Concept Paper

Not peer-reviewed version

Towards Setting Minimum and Optimal Data to Report for Malaria Molecular Surveillance with Targeted Sequencing: The “What” and the “Why”

Jonathan Juliano [*] , Cecile Meier-Scherling , Neeva Wernsman Young , George Tollefson , Sean Connelly , Jonathan Parr , Melissa Conrad , Jacob Sadler , Christopher Hennelly , Ashenafi Assefa , Lucy Okell , Abebe Fola , Karamoko Niare , Jacob Marglous , Kelly Carey-Ewend , Ronald Futila Kyong-Shin , Isabela Gerdes Gyuricza , Gina Cuomo-Dannenburg , Shazia Ruybal-Pesántez , Oliver Watson , Robert Verity [†] , Jeffrey Bailey [*,†]

Posted Date: 21 July 2025

doi: 10.20944/preprints202503.0561.v2

Keywords: malaria; surveillance; molecular epidemiology; data; genomics; FAIR

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 21 July 2025 doi:10.20944/preprints202503.0561.v2

Concept Paper

Towards Setting Minimum and Optimal Data to Report for Malaria Molecular Surveillance with Targeted Sequencing: The "What" and the "Why"

Running Title: MMS Data

Jonathan J. Juliano [1,2,3,4,*], Cecile P. G. Meier-Scherling [5], Neeva Wernsman Young [5,6], George A. Tollefson [5,6], Sean V. Connelly [7], Jonathan B. Parr [1,2,4], Melissa D. Conrad [8], Jacob M. Sadler [1,2], Christopher M. Hennelly [1], Ashenafi Assefa [1,9], Lucy Okell [10], Abebe A. Fola [6], Karamoko Niaré [6], Jacob Marglous [5,6], Kelly Carey-Ewend [2,7], Ronald Futila Kyong-Shin [11,12], Isabela Gerdes Gyuricza [4], Gina Cuomo-Dannenburg [10], Shazia Ruybal-Pesántez [10,13], Oliver J. Watson [10], Robert Verity [10,†] and Jeffrey A. Bailey [5,6,*,†]

[1] Division of Infectious Diseases, Department of Medicine, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA
[2] Institute for Global Health and Infectious Diseases, University of North Carolina, Chapel Hill, NC 27599, USA
[3] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA
[4] Curriculum in Genetics and Molecular Biology, University of North Carolina, Chapel Hill, NC 27599, USA
[5] Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA
[6] Department of Pathology and Laboratory Medicine, Brown University, Providence, RI 02912, USA
[7] MD-PhD Program, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA
[8] Department of Medicine, University of California- San Francisco, San Francisco, CA 94110, USA
[9] Ethiopian Public Health Institute, Addis Ababa, Ethiopia
[10] MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College, London, W12 0BZ, UK
[11] Curriculum in Bioinformatics and Computational Biology, University of North Carolina, Chapel Hill, NC 27599, USA
[12] National Institute of Biomedical Research, Kinshasa, Democratic Republic of the Congo
[13] Instituto de Microbiología, Universidad San Francisco de Quito, Ecuador
[*] Correspondence: Author: jjuliano@med.unc.edu (J.J.J.); jeffrey_bailey@brown.edu (J.A.B.)
[†] Co-senior authors.

Abstract

The COVID-19 pandemic showcased the power of genomic surveillance in tracking infectious diseases, driving rapid public health responses, and global collaboration. This same infrastructure is being leveraged for malaria molecular surveillance (MMS) in Africa to tackle challenges like artemisinin partial resistance and *Plasmodium falciparum* histidine-rich protein 2 and 3 gene deletions. However, variability in reporting sequencing methods and data reporting is currently limiting the validation, comparability, and reuse of data. To maximize the impact of MMS, we propose minimal and optimal data for reporting that are key for validation and maximize transparency and FAIR (Findable, Accessible, Interoperable, Reusable) principles. Rather than focusing on specific data formats, here, we propose what should be reported and why. Moving to reporting individual infection-level polymorphism or microhaplotype data is central to maximizing the impact of MMS. Reporting must adhere to local regulatory practices and ensure proper data oversight and management, preventing data colonialism and preserving opportunities for data generators. With malaria's challenges transcending borders, reporting and adopting standardized practices are essential to advance research and strengthen global public health efforts.

Keywords: malaria; surveillance; molecular epidemiology; data; genomics; FAIR

Manuscript

Pathogens evolve through mutations that can modify their virulence, transmissibility, and response to interventions. These evolutionary changes drive dynamic processes with consequences at regional, national, and global scales. During the COVID-19 pandemic, molecular surveillance was central to our ability to respond to rapidly evolving strains. Large-scale sequencing of SARS-CoV-2 variants provided the ability to quickly identify new variants, allowing the public health community to swiftly monitor their spread and develop and deploy new interventions, such as updating vaccines.[1,2] Moreover, the ability to study transmission using sequencing enabled a better understanding of the risks posed by infected individuals to uninfected populations.[3–7] The terms "sequencing" and "variant" became commonplace in the public's vocabulary, and an appreciation for the impact of these tools broadened. Community efforts such as NextStrain allowed for interactive, close-to-real-time monitoring of pathogen spread by tracking genetic variation and evolution across broad geographic areas, combined with detailed information on the location where isolates were collected.[8] Modeling groups from around the globe were able to leverage this information to provide robust inference of epidemiological dynamics, inform outbreak response, help with predictions for the eventual impact of the pandemic, and infer undetected infections via seropositivity.[9–12] There were challenges associated with genomic surveillance of COVID-19 during the pandemic, including differences in initial methods and quality control, inconsistencies in clinical and demographic data reporting, delays in sharing, and slow compilation, which limited the utility of comprehensive datasets.[13] Since then, the World Health Organization's (WHO's) Global Strategy for Genomic Surveillance of Pathogens with Pandemic and Epidemic Potential, 2022–2032, recommended that pathogen genomic surveillance be introduced in every country of the world.[14] In addition, newly established networks such as the WHO International Pathogen Surveillance Network (https://www.who.int/initiatives/international-pathogen-surveillance-network) and the Africa Pathogen Genomics Surveillance Network (https://africacdc.org/priority-pathogens-and-use-cases-for-genomic-surveillance-in-africa/) are advocating for the use of pathogen genomics to inform public health decision-making.

