

Toward building a comprehensive human pan-genome: The SEN-GENOME project

Amadou Gaye,^{1,6,*} Andrea Regina G. Sene,^{2,6} Macoura Gadjji,³ Alioune Deme,⁴ Aynina Cisse,⁵ and Rokhaya Ndiaye^{2,*}

Summary

The human reference genome (GRCh38), primarily sourced from individuals of European descent, falls short in capturing the vast genetic diversity across global populations. Efforts to diversify the reference genome face challenges in accessibility and representation, exacerbating the scarcity of African genomic data crucial for studying diseases prevalent in these populations. Sherman et al. proposed constructing reference genomes tailored to distinct human sub-populations. Their African Pan-Genome initiative highlighted substantial genetic variation missing from the GRCh38 human reference genome, emphasizing the necessity for population-specific genomes. In response, local initiatives like the Senegalese Genome project (SEN-GENOME) have emerged to document the genomes of historically overlooked populations. SEN-GENOME embodies community-driven decentralized research. With meticulous recruitment criteria and ethical practices, it aims to sequence 1,000 genomes from 31 ethnolinguistic groups, in the fourteen administrative regions of Senegal, fostering local genomic research tailored to the region. The key to SEN-GENOME's success is its commitment to local governance of data, capacity building, and integration with broader pan-genome projects in Africa. Despite the complexities of data harmonization and sharing, our collaborative efforts are aligned with common goals, ensuring steady progress toward a comprehensive human pan-genome. We invite and welcome collaboration with other research entities to achieve this shared vision. In summary, local initiatives such as SEN-GENOME are pivotal in bridging genomic disparities, offering pathways to equitable and inclusive genomic research. Collaborative endeavors guided by a collective vision for human health will propel us toward a more encompassing understanding of the human genome and better health through genomic medicine.

Background

The human reference genome (GRCh38), constructed by the Human Genome Project and maintained by the Genome Reference Consortium (GRC),^{1,2} is derived from a mosaic of sequences obtained from 20 anonymous volunteers primarily of European descent. This single-linear sequence, which is predominantly from one of the five individuals, does not fully capture the genetic diversity present in global populations. Nonetheless, it has served as a foundational resource for the scientific community, facilitating countless discoveries and advancements in genomic research. Despite efforts to enhance the diversity of the reference genome by incorporating approximately 998 alternate contigs and 472 scaffolds (chained contigs) into the current human genome assembly,^{2,3} these additions are frequently overlooked in genetic ana-

lyses, even though they can overlap with genomic loci important for disease susceptibility. Furthermore, the process of integrating this diversity is not clear and is primarily managed by the GRC.

Centralized human pan-genome projects^{4–8} have played a pivotal role in cataloging genetic diversity on a global scale. However, despite their utility, these initiatives face inherent limitations, particularly in accessing and representing diverse and isolated populations, with African populations being significantly underrepresented in genomic data compared to those of European or East Asian ancestry. This disparity arises from a combination of factors, including limited resources and funding for genomics research in African countries. As a consequence, there is a notable paucity of data on the genetic diversity of African populations, hindering

efforts to study the genetics of diseases that disproportionately affect these populations and accurately identify genetic risk factors for such diseases.

To address the lack of diversity in the current reference genome, Sherman et al. proposed a solution in 2019 that involves assembling “reference genomes for human sub-populations.”⁴ They constructed an African pan-genome (APG) using data from 910 deeply sequenced individuals from African diaspora populations. Their analysis revealed that approximately 10% of the genome of African diaspora populations is missing from the standard human reference genome.⁴ These results underscore the need for reference genomes that accurately capture the genetic diversity of distinct human populations. For example, research indicates that individuals who are ultrarapid metabolizers of the enzyme CYP2D6, including certain African populations,

¹Department of Integrative Genomics and Epidemiology, School of Graduate Studies, Meharry Medical College, Nashville, TN, USA; ²Division of Human Genetics, Faculty of Medicine, Pharmacy and Odontology, University Cheikh Anta Diop, Dakar, Senegal; ³Service of Biological Hematology & Oncology-Hematology, Faculty of Medicine, Pharmacy and Odontology Stomatology, University Cheikh Anta Diop, Dakar, Senegal; ⁴Laboratory of Prehistory and Cultural Heritage, Department of History, Faculty of Arts and Humanities, University Cheikh Anta Diop, Dakar, Senegal; ⁵Senegalese National Academy of Science and Technology, Dakar, Senegal

⁶These authors contributed equally

*Correspondence: amadou.gaye@mmc.edu (A.G.), rokhaya9.ndiaye@ucad.edu.sn (R.N.)
<https://doi.org/10.1016/j.ajhg.2024.08.017>.

