OPEN ACCESS

COMMENTARY

The Innovation Medicine

Discover novel biomarkers using population-scale whole-genome sequencing

Jingyi Zhao[1,2] and Sipeng Shen[1,2,3,*]

[1]Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing 211166, China

[2]Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Jiangsu Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing 211166, China

[3]Changzhou Institute for Advanced Study of Public Health, Nanjing Medical University, Changzhou 213022, China

*Correspondence: sshen@njmu.edu.cn

Received: September 15, 2025; Accepted: November 3, 2025; Published Online: November 5, 2025; https://doi.org/10.59717/j.xinn-med.2025.100168

Citation: Zhao J. and Shen S. (2025). Discover novel biomarkers using population-scale whole-genome sequencing. *The Innovation Medicine* **3**:100168.

Genetic studies have long sought to unravel how inherited variation influences human health and disease. Genome-wide association studies (GWAS) have identified thousands of common variant loci, while exome-wide association studies (ExWAS) have implicated rare coding mutations such as BRCA1, BRCA2, CHEK2, and ATM in cancer predisposition. Yet these efforts collectively illuminate only a narrow segment of the genomic landscape: over 95% of human genetic variation resides in the non-coding genome, and the vast majority of variants are rare. Their role in complex disease has remained largely inaccessible due to the limited resolution of single-nucleotide polymorphism (SNP) arrays and exome capture.

A turning point arrived when the UK Biobank (UKB) released whole-genome sequences (WGS) for 490,640 participants.[1] This unprecedented dataset—by far the largest short-read WGS resource assembled—captures both coding and non-coding regions, including common and rare single-nucleotide variants (SNVs), and structural variants (SVs), offering a panoramic view of the human genome. With more than 1.5 billion variants catalogued, UKB WGS provides the statistical power and phenotypic breadth to connect previously invisible forms of variation to human health outcomes. It is no exaggeration to say that this release marks the beginning of a new era in population genomics (Figure 1).

WHAT WE CAN NOW SEE

The expansion from arrays and exomes to population-scale WGS shifts discovery beyond common, well-tagged variants toward a richer spectrum of potential biomarkers. WGS uncovers rare and ultra-rare SNVs, SVs, and somatic events that were previously inaccessible. These classes of variants are not merely incremental additions; they often carry larger effect sizes, display ancestry-specific allele frequencies, and provide sharper mechanistic hypotheses.

Examples from recent studies underscore these possibilities. A multi-ancestry sequencing-based GWAS of C-reactive protein (CRP) across more than 513,000 genomes identified over a hundred independent signals, implicated 151 genes—55 of them novel—and highlighted functional enrichment in blood tissues and epigenomic marks.[2] The study demonstrated how cross-ancestry fine-mapping can narrow causal intervals and sharpen biological interpretation. Similarly, a large ExWAS of lung cancer revealed novel deleterious missense variants in TET3 and POT1. These variants not only carried substantial effect sizes but also connected genetic risk to intermediate traits such as telomere length and emphysema, illustrating the translational value of sequencing-based discovery.

UKB WGS itself expands this repertoire further. Each individual carries, on average, more than 13,000 SVs, with rare SVs tending to be longer and more likely to disrupt functional elements. In addition, somatic mutations in blood cells—clonal hematopoiesis of indeterminate potential (CHIP)—can now be studied systematically in hundreds of thousands of individuals, linking genomic mosaicism to cardiovascular disease and aging. Together, these findings show that population-scale WGS not only catalogs variation but also reveals entirely new categories of biomarkers that extend beyond common SNPs.