Molecular surveillance has now become integral to understanding multiple pathogens of global public health significance, including malaria. Malaria is an endemic disease across the world, particularly in Africa, where the vast majority of the cases and approximately half a million deaths occur annually. Over the last 5 years, a significant investment has been made in malaria molecular surveillance (MMS) by foundations, such as the Gates Foundation; by public health agencies, such as Africa CDC and the US CDC; through research agencies, such as the United States National Institutes of Health (NIH), Japan International Cooperation Agency (JICA), and European and Developing Countries Clinical Trials Partnership (EDCTP); through use of Global Fund support to national malaria control programs; and by direct investment by countries. Molecular and genomic surveillance of malaria will impact both public health decisions and our fundamental understanding of the parasite's biology. Multiple use cases for molecular surveillance have been outlined previously, including:[15]

- Identifying the molecular mechanism/origin of drug and diagnostic resistance
- Monitoring the prevalence/frequency and spread of drug or diagnostic resistance markers
- Classifying outcomes in therapeutic efficacy studies (TESs) as reinfection, recrudescence, or, in the case of *P. vivax,* relapse
- Estimating transmission intensity
- Estimating the connectivity and movement of parasites between geographically distinct populations

- Classifying malaria cases as locally acquired or imported from another population
- Reconstructing granular patterns of transmission

The effective use of MMS is challenged by multiple issues. One central challenge is the timeliness of results being communicated to partners who actually enact policy, such as National Malaria Control Programs and the World Health Organization. There are multiple reasons for this: 1) the complex nature of datasets can lead to long analysis times, 2) data is sometimes held for conferences or high-impact manuscripts, or 3) data may not be generated in-country or may not have involvement of key stakeholders. This highlights the importance of "How" MMS is done and by "Who". There is an ongoing push to facilitate the transition to African institutions, including government institutions, to conduct the work and manage the data. This is an essential process to make MMS as impactful as possible, as early data sharing can have direct impacts. An example of the impact of early data sharing and academic production is highlighted by the recent co-publication of a Lancet Microbe and Lancet Infectious Diseases article about ART-R in Tanzania. In 2024, we reported on the prevalence of ART-R in Tanzania from 2021.[16] While this data did take time, nearly 3 years after collection, to reach the broader academic audience, it had been shared with the Tanzania NMCP and in-country WHO representatives. This led to a therapeutic efficacy study, which was published by Ishengoma et al. at essentially the same time.[17] This robust interaction between surveillance and evaluation occurred because the primary driver of this work was a strong collaboration of officials from the Ministry of Health and MMS experts (local and international) under the leadership of the National Institute for Medical Research (NIMR) in Tanzania. "How is MMS conducted" and "Who is conducting it" have their own challenges, but this perspective focuses on two different questions: the "What" in terms of data presentation and the "Why" of the selection of the data to be shared.

Malaria genomics is challenging given the complexity of the pathogen and the infection. Malaria genomics has blossomed with the emergence of next-generation sequencing. Akin to COVID-19, the increased sequencing capabilities have allowed unprecedented insight into parasite genetics and dramatically improved our ability to monitor emerging threats such as drug and diagnostic resistance. However, compared to viral genomics, *Plasmodium spp.* (and other eukaryotic pathogens) present differing and more numerous challenges. This includes high prevalence of infections with low parasite densities, where samples are overwhelmingly human host DNA. Low parasitemia samples are often challenging to sequence and require enrichment techniques to capture enough parasite genomic material.[18] Mixed infections with minor strains or multiple species that can occur at low relative abundances limit their detection relative to the sequencing error rate.[19] Additionally, eukaryotic recombination continually reassorts parts of the parasite genome, obfuscating relationships between parasites and making tracking transmission chains or the parasite origins difficult. Beyond complex biology, the standardization of methods is challenging. Unlike SARS-CoV-2, it is not practical to sequence the entire genome all the time. While *Plasmodium spp.* genomes are actually on the smaller size relative to other eukaryotes, at only 23 megabases (Mb) for *P. falciparum;* this size still presents challenges vis-à-vis the cost-effectiveness of leveraging only whole-genome sequencing (WGS) for surveillance. The cost of WGS remains prohibitively expensive for large-scale MMS and instead, less-expensive targeted sequencing approaches, such as multiplexed amplicon deep sequencing or molecular inversion probes (MIPs),[20–26] are often employed. Targeted sequencing methods are also more sensitive and better for low parasitemia samples.[27] To date, compared to WGS, bioinformatic tools have been relatively ad hoc and optimized for specific questions of interest.[19,28–30] Thus, a major challenge with targeted MMS is the heterogeneity in assays and analyses, and the need to ensure that data can be properly validated as well as combined and analyzed in a rigorous way.

All next-generation sequencing targeted and whole-genome approaches used for MMS essentially generate the same deep sequencing data. Therefore, the data reported can be highly similar regardless of the approach. And there are universal metrics that impart measures of quality of reads, coverage, and genotyping calls. Data reporting formats and standards for the genetic information can be used across single target amplicon deep sequencing[31], simple amplicon panels[20], highly complex amplicon panels[23], and MIPs, as well as whole genome sequencing.[25] These guidelines can not be fully applied to some older methods used for MMS, like microsatellites and Sanger sequencing.

MMS is poised to help address urgent emerging challenges. Malaria control programs in Africa face two threats where MMS will be particularly helpful. First, the emergence and spread of artemisinin partial resistance (ART-R) is a growing, major public health challenge. Since its first report in 2014 in Rwanda, mutations in the *Plasmodium falciparum* kelch13 (K13) protein associated with ART-R have been detected across multiple African countries. Multiple WHO-candidate and validated ART-R markers have been found along the Rift Valley, reaching from Eritrea to Rwanda.[32–34] These mutations have reached a prevalence of over 20% in many of these areas.[32–34] K13 polymorphisms have also been sporadically found across Africa in other locations and are cropping up now in Southern Africa.[35] However, the malaria community is currently in an advantageous position compared to that during the emergence of resistance to previous antimalarials. Unlike before, we have identified ART-R resistance markers before the partner drugs co-formulated with artemisinin derivatives have started to fail. Thus, we can leverage MMS to: 1) evaluate and monitor the spread of resistance based on the K13 mechanism; 2) study parasite evolution and fitness associated with K13 mutations and mutations associated with artemisinin combination therapies (ACTs); and 3) directly monitor the impact of interventions put in place to reduce the spread of ART-R and preserve the efficacy of ACTs. Further, genomics, particularly targeted deep sequencing, has the ability to provide “molecular correction” in Therapeutic Efficacy Studies (TESs) by distinguishing new infections from recrudescences (treatment failures) to determine if parasites found before and after treatment are the same strain.[30,36] The second challenge is the emergence of “diagnostic-resistant” parasites in the Horn of Africa, where parasites have a deletion of the genes encoding histidine-rich proteins 2 and 3 (HRP2 and HRP3). HRP2 and its paralogue HRP3 are the primary antigens detected by *P. falciparum* malaria rapid diagnostic tests (RDTs) in Africa. Parasites lacking the genes encoding these proteins are not detected by these widely used RDTs. These parasites are spreading and have risen to high prevalence in the region, causing malaria control programs in countries like Eritrea and Ethiopia to push towards alternative diagnostics.[37–40]