© 2024 American Society of Human Genetics. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.



convert codeine to morphine at a faster rate, increasing the risk of opioid toxicity.⁹ Initially, this discovery, which was predominantly based on studies involving European populations, did not consider the significant differences in drug metabolism among various ethnic groups. This oversight resulted in adverse effects when standard codeine dosages were administered to these populations. Similarly, the effectiveness of the antiplatelet medication clopidogrel is greatly impacted by the CYP2C19 genotype.¹⁰ Research has revealed that patients with specific genetic variants that are more prevalent in non-European populations such as East Asians and Africans exhibit a diminished response to the drug. Consequently, these patients can experience higher rates of cardiovascular incidents.¹¹ This issue was initially missed because most pharmacogenomic research concentrated on European populations. Sherman et al. emphasized the importance of creating reference genomes for all human populations, as variations in the missing regions may not be detected by studies using the current human reference genome.

The SEN-GENOME project

In light of these challenges and recommendations, there is a compelling need for local initiatives to complement centralized efforts and bridge the gaps in genomic representation. One such initiative is the Senegalese Genome project, SEN-GENOME. By focusing on underrepresented populations and leveraging local expertise and resources, SEN-GENOME will address specific health challenges that disproportionately affect Senegalese populations, such as sickle cell disease,^{12,13} hypertension,^{14,15} and diabetes.^{16,17} These conditions can be better managed through genetic research and precision medicine tailored to the unique genetic profiles of these populations. Moreover, SEN-GENOME will enhance genetic diversity assessment, ensuring that healthcare benefits are more equitably distributed. This initiative will ultimately contribute to building a more

comprehensive and inclusive human pan-genome, advancing our understanding of human genetic variation, facilitating precision medicine initiatives, and promoting health equity on a global scale.

SEN-GENOME endeavors to map the genomes of Senegalese populations, including groups absent in current genomic initiatives, offering insights into the region's genetic landscape. Its primary aim is to construct a reference genome tailored to the Senegal population, fostering local genomic research and aiding precision medicine efforts tailored to the local population.

SEN-GENOME, represents a grass-roots effort in decentralized genomic research, providing a model for African scientists to document the genomes of historically overlooked populations. Supported by the Senegalese National Academy of Sciences and Technology and housed within the Division of Human Genetics at Université Cheikh anta Diop (Dakar, Senegal), SEN-GENOME is a community-driven project with a practical approach. Our community-based prospective study recruited 1,015 individuals (452 men and 563 women) aged 18 to 92 years, representing 31 ethnolinguistic groups (Figure 1) documented by the Senegalese National Statistics and Demography Agency. Recruitment took place between November 2022 and May 2023, encompassing 80 predominantly mono-ethnic villages and the city of Dakar, across Senegal's 14 administrative regions. The inclusion criteria required participants to reside in predominantly mono-ethnic villages for at least 18 years, have native ancestry across at least three generations from the same village, have no known health issues, and be able to provide written informed consent. Phenotypic data were collected using a questionnaire administered via tablets with KoboToolbox software (<https://www.kobotoolbox.org/>). The questionnaire covered socio-demographic parameters, health, behavior, lifestyle data (smoking status, alcohol consumption, physical activity, sleep

quality), and medical history. For each participant, two blood samples were collected in EDTA tubes for DNA extraction, blood cell count, and plasma storage. Additionally, blood samples were collected in dry tubes for lipid parameter analysis and serum storage. All biological samples are securely stored at -80°C . Anthropometric parameters measured included weight, height, waist circumference, and hip circumference. Body mass index (BMI) was automatically calculated using the formula: $\text{BMI} = \text{weight (kg)} / \text{height (m)}^2$. We also calculated waist-to-height ratio and waist-to-hip circumference ratio. We also measured physiological parameters including blood pressure, heart rate, and peripheral temperature.