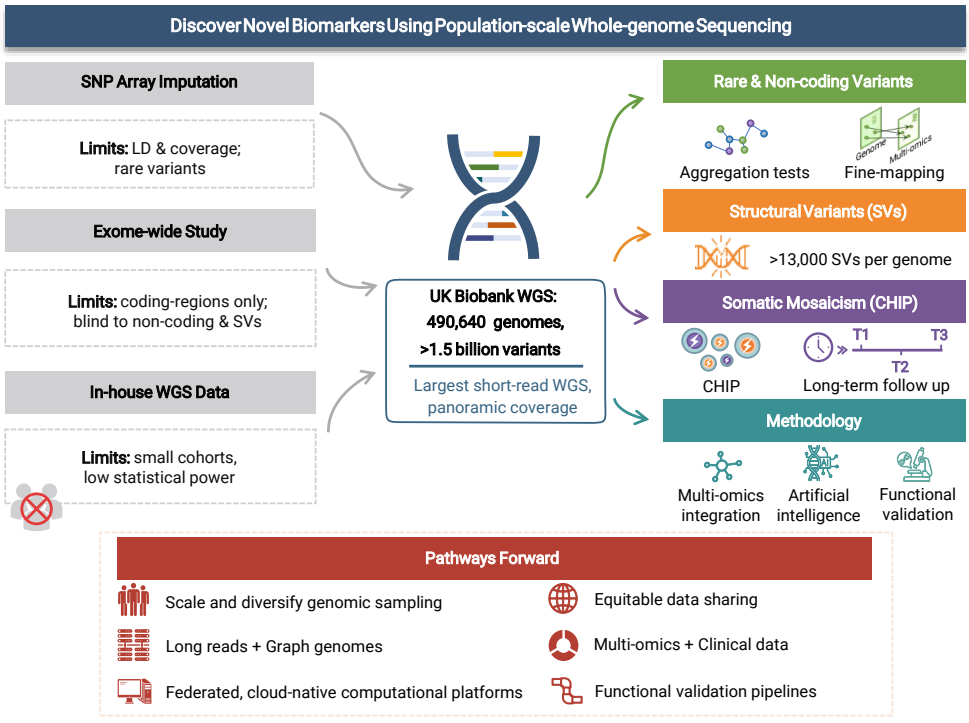

Figure 1. A framework for novel biomarker discovery using population-scale whole-genome sequencing. The diagram illustrates the paradigm shift from traditional genomic methods to the large-scale analysis enabled by UK Biobank whole-genome sequencing (WGS) project Left: limitations of previous approaches—SNP-array–based GWAS/imputation, Exome-Wide Association Studies, and small in-house WGS—restricted largely to common or coding variants, with constrained sample sizes, reduced power, and blind spots for non-coding regions and structural variation. Right: variant classes and analytic strategies now enabled at scale: rare and non-coding variants, structural variants, and somatic mosaicism. Methodology highlights the necessary evolution in analytics, including the integration of multi-omics data, artificial intelligence, and functional validation to derive biological mechanisms. Bottom: Pathway forward that combines large, diverse cohorts, adopt long-read sequencing and graph-based references, integrate multi-omics with clinical data, leverage cloud-native computing, and build robust functional-validation and equitable data-sharing frameworks.

HOW WE MUST ANALYZE DIFFERENTLY

The scale and diversity of WGS demand analytical innovation. Rare variants pose particular statistical challenges: single-variant analyses often lack power, necessitating gene-based methods such as burden tests and frameworks such as variant-Set Test for Association using Annotation infoRmation (STAAR).[3] These approaches aggregate signals across a functional unit, increasing sensitivity to detect modest but biologically important effects of rare alleles.

Multi-ancestry cohorts add further depth. Cross-ancestry fine-mapping, as illustrated in the CRP study,[2] leverages differences in linkage disequilibrium to refine causal inference and improve generalizability. UKB, despite its size, remains predominantly European in ancestry; thus, integration with other large-scale initiatives such as TOPMed and All of Us is essential for equitable discovery.

A complementary strategy is imputation using large-scale sequencing reference panels. Recent work has shown that imputing rare variants into array-based cohorts dramatically increases effective sample size for rare variant association studies.[4] This synergy between WGS and imputation maximizes discovery potential in contexts where disease cases remain scarce.

Beyond association testing, WGS invites functional interpretation. Integration with transcriptomic, proteomic, and epigenomic data helps connect statistical signals to biological mechanisms. Artificial intelligence (AI) and machine learning (ML) now enhance causal inference and gene prioritization.

However, computational inference alone is insufficient to establish causality. Functional validation represents the next critical layer of analysis. Genome editing technologies such as CRISPR screens can experimentally test gene–trait hypotheses at scale, while organoid and induced pluripotent stem cell models enable the evaluation of variant effects in relevant tissue contexts. Single-cell perturbation assays and spatial transcriptomics now allow causal variants to be mapped onto specific cellular states. Together, these complementary experimental systems ensure that sequencing discoveries can move from statistical signals toward mechanistic and therapeutic understanding.

WHAT REMAINS INVISIBLE

Despite these advances, critical gaps remain. Case numbers for many diseases within UKB are still limited, restricting power for rare variant discovery. SVs are incompletely resolved by short-read data; repetitive regions, inversions, and complex rearrangements remain elusive. The emerging Human Pangenome Reference Consortium and graph-based genome assemblies offer promising solutions to capture structural diversity across ancestries and reduce reference bias. Complementary long-read sequencing technologies, including PacBio HiFi and Oxford Nanopore, are increasingly indispensable for comprehensive variant detection (see: https://github.com/lts4jinnie/TIME_COMMENTARY/blob/main/TIME_COMMENTARY_V6.pdf).