How is MMS data currently being collated? Current efforts to collate data on molecular threats for malaria control primarily occur through three mechanisms. The WorldWide Antimalarial Resistance Network (WWARN, www.wwarn.org) was formed in 2009 and has been actively drawing both from the published literature and through direct collaboration with investigators to populate interactive antimalarial resistance mapping tools. While these visualizations and the underlying data have been a valuable resource, the timely integration of data from large genomic epidemiology studies has lagged behind. This is partly due to the labor involved in manual extraction and the lack of consistency or adequacy in the data reported. The WHO also maintains the Malaria Threats Map and interactive dashboards (https://apps.who.int/malaria/maps/threats/) for both antimalarial resistance and *hrp2/3* deletion. Similar challenges exist for these tools and the underlying data, particularly with respect to timely data integration. Lastly, individual projects or national malaria control programs have developed dashboards to collate and visualize data. Given the potential benefits of combined data analysis, making the collation of data easier is critical to maximize and streamline its use.

Here, we focus on what data should be reported both minimally and optimally for MMS. Despite the explosion in high-throughput sequencing, the malaria community has not addressed what key data needs to be reported to properly leverage this information for control and elimination programmatic priorities, as well as further scientific investigation. This has led to haphazard and ambiguous publication of data or limited release of the underlying data that lessens its impact. Here, we focus on the “what” and “why” of data reporting rather than the “how” - avoiding the second (and often unintentionally intertwined) step regarding data format and storage that often unfortunately complicates these discussions. The goal is to recommend data reporting standards, recognizing that individual data generators must abide by local regulatory and data management requirements. In addition, different data generators may have different resources, either financial or personnel, available that can work toward depositing data. Currently, no data generator is meeting these standards on a routine basis, including the authors. These perspectives and proposed standards are the vision of the authors, as researchers who have worked extensively on MMS and with multiple partners, and are not meant to represent NMCPs or the WHO. Beyond basic reporting, the authors felt the standards proposed would be beneficial for the scientific community to ensure rigorous findings and allow broader use of the data. This is meant to set an “initial goal post” towards which the community can more broadly refine and move towards together. **Table 1** highlights requirements for depositing MMS data into repositories to ensure rigor, reproducibility, and reuse. **Table 2** highlights proposed reporting for sequencing data and metadata to allow for more rapid reuse towards MMS goals, at the national and regional level, as well as for academic purposes to help advance our understanding of malaria. It is important to note that there are many common metrics used in MMS, such as multiplicity of infection (MOI) or complexity of infection (COI), that are commonly used as part of data analysis. It is obviously good to report these metrics, but they were not included in the minimal standards for reporting. This is in part because comparison across studies using different methods to assess these metrics may not be the best option, but rather having the data available in a way to allow reuse to jointly calculate these metrics across studies would be more robust.

Table 1. Public repository deposition for rigor, reproducibility and reuse.

Variable	Minimum Standard	Optimal Standard
	Study and Participant MetaData	
Raw Sequence	All studies should provide underlying raw sequencing data for reproducibility of findings by others.	Same as minimum.
	Raw sequencing data are the key to true reproducibility and validity of any study and should be required. Without raw data, inappropriate analyses leading to called variants or microhaplotypes can never be properly addressed. This also optimizes data for use for other scientific questions.	
Metadata	All key variables as deemed de-identified used in study for the published work deposited in a sustainable uncontrolled public database (e.g., open access).	All key variables deposited in a public controlled database that allow full reanalysis and validation of the study deposited in a sustainable uncontrolled public database (e.g., access needs approval as may contain identifiable data)
	Full metadata can potentially lead to participant identification -- although the risk of negative impact to study participants is low given malaria is a common, unstigmatized disease. Optimally, all data exactly as used in published analyses is deposited into a controlled database that allows for registered, vetted scientists to reproduce, validate, and extend work.	
Methods/Code	Detailed methods used for processing sequence data, programs, settings, filtering, and	Fully reproducible coding pipeline that takes data and produces all results and figures from the main analysis.

	analysis with metadata.	
	While detailed written methods are key, for analysis the exact code used to analyze data and generate figures allows others to examine and check methods. Deposition of code in GitHub or similar platform is obligatory. New developing methods for code reproducibility, such as Code Ocean compute capsules, are being implemented.[60]	
SequencingPanel/Assay	Genomic locations sequenced and genotyped.	Complete description of panel target regions and any filtered regions that may have been ignored due to high-levels of known sequencing error.
	Understanding the gene or genomic locations assayed by a panel allows for better integration of data. Panel design should be deposited in an easily accessible public database that is fixed in version at the time of the study. Combined with microhaplotype or allele depth, this allows for retrospective determination of reference genotypes for new mutations found later--since the older study would have only found wild type and thus not have reported a nonvariant site. Filtered regions removed due to difficult-to-assess repeats or error-prone sequences are important since underlying variation found in subsequent studies in these regions would need reanalysis. Microhaplotypes and their within-sample counts represent a compact format that is lossless and easily encodes how well a missing mutation in earlier samples sets was missed.	
Controls	Set of parasite standards to provide context of sensitivity and specificity; All studies should be run with negative controls (e.g., human DNA or water).	In addition to controls, random replicate samples to assess assay variation in 5-10% of samples.
	Laboratory-derived controls ensure consistent assay performance but cannot address the sample quality for a given experiment. Thus, repeating a percentage of samples (biological replicates) and assays (technical replicates) provides a more robust assessment of a given sample set. Ultimately, replicates (duplicates or even triplicates) can help control for noise and jackpot events, although these efforts increase costs.	

Table 2. Specific metadata and variant data and measures/statistics for Malaria Molecular Surveillance.

Variable	Minimum Standard	Optimal Standard
	Study and Participant MetaData	
Date of collection	Month and year of collection; Start and end date of study (maximal aggregation over a year).	Individual collection date (jittered if malaria diagnosis date is considered identifying information to maintain longitudinal order at site).
Location of collection	Collection site or aggregated neighboring collection sites with GPS coordinate of clinic used or centroid of neighborhood. *Clinic, village or town should be easily*	Highest resolution data possible (GPS location of household, clinic of collection, town/city of collection) at individual level data (jittered if considered potentially