SEN-GENOME's approach emphasizes local engagement and ethical research practices, involving village community leaders and utilizing local languages to communicate project objectives and obtain consent. This community-driven science model presents tangible benefits for African researchers and populations. However, due to budget constraints, SEN-GENOME seeks partnerships for the deep DNA sequencing of collected samples.

Central to SEN-GENOME is the empowerment of local governance of genomic data, as detailed in the ethics, governance, data security, and data use policies (detailed in the next section). This will be realized through hosting data locally at the national high-performance computing (HPC) cluster in Diamniadio, nurturing indigenous leadership, and actively engaging local scientists in the research process. Consequently, any partnerships must align with this fundamental approach. Additionally, the project prioritizes capacity building by providing hands-on training to young and junior researchers, involving them in every step of the research process from participant recruitment to data analysis and hence ensuring sustainable expertise within the region. Moreover, SEN-GENOME aims to contribute to the broader APG Project and complement and extend other initiatives such as the South African Human

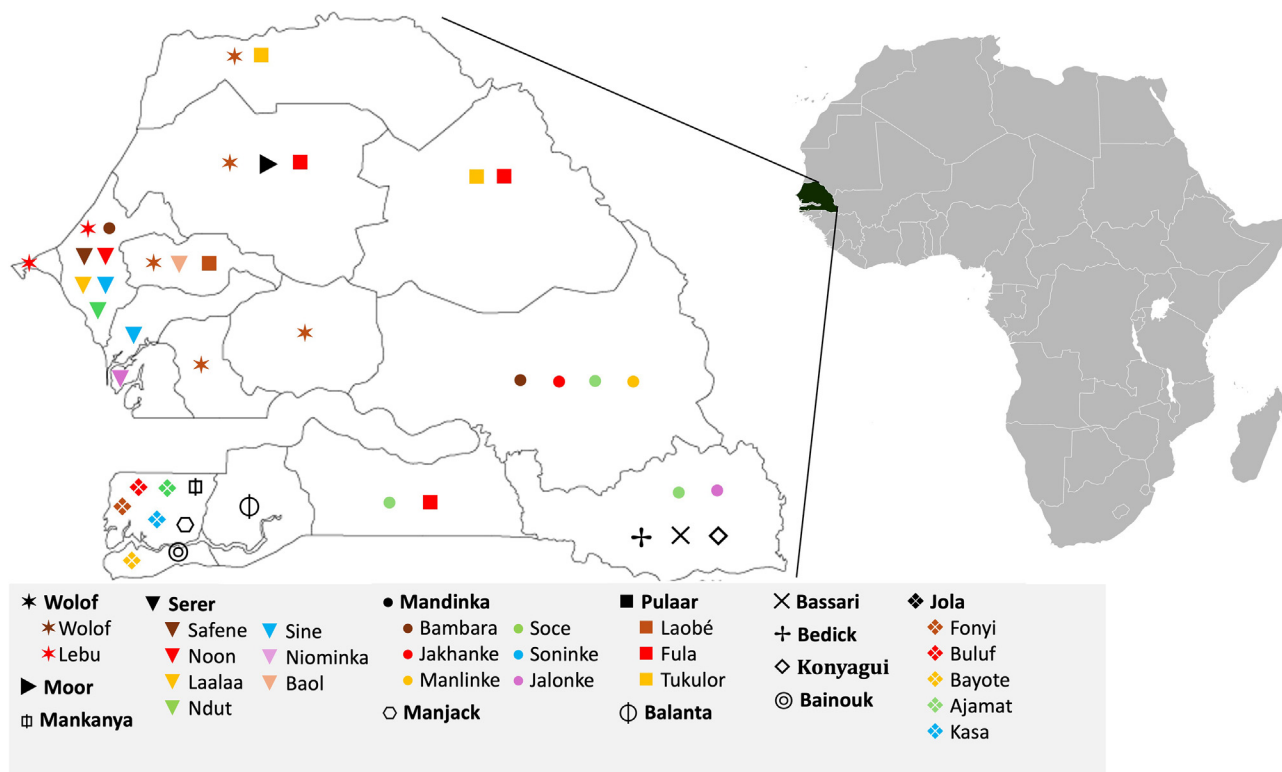


Figure 1. Administrative regions of Senegal where samples were collected

The SEN-GENOME project sampled ethnolinguistic groups across all fourteen administrative regions of Senegal (West Africa), with recruitment conducted within each region. It should be noted that the map labels correspond to administrative regions and not the exact geographical location of the communities. The sampling strategy mirrors the denser population concentration in the eastern regions of the country.