Computational demands are equally formidable. Managing petabyte-scale datasets requires sustained investment in cloud-based infrastructure, high-performance computing, and optimized pipelines for alignment, variant calling, and annotation. Recent developments have introduced cloud-optimized data formats and workflow systems for UKB WGS,[5] substantially reducing storage burden and enabling scalable, distributed analysis."

Ethical and governance considerations represent another frontier. Population-scale genomics amplifies concerns about privacy, data ownership, and secondary use of sensitive health information. Informed consent frameworks must evolve to allow broad yet secure data sharing while protecting participant autonomy. Transparent governance structures, federated data access models, and continuous community engagement are essential to maintain public trust in large-scale genomic initiatives.

Moreover, lack of ancestral diversity persists—over 90% of UKB participants are of European descent—limiting generalizability. Multi-ancestry efforts such as the CRP study demonstrate the scientific and ethical necessity of inclusivity. To address this, gnomAD v4, the Japan Biobank, and national genomic efforts in China have significantly expanded coverage of underrepresented populations, collectively improving variant interpretation and disease relevance worldwide. These initiatives complement Western-led resources and exemplify the global shift toward equity in genomic medicine.

THE PATH FROM GENOMICS TO MEDICINE

The release of UKB WGS heralds a new phase in human genomics, but realizing its full promise will require coordinated advances. Translational examples already illustrate feasibility. The variant rs371858405 on chromosome 10 has been linked to reduced risk of hypothyroidism, suggesting protective alleles can inform disease prevention strategies. Similarly, germline PCSK9 loss-of-function variants identified through sequencing studies have guided the development of PCSK9 inhibitors for hypercholesterolemia—one of the most successful stories of genomic translation. In oncology, pathogenic BRCA1 carriers now benefit from PARP inhibitor therapy, exemplifying how rare variant discovery reshapes treatment paradigms. Furthermore, CHIP-associated mutations (DNMT3A, TET2) discovered in blood sequencing are emerging biomarkers for cardiovascular and hematologic risk stratification.

To advance this continuum, three directions are pivotal. First, multi-omics integration provides the functional context to interpret genetic associations across transcriptomic, proteomic, and metabolomic layers. Second, AI-driven frameworks can accelerate causal inference and drug target identification by predicting variant pathogenicity and network effects. Third, functional and clinical validation pipelines—ranging from CRISPR perturbation screens to organoid modeling and high-content phenotyping—must be systematically incorporated to transform associations into interventions.

SUMMARY

In summary, the UK Biobank's 490,640 whole genomes constitute a landmark achievement, shifting human genetics from catalogues of variants to comprehensive maps of human diversity. Translating this map into mechanism and medicine will require innovation, validation, and above all, collaboration.

REFERENCES

1. Carss K., Halldorsson B. V., Hou L., et al. (2025). Whole-genome sequencing of 490,640 UK Biobank participants. *Nature* **645**:692–701. DOI:10.1038/s41586-025-09272-9
2. Li H., Zhao J., Dai J., et al. (2025). Multi-ancestry sequencing-based genome-wide association study of C-reactive protein in 513,273 genomes. *Nat. Commun.* **16**:3892. DOI:10.1038/s41467-025-59155-w
3. Li X., Li Z., Zhou H., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**:969–983. DOI:10.1038/s41588-020-0676-4
4. Dai J., Zhang Y., Gao Y., et al. (2025). Boosting the power of rare variant association studies by imputation using large-scale sequencing population. *Genomics Proteomics Bioinformatics* **17**:qzaf084. DOI:10.1093/gpbjnl/qzaf084
5. Li X., Wood A. R., Yuan Y., et al. (2025). Streamlining large-scale genomic data management: Insights from the UK Biobank whole-genome sequencing data. *Cell Genom.* **18**:101009. DOI:10.1016/j.xgen.2025.101009

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (82373685 to S.S.), Open Research Fund Program of Changzhou Institute for Advanced Study of Public Health, Nanjing Medical University (No. CPHM202301 to S.S.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript

AUTHOR CONTRIBUTIONS

JZ drafted the manuscript, SS supervised the work. All authors contributed to and approved the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.