	attainable.	identifying information).
Age at time of collection	Age in years at time of collection.	Age in years and months or years to a single decimal place (at study start if longitudinal).
Sex	As collected by the study.	As collected by the study.
Treatment status	Pre-treatment or post-treatment. *Important to understand if frequencies or prevalences of drug resistance mutations could be skewed due to recent drug pressure in the individual.*	Complicated studies with multiple time points should delineate timing of sampling -- e.g., TES.
Sampling strategy	Symptomatic, asymptomatic, community, clinic, etc. on a study level.	Assigned to each individual sample in cases of complex study design.
Travel information	If available.	Provide all travel information available at individual level data.
Sequencing and Genotyping Data		
Variant/ haplotype calls	Nucleotide or amino acid change at variant sites called. Heterozygous or homozygous calls of known public health import. Individual level data should be reported for specific mutations, including validated resistance variants without observed variant genotypes. FAIR format variant calls such as VCF or preferably gVCF provided supplementally. *With next-generation sequencing (NGS) data, reporting within-sample allele frequencies is important.* This data should be unfiltered (no sites or samples removed beyond baseline initial variant e calling.	Individual-level full microhaplotypes if generated and genotyping data (amino acid and nucleotide, indels, etc.) across all regions sequenced/variants called and provided in FAIR formats. Development of microhaplotypes that maintain linkage information and are optimal. *Microhaplotypes and, to a slightly lesser extent, full GVCFs allow for examination of potential new mutations that might be captured but otherwise not recognized initially*.This data should be unfiltered (no sites or samples removed beyond baseline initial variant e calling.
Read or UMI depth	Number of reads or unique molecular identifiers (UMI) informing each genotype (SNP or combination of SNPs) reported at each locus by individual. Total number of reads per locus reported. *This is key to any quality assessment to know how much weight each sample gets.*	Number of reads or unique molecular identifiers (UMI) informing each full haplotype called (not just those reported in the manuscript) at each loci by individual. Total number of reads per loci reported. *Read depth provides a limited approximation of the information content, whereas UMIs provide a fuller accounting traceable to individual molecules of template in the sample.*
Frequency (population and within sample of allele or variant)	Average allele frequency for aggregate site/region.	Within-sample allele frequencies for each participant; these can be calculated directly or from read depth/UMI counts.
	Allele frequency is not always reported compared to prevalence. However, frequency is much more robust to assessing sequencing error or low-level contamination. For instance, presume in 100 samples there are 10 samples with errors reporting K13 C580Y at a within sample allele	

	frequency of 1% each. For those 100 samples it would result in a reported prevalence of 10%, but only an average population allele frequency for 580Y of 0.1%. There is concern that such errors occur when there is a high percentage of mixed infections for a given mutation.

The ultimate goal should be the deposition of de-identified individual participant sequence data and metadata into the public realm. FAIR (Findable, Accessible, Interoperable, Reusable) principles should be followed to the greatest extent to allow the most extensive use of data for other scientific questions.[41–43] The lack of such data inhibits the forward advance of science and therefore prompted the US National Institutes of Health (NIH) to require this reporting standard for grants submitted after 2023. However, many publications still leave data available upon request, which often results in delays in access and additional unstated requirements such as authorship stipulations by the data holders. MMS supported by other funders or carried out by government agencies may not be mandated to deposit raw data in a similar way.

To maximize reproducibility and reuse, both raw sequencing data and initial variant calls should be shared with the community. Sharing raw sequencing data at the individual participant level would allow the fullest reuse and should be the gold standard for data sharing. However, full computational reanalysis would be laborious and costly, limiting its reuse by many groups. Therefore, the reporting of initial variant data with publication would allow the community to take full advantage of the previous analysis of sequencing data to integrate across data sets. While some initial filtering for quality, depth, or missingness may be part and parcel of variant or microhaplotype calling, providing the baseline calls prior to more stringent filtering that may be needed for subanalyses. This initial call data allows others to reproduce results without downloading and recalling variation, and allows for appropriate filtering for any additional analyses. Data on read depth per amplicon and raw numbers of reads supporting each allele variant are needed. This importantly maintains the within-sample variant frequencies in cases of mixed infections. In addition, if available, data on both individual simple variants [e.g., single-nucleotide polymorphisms (SNP) and small indels] and overall microhaplotypes (a segment of DNA containing 2 or more mutations) should be reported. Provision of the exact microhaplotypes (the full sequence or all variation from a single amplicon) is the optimal format for polymerase chain reaction (PCR)-based targeted sequencing as it preserves the observed linkage between polymorphisms which can contain important information–including level of resistance, origin of a mutation, spatial or relatedness information that might otherwise be lost. For example, *P. falciparum* dihydrofolate reductase (*dhfr*), *P. falciparum* dihydropteroate synthase (*dhps*), *P. falciparum* chloroquine resistance transporter (*crt*), and *P. falciparum* multidrug resistance protein 1 (*mdr1*) [reviewed in [44–48]]. Similarly, measures of copy number, whether from sequence data or other methods (e.g., qPCR) should be reported as a continuous copy number for the sample (not a rounded value), allowing the overall variance to be further assessed. It is essential that variant reporting also meets FAIR requirements.[43] Reporting of underlying read depth for microhaplotypes in each sample allows others to potentially assess copy number if not previously done. For instance, the publication of data only as plots of site frequency, common in the literature, does not meet these standards. To facilitate this, data reporting standards are needed, particularly for this level of data, meaning it is key to define what data should be minimally and optimally reportable (**Table 2**). There are well-developed public locations for the submission of raw sequencing data, such as the National Center for Biotechnology Information's (NCBI's) Short Read Archive (SRA) and the European Nucleotide Archive (ENA). At a minimum, initial variant data should be presented as machine-readable tables deposited as supplemental, publicly available data. The selection of storage for variant data are more variable and newer, flexible centralized repositories for disparate data, such as Zenodo (zenodo.org), housed at CERN, can provide storage for variant calls that are currently evolving. Individual project-based variant viewers have also been leveraged.[49]

Certain aspects of metadata are critical to future use. Data sharing must comply with local ethical board requirements, with all individual-level data shared in de-identified formats. Within these constraints, the most precise information allowable that complies with these regulations should

doi:10.20944/preprints202503.0561.v2

be made available. First, the geographic location of sample collections should be as precise as possible to allow for accurate mapping, intervention deployment, and understanding of impacts.[25,50] This may include jittered household-level data, but may be limited to the health district of collection. Second, approximate dates of collection should be shared, as seasonality is important for malaria transmission and temporal trends. It is expected that the prevalence of mutations may vary depending on when in the transmission season they are collected.[50–52] Precise dates should not be shared, but month or season and year may be allowable. Third, demographics play an important role in malaria risk; thus age and sex should be considered essential components.[53,54] Fourth, human mobility is a critical aspect for studying malaria importation and migration of parasites; hence travel data should be included where available.[55] Fifth, treatment history (e.g., pre-treatment sample versus post-treatment sample) is critical for understanding the data, in particular for therapeutic efficacy studies (TES) where recurrent parasitemia is occurring and where post-treatment samples are not representative of the background frequency of mutations due to the selective pressure within individuals making resistant parasitemia more likely in these samples. Lastly, the clinical status of the individuals (e.g., is this a study of clinical cases or asymptomatic cases) should be reported. Similar to variant calls, public repositories like Zenodo provide flexible options for metadata storage. Other options exist, such as university-run data repositories, an example being the UNC Dataverse (https://dataverse.unc.edu/), or within publications themselves.

