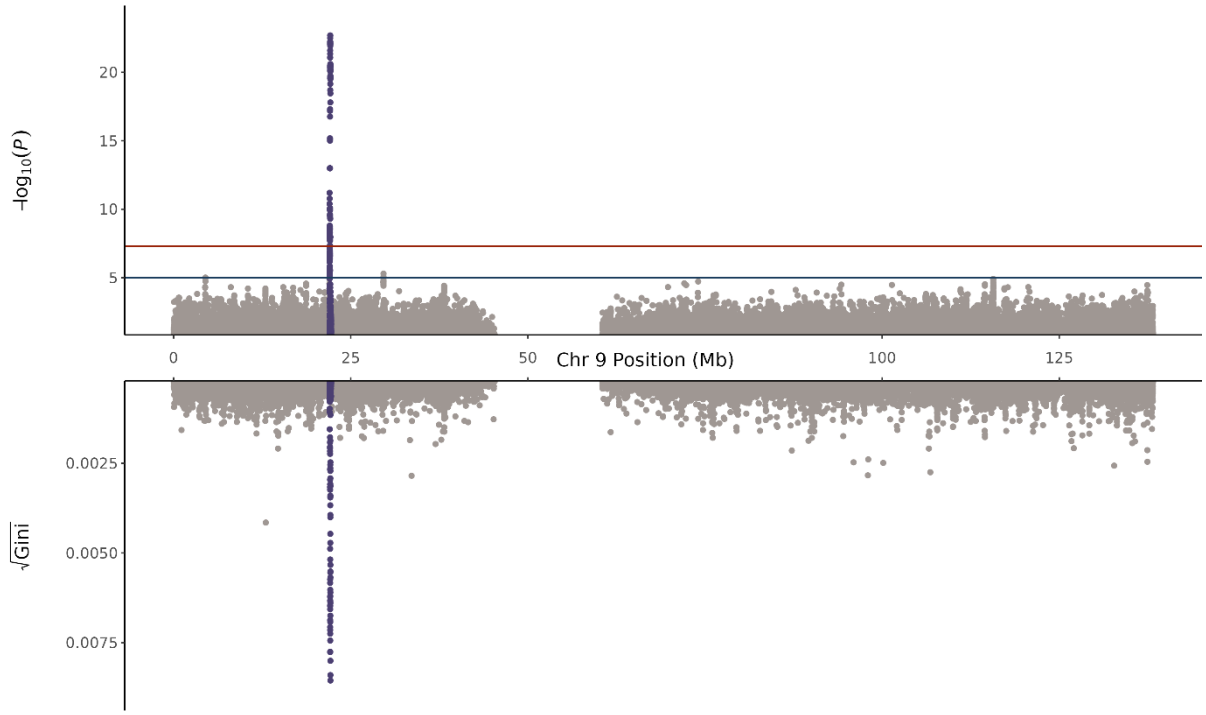

Figure 1. Miami Plot of Chromosome 9 for single-SNP Logistic Regression (Top) and VariantSpark (Bottom). The locus 9p21.3 reached genome-wide significance in the logistic regression analysis and is the top association from VariantSpark. Note that the Gini score (bottom *y*-axis) is used to measure variable importance and while it is analogous to the traditional P-values of logistic regression to rank associations relative to other associations, it does not include multiple testing and does not have guiding significance thresholds.