Genome Programme,¹⁸ the Nigerian 100K Genome Project,¹⁹ and the north African (Tunisia, Morocco, and Algeria) initiative PerMediNA.²⁰

The HPC located in Diamniadio, which was the third highest performing computer cluster of its kind in Africa at its acquisition in 2019, plays an important role in this endeavor. The facility has 10 terabytes of user data storage, 1.1 petabytes of scratch storage, and uses a high-performance InfiniBand Enhance Data Rate interconnection network for data archiving. This TAQUEY supercomputer boasts a total of 246 nodes across Skylake (SKL; 80 GB of RAM per node), Knights Landing (KNL; 128 GB of RAM per node), and Graphics Processing Unit (GPU; 80 GB of RAM per node) partitions. It provides high-speed, low-latency connectivity for compute nodes and storage systems with a bandwidth of up to 100 Gbps per link. This substantial computational power and storage capacity

support large-scale genomic data processing and analysis (<https://cineri.sn/carasteriqtiques/>).

While local initiatives like SEN-GENOME offer immense potential, they also face inherent challenges to establishing a global human genome reference, particularly in data sharing, data integration, and ethico-legal and social barriers. Different initiatives may employ diverse protocols, consent and strategies, complicating efforts for seamless data harmonization. Additionally, effective data sharing requires robust infrastructure and a willingness to contribute to the common good, which may be hindered by potential competing interests.

However, these challenges are not insurmountable. The global scientific community is moving toward harmonized platforms and standards, facilitating data integration and interoperability. Collaborations driven by shared research questions and goals incentivize data sharing, fostering a

win-win scenario for all stakeholders involved.

In conclusion, local initiatives such as SEN-GENOME are indispensable building blocks toward achieving a comprehensive human pan-genome. By addressing the unique genomic diversity of underrepresented populations and empowering local communities, these initiatives pave the way for equitable and inclusive genomic research. As we navigate the challenges of data integration and sharing, collaborative efforts guided by a shared vision of advancing human health will propel us toward a more comprehensive understanding of the human genome.

SEN-GENOME's compliance and data management policy

A comprehensive informed consent (see [supplemental note](#)) is administered to participants in their native language. The participants are provided with detailed information about the

study, its purpose, procedures, risks, and benefits. The consent forms are reviewed and approved by an independent ethics review board of the university Cheikh Anta Diop. The participants have the right to withdraw from the study at any time without any consequences. The project maintains strict participant confidentiality by anonymizing data, keeping personally identifiable information separately, and using secure methods to store the data. The identifiable data are only accessible to authorized personnel. Consultations with community leaders and representatives are conducted to ensure that the project and its procedure respect local cultural and ethical standards. Continuous feedback loop with community leaders was maintained to address any concerns or issues that arise during the study. All research activities undergo a review by an ethics committee to ensure adherence to ethical standards. Independent audits will be conducted to verify compliance with ethical guidelines.

There is a governance board consisting of local scientists, and representatives from the Division of Human Genetics at Université Cheikh Anta Diop. The board meets regularly to review project progress, address ethical and governance issues, and make decisions regarding data access and use. There is a clear and transparent processes for decision making, including documented meeting minutes and decisions. Stakeholders, including the National Academy of Sciences and Techniques of Senegal, are informed of major decisions and changes to the project. To ensure accountability, the project set up regular reporting to funding bodies, partners, and the community on the project's progress and outcomes. Those reports allow for stakeholders to raise concerns or provide input on the project's activities. A major emphasis of the project is building local capacity through training and development programs for researchers and staff. This is partly achieved by the involvement of local researchers in all aspects of the project to ensure knowledge transfer and sustainability.

The data will be stored at the national HPC cluster in Diamniadio with robust security measures. The infrastructure uses advanced encryption methods for data storage and transfer. The strict access control mechanisms ensure only authorized personnel can access sensitive data. Multi-factor authentication and regular access audits are implemented to prevent unauthorized access. Standard backup and recovery procedures are in place to prevent loss in case of technical failures. There is also a disaster recovery plan to ensure data integrity and availability in the event of a system failure, as well as regular security assessments and audits to ensure compliance with best practices and international data security standards. The National Data Center, which oversees the HPC, has protocols for responding to data breaches or security incidents that include an immediate notification of affected individuals and stakeholders in case of a data breach.