Multiple technical aspects need to be reported to assess the quality of variant data and samples. First, absolute values of sequencing reads (or the depth) are needed to estimate within-sample allele frequencies for mixed infections. Owing to malaria infections by multiple parasite strains, the observed prevalence of a resistance marker within a population depends on both its frequency and the average number of mixed infections in the population. Mixed infections are often not reported or estimated. Absolute values of sequencing reads therefore allows for more suitable comparisons between studies conducted in different transmission intensities. These underlying values can also importantly allow for quality reassessment, particularly in large studies where false positives become more likely due to the high number of tests. For example, if a variant allele is occurring only at low frequency within samples across a large number of samples, this raises concern for false-positive calls due to sequencing error or contamination. The read depth also allows for better filtering for secondary analyses that may be prone to different levels of allowable error. Lastly, this detail allows for the assessment of sample duplication in studies.

Robust controls should be reported for all MMS, preferably standardized panels that can be used across labs, such as those being developed by WHO and others.[56,57] In addition, quality control should be conducted by repeating 5-10% of samples and reporting the genotype concordance. The use of replicate samples can improve data quality and interpretability in two ways: 1) helping to assess sample quality and 2) helping to improve within-sample allele frequency estimates. In the first case, replicates are necessary as the quality of samples can vary significantly depending on the source. While some studies may have detailed chains of custody with well-documented storage conditions, many others have less reliable information. For example, large national surveys like demographic health surveys (DHSs) have dried blood spots that pass through multiple hands, have unclear storage conditions in the field, and have been used for other assays before being available for MMS.[25,26] Many clinic-based MMS systems or TESs leverage health site care workers to collect samples and store them with minimal ongoing supervision. In the second case, the uncertainty of within-sample allele frequency estimates due to jackpotting (overamplification of certain alleles due to chance or biases), provides a strong argument to be made for replicating all samples when accurate measurement of allele variant absence, presence, or frequency is essential to the question (e.g., TESs), as replicates of 5-10% indicate what percent of samples may go awry but not which samples.[58]

Specific technology and kits used for sequencing, as well as details of filtering and data processing, are needed. When research methods are described only in piecemeal fashion—listing programs and key parameters—important details are often omitted, making it difficult to reproduce or validate findings. Sharing the exact code, such as through GitHub or another repository, is a step forward, but challenges can still arise if others are not using the same platform or environment. Installing and running the software for a pipeline is often a challenge. Issues around bioinformatics code and genomics data analysis tool usability and availability have been recently reviewed using defined software standards criteria[43] and there has been work to improve the usability of software for MMS.[59] Beyond this, leveraging containers that can encapsulate and make portable the entire analysis pipeline for a dataset or manuscript such as Docker and Singularity are even more optimal as well as newer shared cloud environments such as Code Ocean web-browser-based "compute capsules" adopted by Nature Journals, all of which allow code to be run in an encapsulated environment with proper version control from a web browser (https://codeocean.com/resources/nature-partnership).[60]

Finally, data are often used across multiple studies. In this case, denotation in the data table of where else the data have already been presented and made publicly available is essential. There are instances where the same data used in multiple publications enters into meta-analyses multiple times, leading to bias.[61]. Subsequent publications should be required to include standalone supplementary data for the new sequencing data generated by this study as well as PMID or other reference to the previous reporting of the data and delineate the extent of sample or data reuse. This can also help merge data, e.g., variant data for different genes published separately, and is also good practice for peer review of the individual report as it should be clear when data or samples from previous work are being reused. Ultimately, reporting of individual-level data would guard against duplication, given the ability to have unique identifiers per infected sample that transcend individual publications.

Of note, MMS does leverage non-targeted approaches as well. Here, we have focused on reporting targeted sequencing data because that is the most widely generated data in MMS. However, certain aspects of MMS require more detailed data that is best generated with whole-genome sequencing. In particular, studying flanking regions of important genes to understand their origin and spread from a specific origin is potentially useful information. Deciding the best approaches and analysis methods for this is complicated and beyond the scope of this piece. The size of the haplotype block may vary based on the type and age of the selective sweep involved. Currently, WGS represents a fraction of data generated for MMS. However, in general, similar principles of individual-level metadata and genetic data in public repositories would still apply.

While the goal of sharing individual-level data are widely supported, several challenges must be addressed. First, researchers often hesitate to share data due to concerns about being "scooped" on potential analyses. To mitigate this, mechanisms should be established to ensure that data generators have priority in publishing planned analyses.[62] Second, not all studies obtain consent from participants for public sharing of individual-level data, raising ethical and privacy concerns, including the risk of identifiability. To protect participants, shared data should adhere to appropriate consent protocols and be presented in a way that minimizes re-identification risks, especially with accurate geolocation and demographic information of samples. Privacy protections can be further strengthened through the use of secure data repositories with tiered access, where only approved researchers can access specific datasets under predefined conditions.[63] Third, data colonialism remains a significant concern. Achieving equitable data sharing requires a fundamental shift in research culture, alongside financial and policy support from funders, research institutions, journals, and governments.[62,64] This involves integrating equity into data-sharing policies, recognizing all intellectual contributions to research, and aligning academic recognition with data-sharing mandates to ensure appropriate rewards for meta-analyses, data sharing, and capacity-building efforts. Investments in human resources, infrastructure, and collaborative networks are also necessary to strengthen data curation and secondary data analysis capacity in low- and middle-income countries, and to develop sustainable and inclusive platforms for complex data integration and analysis. Finally, concerns about commercialization and benefit sharing must be addressed. Publicly available data may be leveraged by companies to develop profitable products and services without direct benefits to the communities that provided the data. Ensuring that individuals and communities share in the benefits of research outcomes is essential, and mechanisms such as licensing agreements for data use can help address these issues.