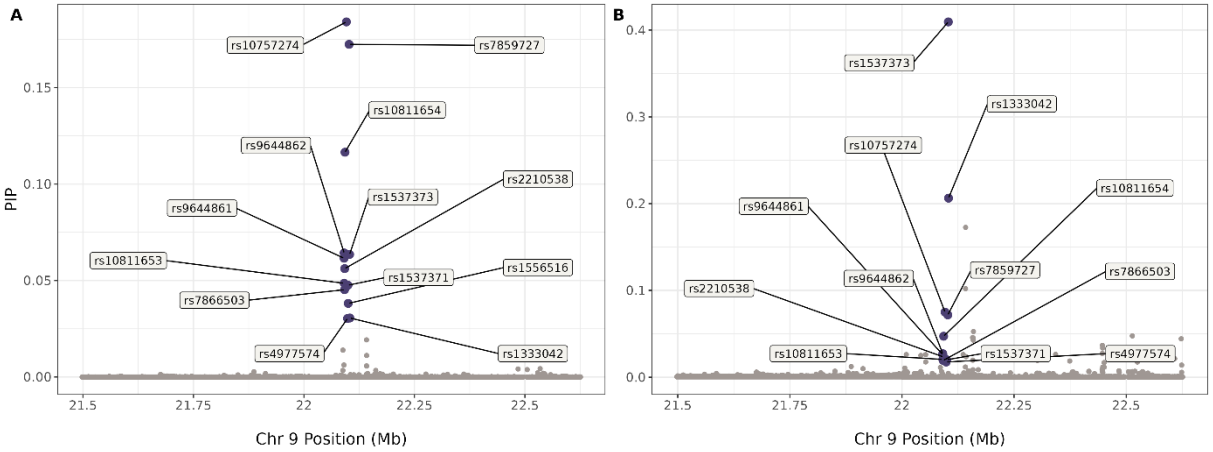

Figure 2. Scatterplot of SNVs in the 9p21.3 locus with Posterior Inclusion Probability (PIP) (y-axis) against Chromosome 9 Position (x-axis) for FINEMAP (A) and SuSie (B). The first credible set of both approaches include the same 13 SNVs which are highlighted in dark purple and annotated with *rs*IDs.

The same pipeline was also applied to chromosome 21, further testing the differences between traditional LR and ML on this new data resource. Figure 3 shows that sLR did not find any significant genome-wide SNVs (Supplementary Table 7) while REGENIE found one significant genome-wide SNV, *rs*28451064 (Supplementary Table 8) in the known CAD locus 21q22.11 (*KCNE2*)[17]. VariantSpark also identifies this SNV as the top association (Supplementary Table 9). The *rs*28451064 SNV is intergenic between the long non-coding RNA (lncRNA), *LINC00310*, and the *KCNE2* gene and has been previously associated with CAD as a putative functional SNP[18,19].

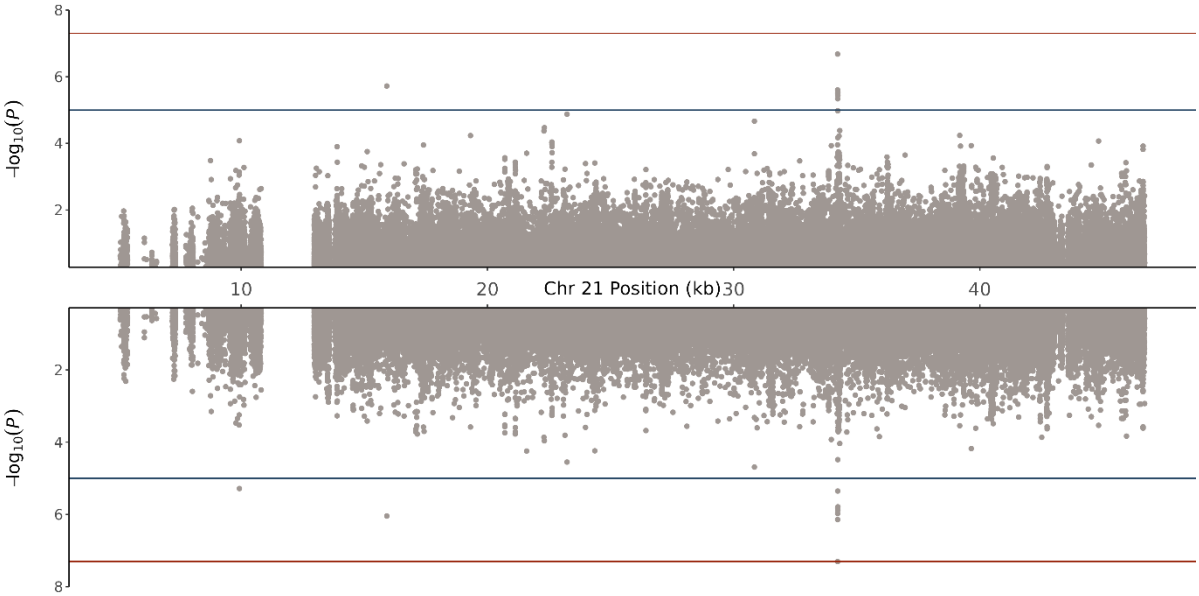

Figure 3. Miami Plot of Chromosome 21 for Logistic Regression (Top) and REGENIE (Bottom). The locus 21q22.11, which has been previously associated with CAD, is only showing marginal association in the single-SNP logistic regression analysis. Conversely, a single SNP in the locus reached genome-wide significance in the REGENIE analysis.