The project team, in collaboration with the governance board, is responsible for approving requests for data reuse. All requests for data reuse must be submitted with a detailed research proposal outlining the intended use of the data. Then, the proposal is reviewed by the project team to ensure alignment with the study's goals and ethical standards. Proposals are reviewed by a subcommittee within the governance board, which includes members with expertise in genomics and bioethics; the scientific merit, ethical considerations, and potential impact of the proposed research is then evaluated. Additionally, proposals must also receive approval from the National Ethics Committee to ensure they comply with national and international ethical standards. Only authorized researchers with approved research proposals can access the project's data. Access permissions are granted based on the principle of least privilege, ensuring users only have access to the data specified in their approved research proposal. All personal identifiers are removed or anonymized before data sharing and reviewed to ensure no residual identifying information is present. Researchers must

sign a data use agreement (DUA) that outlines the terms and conditions of data reuse. The DUA includes clauses on data security, privacy protection, and the prohibition of data sharing with third parties without explicit permission. A transparent process for tracking data access and sharing is maintained. Researchers are encouraged to share their findings with the original data providers and the wider community, ensuring that the benefits of data reuse are disseminated. Any publications resulting from the data reuse must acknowledge the SEN-GENOME project.

Participants receive comprehensive information about the study, including its purpose, procedures, risks, and benefits, in their native language. The consent forms are reviewed and approved by the independent ethics review board of the University Cheikh Anta Diop. Participants can withdraw from the study at any time without any consequences. With respect to confidentiality and data anonymization, the SEN-GENOME project maintains strict participant confidentiality by anonymizing data, storing personally identifiable information separately, and using secure methods for data storage. Identifiable data are accessible only to authorized personnel. The project consults the communities by engaging with community leaders and representatives to ensure the project respects local cultural and ethical standards. A continuous feedback loop with community leaders addresses any concerns during the study.

Acknowledgments

The authors are grateful to the Senegalese National Academy of Science and Technology.

Declaration of interests

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used chatGPT in order to check

and correct language spelling and grammar. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Supplemental information

Supplemental information can be found online at <https://doi.org/10.1016/j.ajhg.2024.08.017>.