What about the question "how to represent genomic data?" Here we have focused on "what" data should be reported and "why" rather than "how". There are multiple potential data formats available for reporting data, in particular for genomic data. For whole-genome data or SNP data, standard formats such as variant call format (VCFs) are an excellent choice. Targeted sequencing also often employs these same formats but data are lost for non-variant regions and haplotypic linkage in mixed infections. Better formats for reuse include GVCFs (or more compressed ReblockGVCFs) that keep information for non-variant regions. Ultimately, preserving microhaplotypes, the most direct representation of error-corrected sequenced PCR products, leads to minimal loss of information compared to the raw sequence data. Drawing from malaria research, one format, Portable Microhaplotype format (PMO) (https://plasmogenepi.github.io/PMO_Docs/), is an attractive lossless (full microhaplotypes) intermediate under active development to promote harmonization of analysis pipelines with negligible raw data loss. It also aims to capture, in a regimented way, appropriate metadata relevant to malaria epidemiology and MMS. Longer-term, standardization by the community of robust, well-defined epidemiological vocabulary will be useful. As the PMO epidemiological metadata is potentially a standalone format, it could be adopted for WGS and other next-generation sequencing pipelines and not solely limited to projects generating microhaplotypes. Other data standards may be needed at different points for downstream analyses; for example, the STAVE (https://github.com/mrc-ide/STAVE) package aims to provide a flexible and convenient format for site and/or temporal aggregate genotype data that can be used in prevalence mapping. These formats and related tools importantly can speed analysis within laboratories and could particularly aid programs working with multiple labs to integrate their own data rather than trying to merge unstandardized data across studies. These formats have the potential to be used beyond malaria research, and for targeted sequencing across multiple organisms, forming the backbone of a unified data analysis ecosystem for targeted sequencing.

The malaria community must mount a coordinated response to the emerging threats of antimalarial drug and diagnostic resistance in Africa. The power of data sharing can be harnessed to provide critical insights into parasite biology and drivers of the spread of resistance that extend well beyond what individual studies can achieve alone. Given that parasites do not respect political borders, the malaria community must also work across borders (and studies). Striving for the highest quality malaria MMS data and reporting is a critical step toward overcoming challenges facing the Africa region and improving the public's health.

Author contributions: Conception and design: JJJ, RV, and JAB; Drafting manuscript: All authors; Revision and final approval: All authors; Accountability: JJJ and JAB. Generative AI was used in the writing of this manuscript. The authors take full responsibility for the content.

Funding: This prospective was funded by the National Institutes for Allergy and Infectious Diseases (R01AI156267 to JAB and JJJ, R01AI139520 to JAB, R01AI155730 to JJJ and JTL, R01AI173558 to MC, R01AI177791 to JBP, K24AI134990 to JJJ). LO, GC-D, SR-P, OJW and RV acknowledge funding from the MRC Centre for Global Infectious Disease Analysis (reference MR/X020258/1), funded by the UK Medical Research Council (MRC). OJW is also supported by an Imperial College Research Fellowship sponsored by Schmidt Sciences. This UK funded award is carried out in the framework of the Global Health EDCTP3 Joint Undertaking. This work was partially supported by the Bill & Melinda Gates Foundation (INV-050353 to JBP). The views expressed do not reflect those of the funders.

Ethics Statement: Not applicable.

Data Availability: Not applicable.

Acknowledgements: We thank UNC's Winston House for hosting us during the Infectious Disease Epidemiology and Ecology Laboratory (IDEEL) 2025 London Hackathon, where this manuscript was conceived.

Conflicts of Interest: JBP reports research support from Gilead Sciences, non-financial support from Abbott Laboratories, and consulting for Zymeron Corporation, all outside the scope of the manuscript. RV acknowledges consulting for I C Consultants Limited, outside the scope of this manuscript. All other authors declare no competing interests. This manuscript has been submitted as a preprint (doi pending).

References

1. Kames J, Holcomb DD, Kimchi O, DiCuccio M, Hamasaki-Katagiri N, Wang T, Komar AA, Alexaki A, Kimchi-Sarfaty C., 2020. Sequence analysis of SARS-CoV-2 genome reveals features important for vaccine design. *Scientific Reports 10*: 1–11
2. Abera A, Belay H, Zewude A, Gidey B, Nega D, Dufera B, Abebe A, Endriyas T, Getachew B, Birhanu H, Difabachew H, Mekonnen B, Legesse H, Bekele F, Mekete K, et al., 2020. Establishment of COVID-19 testing laboratory in resource-limited settings: challenges and prospects reported from Ethiopia. *Glob Health Action 13*: 1841963
3. Wang L, Didelot X, Yang J, Wong G, Shi Y, Liu W, Gao GF, Bi Y., 2020. Inference of person-to-person transmission of COVID-19 reveals hidden super-spreading events during the early outbreak phase. *Nature Communications 11*: 1–6
4. Zhang W, Govindavari JP, Davis BD, Chen SS, Kim JT, Song J, Lopategui J, Plummer JT, Vail E., 2020. Analysis of Genomic Characteristics and Transmission Routes of Patients With Confirmed SARS-CoV-2 in Southern California During the Early Stage of the US COVID-19 Pandemic. *JAMA Network Open 3*: e2024191
5. Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H, Yang J, Xing F, Liu J, Yip CC-Y, Poon RW-S, Tsoi H-W, Lo SK-F, Chan K-H, Poon VK-M, Chan W-M, et al., 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet (London, England) 395*: 514
6. Bedford T, Greninger AL, Roychoudhury P, Starita LM, Famulare M, Huang ML, Nalla A, Pepper G, Reinhardt A, Xie H, Shrestha L, Nguyen TN, Adler A, Brandstetter E, Cho S, et al., 2020. Cryptic transmission of SARS-CoV-2 in Washington state. *Science (New York, NY) 370*
7. Cerami C, Popkin-Hall ZR, Rapp T, Tompkins K, Zhang H, Muller MS, Basham C, Whittelsey M, Chhetri SB, Smith J, Litel C, Lin KD, Churiwal M, Khan S, Rubinstein R, et al., 2022. Household Transmission of Severe Acute Respiratory Syndrome Coronavirus 2 in the United States: Living Density, Viral Load, and Disproportionate Impact on Communities of Color. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 74*
8. Anon. Website. Available at: https://academic.oup.com/bioinformatics/article/34/23/4121/5001388. Accessed
9. Knock ES, Whittles LK, Lees JA, Perez-Guzman PN, Verity R, FitzJohn RG, Gaythorpe KAM, Imai N, Hinsley W, Okell LC, Rosello A, Kantas N, Walters CE, Bhatia S, Watson OJ, et al., 2021. Key epidemiological drivers and impact of interventions in the 2020 SARS-CoV-2 epidemic in England. *Sci Transl Med 13*
10. Perkins TA, España G., 2020. Optimal Control of the COVID-19 Pandemic with Non-pharmaceutical Interventions. *Bulletin of Mathematical Biology 82*: 1–24
11. Walker PGT, Whittaker C, Watson OJ, Baguelin M, Winskill P, Hamlet A, Djafaara BA, Cucunubá Z, Olivera Mesa D, Green W, Thompson H, Nayagam S, Ainslie KEC, Bhatia S, Bhatt S, et al., 2020. The impact of COVID-19 and strategies for mitigation and suppression in low- and middle-income countries. *Science 369*: 413–422
12. Chiu WA, Ndeffo-Mbah ML., 2021. Using test positivity and reported case rates to estimate state-level COVID-19 prevalence and seroprevalence in the United States. *PLOS Computational Biology 17*: e1009374
13. Ling-Hu T, Rios-Guzman E, Lorenzo-Redondo R, Ozer EA, Hultquist JF., 2022. Challenges and Opportunities for Global Genomic Surveillance Strategies in the COVID-19 Era. *Viruses 14*