These results highlight that the era of ML in genomics could be ushered in by the unprecedented data volumes, both in cohort size and density, that UKB-style mega-biobanks offer. In particular, the capability to capture complex genome-wide interactions hold great promise for polygenic risk scores[20–23]. VariantSpark may be of particular value here due to its ease of use in identifying the relevant associations in a one step process, compared to REGENIE, which requires an initial genome-wide pre-processing step to first estimate background polygenic effects (Supplementary Material 1), or sLR followed by fine-mapping, which misses associations that lack independent individual association.

UKB's RAP and DNAnexus have pioneered access to large-scale individual-level data by overcoming the risks of moving data, such as corruption, inability to enact consent changes, and lack of standardisation. Supporting this paradigm-shift, we have formulated three considerations for researchers to efficiently process the UKB's resource on RAP and other such datasets (e.g., All Of Us[24]).

Firstly, parallelisation is crucial for efficient computation over the large number of pVCFs in mega-biobank-scale WGS cohorts. Currently, the web UI of RAP and available tutorials are tailored towards processing the genomic data sequentially or in small batches. This was well suited to earlier and smaller versions of the UKB datasets (e.g., array imputed data), however, with 145,207 pVCFs this has become impractical.

To address this challenge, we designed a scalable orchestration engine called RAP parallelisation orchestration engine template (RAPpoet), to automatically parallelise our pipeline across the pVCFs using the DNAnexus CLI (Supplementary Figure 2; Supplementary Materials S2). A 'driver' script configures the cloud environment and co-ordinates the 'worker' scripts on individual compute instances. The worker script then defines the functions that need to be run over the individual files, e.g. quality control, sLR, or ML tasks. Akin to high-performance compute scheduling systems, this worker-driver architecture allows massively parallel workloads while minimising monitoring overhead (centralised coordination). It also enables multitasking within instances, thereby optimising the utilisation of compute resources.

The upcoming inclusion of pre-processed PLINK and BGEN format files in the RAP will collapse individual pVCF files into larger cohort files. While there may be fewer files, the data volume may exhaust compute resources and still requires these monolithic files to be split into distributed workloads, which can be done with RAPpoet.

Secondly, cloud compute architecture needs to be tuned to the workload, file size and time constraints at hand. On the RAP, analyses are carried out on AWS Elastic Compute Cloud (EC2) instances with different options for storage, memory capacity, and number of cores.

In this study, we benchmarked different architectures (Supplementary Material S3). By doing so, we reduced costs by 44% (£0.052 per file vs £0.029) and enhanced CPU utilisation for step one of our pipeline (Supplementary Tables 10 and 11). We also reduced runtime by 40% (2.04 minutes per file vs 1.22 minutes) for step two of our pipeline (Supplementary Tables 12 and 13).

Another important factor of cost optimisation is the risk of information loss when using 'spot' instances (low-priority tasks), which are cheaper but have the risk of external termination compared to the more expensive 'on-demand' instances (high-priority tasks) with full availability. Workflows wanting to use low-priority tasks need to implement strategies such as checkpointing to capture machine images, which allows for the process to resume from the last captured state before termination.

Lastly, data privacy and security must be balanced with the benefits of making individual-level data accessible to researchers, as demonstrated in this study. Implementing federated access, including appropriate governance and seamless integration of dynamic consent layers, requires design-choices from conception[25]. Ideally such systems also enable interactive queries across genomes and meta data without full access to the raw data, such as enabled through sBeacon, the cloud-based implementation of the GA4GH Beacon protocol[26].

In conclusion, this study demonstrates that ML approaches are more sensitive in detecting associations from dense mega-biobank datasets compared to LR. As the field continues to adopt cloud-based platforms like the RAP to exclusively manage access to large-scale datasets, their responsibility will be to equitably support researchers in conducting innovative and technologically complex genomic studies.

Data Availability

Access to the UK Biobank dataset is upon application as described at http://www.ukbiobank.ac.uk/using-the-resource and with permission of UKB's Research Ethics Committee. This study was conducted under the approved application 27483.