References

1. International Human Genome Sequencing Consortium (Oct 21 2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945. <https://doi.org/10.1038/nature03001>.
2. Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., et al. (2011). Modernizing reference genome assemblies. *PLoS Biol.* 9, e1001091. <https://doi.org/10.1371/journal.pbio.1001091>.
3. Schneider, V.A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H.C., Kitts, P.A., Murphy, T.D., Pruitt, K.D., Thi- baud-Nissen, F., Albracht, D., et al. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* 27, 849–864. <https://doi.org/10.1101/gr.213611.116>.
4. Sherman, R.M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M.P., Chavan, S., Vergara, C., Ortega, V.E., et al. (2019). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* 51, 30–35. <https://doi.org/10.1038/s41588-018-0273-y>.
5. Wang, Q., Pierce-Hoffman, E., Cummings, B.B., Alföldi, J., Francioli, L.C., Gauthier, L.D., Hill, A.J., O'Donnell-Luria, A.H., et al.; Genome Aggregation Database Production Team; and Genome Aggregation Database Consortium (2020). Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* 11, 2539. <https://doi.org/10.1038/s41467-019-12438-5>.
6. Liao, W.W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J.K., Monlong, J., Abel, H.J., et al. (2023). A draft human pangenome reference. *Nature* 617, 312–324. <https://doi.org/10.1038/s41586-023-05896-x>.
7. 1000 Genomes Project Consortium, Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. <https://doi.org/10.1038/nature15393>.
8. Cho, Y.S., Kim, H., Kim, H.M., Jho, S., Jun, J., Lee, Y.J., Chae, K.S., Kim, C.G., Kim, S., Eriksson, A., et al. (2016). An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes. *Nat. Commun.* 7, 13637. <https://doi.org/10.1038/ncomms13637>.
9. Sim, S.C., Kacevska, M., and Ingelman-Sundberg, M. (2013/02/01 2013). Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. *Pharmacogenomics J.* 13, 1–11. <https://doi.org/10.1038/tpj.2012.45>.
10. Furuta, T., Iwaki, T., and Umemura, K. (2010). Influences of different proton pump inhibitors on the anti-platelet function of clopidogrel in relation to CYP2C19 genotypes. *Br. J. Clin. Pharmacol.* 70, 383–392. <https://doi.org/10.1111/j.1365-2125.2010.03717.x>.
11. De, T., Park, C.S., and Perera, M.A. (2019). Cardiovascular Pharmacogenomics: Does It Matter If You're Black or White? *Annu. Rev. Pharmacol. Toxicol.* 59, 577–603. <https://doi.org/10.1146/annurev-pharmtox-010818-021154>.
12. Ndour, E.H.M., Mnika, K., Tall, F.G., Seck, M., Ly, I.D., Nembaware, V., Mazandu, G.K., Sagna Bassène, H.A.T., Di-one, R., Ndongo, A.A., et al. (2022). Biomarkers of sickle cell nephropathy in Senegal. *PLoS One* 17, e0273745. <https://doi.org/10.1371/journal.pone.0273745>.
13. Faye, B.F., Kouame, K.B., Seck, M., Diouf, A.A., Gadj, M., Dieng, N., Touré, S.A., Sall, A., Toure, A.O., and Diop, S. (2018). Challenges in the management of sickle cell disease during pregnancy in Senegal, West Africa. *Hematology* 23, 61–64. <https://doi.org/10.1080/10245332.2017.1367534>.
14. Macia, E., Gueye, L., and Duboz, P. (2016). Hypertension and Obesity in Dakar, Senegal. *PLoS One* 11, e0161544. <https://doi.org/10.1371/journal.pone.0161544>.
15. Traore, P.W.H.B., Tine, J.A.D., Bassoum, O., Kane, A., and Faye, A. (2023). Associated factors with hypertension, known poorly controlled hypertension, and newly diagnosed hypertension among people aged 18-70 in Senegal. *J. Public Health Afr.* 14, 2538. <https://doi.org/10.4081/jphia.2023.2538>.
16. Guissé, P.M., Sall, S.A.B., Niang, T., Doucouré, T.S., Mboup, M.C., Ngaidé, A.A., and Mbaye, A. (2024). [Acute coronary syndromes in diabetes mellitus : A comparative study between diabetics and non-diabetics patients in Senegalese urban environment]. *Ann. Cardiol. Angiol* 73, 101767. <https://doi.org/10.1016/j.ancard.2024.101767>.
17. Diallo, A.I., Dieng, C.M., Tine, J.A.D., Bassoum, O., Diongue, F.B., Ba, M.F., Ndiaye, I., Ndiaye, M., Faye, A., and Seck, I. (2024). Factors associated with diabetes knowledge, attitudes and practices among people aged 18 and over in the commune of Niakhene in Senegal. *PLOS Glob. Public Health* 4, e0002265. <https://doi.org/10.1371/journal.pgph.0002265>.
18. Choudhury, A., Ramsay, M., Hazelhurst, S., Aron, S., Barden, S., Botha, G., Chimusa, E.R., Christoffels, A., Gamielien, J., Sefid-Dashti, M.J., et al. (2017). Whole-genome sequencing for an enhanced understanding of genetic variation among South Africans. *Nat. Commun.* 8, 2062. <https://doi.org/10.1038/s41467-017-00663-9>.
19. Fatumo, S., Yakubu, A., Oyedele, O., Popoola, J., Attipoe, D.A., Eze-Echesi, G., Modibbo, F.Z., Ado-Wanka, N., et al.; 54gene Team; and NCD-GHS Consortium (2022). Promoting the genomic revolution in Africa through the Nigerian 100K Genome Project. *Nat. Genet.* 54, 531–536. <https://doi.org/10.1038/s41588-022-01071-6>.
20. Hamdi, Y., Boujemaa, M., Ben Aissa-Haj, J., Radouani, F., Khyatti, M., Mighri, N., Hannachi, M., Ghedira, K., Souiai, O., Hkimi, C., et al. (2024). A regionally based precision medicine implementation initiative in North Africa: The Per-MediNA consortium. *Transl. Oncol.* 44, 101940. <https://doi.org/10.1016/j.tranon.2024.101940>.