14. Anon. WHO global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022-2032. Available at: https://www.who.int/initiatives/genomic-surveillance-strategy. Accessed
15. Dalmat R, Naughton B, Kwan-Gett TS, Slyker J, Stuckey EM., 2019. Use cases for genetic epidemiology in malaria elimination. *Malaria Journal 18*: 1–11
16. Juliano JJ, Giesbrecht DJ, Simkin A, Fola AA, Lyimo BM, Pereus D, Bakari C, Madebe RA, Seth MD, Mandara CI, Popkin-Hall ZR, Moshi R, Mbwambo RB, Niaré K, MacInnis B, et al., 2024. Prevalence of mutations associated with artemisinin partial resistance and sulfadoxine-pyrimethamine resistance in 13 regions in Tanzania in 2021: a cross-sectional survey. *Lancet Microbe 5*: 100920
17. Ishengoma DS, Mandara CI, Bakari C, Fola AA, Madebe RA, Seth MD, Francis F, Buguzi CC, Moshi R, Garimo I, Lazaro S, Lusasi A, Aaron S, Chacky F, Mohamed A, et al., 2024. Evidence of artemisinin partial resistance in northwestern Tanzania: clinical and molecular markers of resistance. *Lancet Infect Dis 24*: 1225–1233
18. Oyola SO, Ariani CV, Hamilton WL, Kekre M, Amenga-Etego LN, Ghansah A, Rutledge GG, Redmond S, Manske M, Jyothi D, Jacob CG, Otto TD, Rockett K, Newbold CI, Berriman M, et al., 2016. Whole genome sequencing of Plasmodium falciparum from dried blood spots using selective whole genome amplification. *Malar J 15*: 597
19. Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA., 2018. SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. *Nucleic Acids Res 46*: e21
20. Sadler JM, Simkin A, Tchuenkam VPK, Gyuricza IG, Fola AA, Wamae K, Assefa A, Niaré K, Thwai K, White SJ, Moss WJ, Dinglasan RR, Nsango S, Tume CB, Parr JB, et al., 2024. Application of a new highly multiplexed amplicon sequencing tool to evaluate antimalarial resistance and relatedness in individual and pooled samples from Dschang, Cameroon
21. Holzschuh A, Lerch A, Gerlovina I, Fakih BS, Al-mafazy A-WH, Reaves EJ, Ali A, Abbas F, Ali MH, Ali MA, Hetzel MW, Yukich J, Koepfli C., 2023. Multiplexed ddPCR-amplicon sequencing reveals isolated Plasmodium falciparum populations amenable to local elimination in Zanzibar, Tanzania. *Nature Communications 14*: 1–16
22. LaVerriere E, Schwabl P, Carrasquilla M, Taylor AR, Johnson ZM, Shieh M, Panchal R, Straub TJ, Kuzma R, Watson S, Buckee CO, Andrade CM, Portugal S, Crompton PD, Traore B, et al., 2022. Design and implementation of multiplexed amplicon sequencing panels to serve genomic epidemiology of infectious disease: A malaria case study. *Mol Ecol Resour 22*: 2285–2303
23. Aranda-Díaz A, Vickers EN, Murie K, Palmer B, Hathaway N, Gerlovina I, Boene S, Garcia-Ulloa M, Cisteró P, Katairo T, Semakuba FD, Nsengimaana B, Gwarinda H, García-Fernández C, Da Silva C, et al., 2024. Sensitive and modular amplicon sequencing of diversity and resistance for research and public health
24. Tessema SK, Hathaway NJ, Teyssier NB, Murphy M, Chen A, Aydemir O, Duarte EM, Simone W, Colborn J, Saute F, Crawford E, Aide P, Bailey JA, Greenhouse B., 2022. Sensitive, Highly Multiplexed Sequencing of Microhaplotypes From the Plasmodium falciparum Heterozygome. *J Infect Dis 225*: 1227–1237
25. Verity R, Aydemir O, Brazeau NF, Watson OJ, Hathaway NJ, Mwandagalirwa MK, Marsh PW, Thwai K, Fulton T, Denton M, Morgan AP, Parr JB, Tumwebaze PK, Conrad M, Rosenthal PJ, et al., 2020. The impact of antimalarial resistance on the genetic structure of Plasmodium falciparum in the DRC. *Nat Commun 11*: 2107
26. Aydemir O, Janko M, Hathaway NJ, Verity R, Mwandagalirwa MK, Tshefu AK, Tessema SK, Marsh PW, Tran A, Reimonn T, Ghani AC, Ghansah A, Juliano JJ, Greenhouse BR, Emch M, et al., 2018. Drug-Resistance and Population Structure of Plasmodium falciparum Across the Democratic Republic of Congo Using High-Throughput Molecular Inversion Probes. *J Infect Dis 218*: 946–955
27. Ruybal-Pesántez S, McCann K, Vibin J, Siegel S, Auburn S, Barry AE., 2024. Molecular markers for malaria genetic epidemiology: progress and pitfalls. *Trends Parasitol 40*: 147–163
28. Early AM, Daniels RF, Farrell TM, Grimsby J, Volkman SK, Wirth DF, MacInnis BL, Neafsey DE., 2019. Detection of low-density Plasmodium falciparum infections using amplicon deep sequencing. *Malar J 18*: 219
29. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods 13*: 581–583