Code Availability

Scripts to run RAPpoet, our scalable orchestration engine on DNAnexus CLI for RAP, is available at https://github.com/aehrc/RAPpoet.

References

1. Weissbrod, O. *et al.* Functionally informed fine-mapping and polygenic localization of complex trait heritability. *Nat. Genet.* **52**, 1355–1363 (2020).
2. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
3. Bayat, A. *et al.* VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data. *GigaScience* **9**, giaa077 (2020).
4. Li, S., Carss, K. J., Halldorsson, B. V., Cortes, A. & Consortium, U. B. W.-G. S. Whole-genome sequencing of half-a-million UK Biobank participants. 2023.12.06.23299426 Preprint at https://doi.org/10.1101/2023.12.06.23299426 (2023).
5. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
6. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).
7. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
8. Wei, B., Liu, Y., Li, H., Peng, Y. & Luo, Z. Effect of 9p21.3 (lncRNA and CDKN2A/2B) variant on lipid profile. *Front. Cardiovasc. Med.* **9**, (2022).
9. Patel, R. S. *et al.* Genetic Variants at Chromosome 9p21 and Risk of First Versus Subsequent Coronary Heart Disease Events. *J. Am. Coll. Cardiol.* **63**, 2234–2245 (2014).
10. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **82**, 1273–1300 (2020).
11. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

12. Zhuang, J. *et al.* Methylation of p15INK4b and Expression of ANRIL on Chromosome 9p21 Are Associated with Coronary Artery Disease. *PLOS ONE* **7**, e47193 (2012).

13. Holdt, L. M. & Teupser, D. Long Noncoding RNA ANRIL: Lnc-ing Genetic Variation at the Chromosome 9p21 Locus to Molecular Mechanisms of Atherosclerosis. *Front. Cardiovasc. Med.* **5**, (2018).

14. Broadbent, H. M. *et al.* Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. *Hum. Mol. Genet.* **17**, 806–814 (2008).

15. Xu, L.-B. *et al.* Rs10757274 gene polymorphisms in coronary artery disease: A systematic review and a meta-analysis. *Medicine (Baltimore)* **99**, e18841 (2020).

16. Assimes, T. L. *et al.* Susceptibility locus for clinical and subclinical coronary artery disease at chromosome 9p21 in the multi-ethnic ADVANCE study. *Hum. Mol. Genet.* **17**, 2320–2328 (2008).

17. Lee, S. M., Nguyen, D., Hu, Z. & Abbott, G. W. Kcne2 deletion promotes atherosclerosis and diet-dependent sudden death. *J. Mol. Cell. Cardiol.* **87**, 148–151 (2015).

18. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).

19. Beaney, K. E. *et al.* Functional Analysis of the Coronary Heart Disease Risk Locus on Chromosome 21q22. *Dis. Markers* **2017**, 1096916 (2017).

20. Elgart, M. *et al.* Non-linear machine learning models incorporating SNPs and PRS improve polygenic prediction in diverse human populations. *Commun. Biol.* **5**, 1–12 (2022).

21. Gunter, N. B. *et al.* Machine Learning Models of Polygenic Risk for Enhanced Prediction of Alzheimer Disease Endophenotypes. *Neurol. Genet.* **10**, e200120 (2024).

22. Xu, Y. *et al.* Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. *Cell Genomics* **2**, 100086 (2022).

23. Plagnol, V. Polygenic score development in the era of large-scale biobanks. *Cell Genomics* **2**, 100088 (2022).

24. Bick, A. G. *et al*. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024).

25. Oliva, A. *et al*. Future-proofing genomic data and consent management: a comprehensive review of technology innovations. *GigaScience*, **Accepted**, (2024).

26. Wickramarachchi, A. *et al*. Scalable genomic data exchange and analytics with sBeacon. *Nat. Biotechnol*. **41**, 1510–1512 (2023).

Acknowledgements

We would like to thank and acknowledge DNAnexus and Amazon Web Services for their administrative support in enabling this study.

Author Contributions

L.M.F.S, N.A.T., and D.C.B conceived the study. L.M.F.S, A.K., M.J.O., and B.H designed the study, analysed the data, and interpreted the results. L.M.F.S, A.K., and M.J.O. wrote the first draft. All authors reviewed, commented on, and approved the final version of the manuscript.

Competing Interests

All authors have no competing interests to declare.