30. Lerch A, Koepfli C, Hofmann NE, Messerli C, Wilcox S, Kattenberg JH, Betuela I, O'Connor L, Mueller I, Felger I., 2017. Development of amplicon deep sequencing markers and data analysis pipeline for genotyping multi-clonal malaria infections. *BMC Genomics 18*: 864
31. Bailey JA, Mvalo T, Aragam N, Weiser M, Congdon S, Kamwendo D, Martinson F, Hoffman I, Meshnick SR, Juliano JJ., 2012. Use of massively parallel pyrosequencing to evaluate the diversity of and selection on Plasmodium falciparum csp T-cell epitopes in Lilongwe, Malawi. *J Infect Dis 206*: 580–587
32. Rosenthal PJ, Asua V, Bailey JA, Conrad MD, Ishengoma DS, Kamya MR, Rasmussen C, Tadesse FG, Uwimana A, Fidock DA., 2024. The emergence of artemisinin partial resistance in Africa: how do we respond? *Lancet Infect Dis 24*: e591–e600
33. Rosenthal PJ, Asua V, Conrad., 2024. Emergence, transmission dynamics and mechanisms of artemisinin partial resistance in malaria parasites in Africa. *Nature reviews Microbiology 22*
34. Ishengoma DS, Gosling R, Martinez-Vega R, Beshir KB, Bailey JA, Chimumbwa J, Sutherland C, Conrad, Tadesse FG, Juliano JJ, Kamya MR, Mbacham WF, Ménard D, Rosenthal PJ, Raman J, et al., 2024. Urgent action is needed to confront artemisinin partial resistance in African malaria parasites. *Nature medicine 30*
35. Martin AC, Sadler JM, Simkin A, Musonda M, Katowa B, Matoba J, Schue J, Simulundu E, Bailey JA, Moss WJ, Juliano JJ, Fola AA., 2025. Emergence and Rising Prevalence of Artemisinin Partial Resistance Marker Kelch13 P441L in a Low Malaria Transmission Setting in Southern Zambia
36. Holzschuh A, Lerch A, Nsanzabana C., 2024. Multiplexed nanopore amplicon sequencing to distinguish recrudescence from new infection in antimalarial drug trials
37. Fola AA, Feleke SM, Mohammed H, Brhane BG, Hennelly CM, Assefa A, Crudal RM, Reichert E, Juliano JJ, Cunningham J, Mamo H, Solomon H, Tasew G, Petros B, Parr JB, et al., 2023. Plasmodium falciparum resistant to artemisinin and diagnostics have emerged in Ethiopia. *Nature microbiology 8*
38. Berhane A, Anderson K, Mihreteab S, Gresty K, Rogier E, Mohamed S, Hagos F, Embaye G, Chinorumba A, Zehaie A, Dowd S, Waters NC, Gatton ML, Udhayakumar V, Cheng Q, et al., 2018. Major Threat to Malaria Control Programs by Plasmodium falciparum Lacking Histidine-Rich Protein 2, Eritrea. *Emerg Infect Dis 24*: 462–470
39. Feleke SM, Reichert EN, Mohammed H, Brhane BG, Mekete K, Mamo H, Petros B, Solomon H, Abate E, Hennelly C, Denton M, Keeler C, Hathaway NJ, Juliano JJ, Bailey JA, et al., 2021. Plasmodium falciparum is evolving to escape malaria rapid diagnostic tests in Ethiopia. *Nat Microbiol 6*: 1289–1299
40. Thomson R, Parr JB, Cheng Q, Chenet S, Perkins M, Cunningham J., 2020. Prevalence of Plasmodium falciparum lacking histidine-rich proteins 2 and 3: a systematic review. *Bull World Health Organ 98*: 558–568F
41. Mathur MB, Fox MP., 2023. Toward Open and Reproducible Epidemiology. *Am J Epidemiol 192*: 658–664
42. Peng RD, Dominici F, Zeger SL., 2006. Reproducible epidemiologic research. *Am J Epidemiol 163*: 783–789
43. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 3*: 1–9
44. Blasco B, Leroy D, Fidock DA., 2017. Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic. *Nat Med 23*: 917–928
45. Fidock DA, Eastman RT, Ward SA, Meshnick SR., 2008. Recent highlights in antimalarial drug resistance and chemotherapy research. *Trends Parasitol 24*: 537–544
46. Ippolito MM, Moser KA, Kabuya J-BB, Cunningham C, Juliano JJ., 2021. Antimalarial Drug Resistance and Implications for the WHO Global Technical Strategy. *Curr Epidemiol Rep 8*: 46–62
47. Conrad MD, Rosenthal PJ., 2019. Antimalarial drug resistance in Africa: the calm before the storm? *Lancet Infect Dis 19*: e338–e351
48. Picot S, Olliaro P, de Monbrison F, Bienvenu A-L, Price RN, Ringwald P., 2009. A systematic review and meta-analysis of evidence for correlation between molecular markers of parasite resistance and treatment outcome in falciparum malaria. *Malar J 8*: 89
49. Vauterin P, Jeffery B, Miles A, Amato R, Hart L, Wright I, Kwiatkowski D., 2017. Panoptes: web-based exploration of large scale genome variation data. *Bioinformatics 33*: 3243–3249

doi:10.20944/preprints202503.0561.v2

50. Soremekun S, Conteh B, Nyassi A, Soumare HM, Etoketim B, Ndiath MO, Bradley J, D'Alessandro U, Bousema T, Erhart A, Moreno M, Drakeley C., 2024. Household-level effects of seasonal malaria chemoprevention in the Gambia. *Commun Med (Lond) 4*: 97
51. Thwing J, Williamson J, Cavros I, Gutman JR., 2024. Systematic Review and Meta-Analysis of Seasonal Malaria Chemoprevention. *Am J Trop Med Hyg 110*: 20–31
52. Deutsch-Feldman M, Aydemir O, Carrel M, Brazeau NF, Bhatt S, Bailey JA, Kashamuka M, Tshefu AK, Taylor SM, Juliano JJ, Meshnick SR, Verity R., 2019. The changing landscape of Plasmodium falciparum drug resistance in the Democratic Republic of Congo. *BMC Infect Dis 19*: 872
53. Nankabirwa J, Brooker SJ, Clarke SE, Fernando D, Gitonga CW, Schellenberg D, Greenwood B., 2014. Malaria in school-age children in Africa: an increasingly important challenge. *Trop Med Int Health 19*: 1294–1309
54. Okiring J, Epstein A, Namuganga JF, Kamya EV, Nabende I, Nassali M, Sserwanga A, Gonahasa S, Muwema M, Kiwuwa SM, Staedke SG, Kamya MR, Nankabirwa JI, Briggs J, Jagannathan P, et al., 2022. Gender difference in the incidence of malaria diagnosed at public health facilities in Uganda. *Malar J 21*: 22
55. Tessema S, Wesolowski A, Chen A, Murphy M, Wilheim J, Mupiri A-R, Ruktanonchai NW, Alegana VA, Tatem AJ, Tambo M, Didier B, Cohen JM, Bennett A, Sturrock HJ, Gosling R, et al., 2019. Using parasite genetic and human mobility data to infer local and cross-border malaria connectivity in Southern Africa. *Elife 8*
56. Anon. WHO external quality assurance scheme for malaria nucleic acid amplification testing. Available at: https://www.who.int/teams/global-malaria-programme/case-management/diagnosis/nucleic-acid-amplification-based-diagnostics/faq-nucleic-acid-amplification-tests. Accessed
57. Cunningham JA, Thomson RM, Murphy SC, de la Paz Ade M, Ding XC, Incardona S, Legrand E, Lucchi NW, Menard D, Nsobya SL, Saez AC, Chiodini PL, Shrivastava J., 2020. WHO malaria nucleic acid amplification test external quality assessment scheme: results of distribution programmes one to three. *Malar J 19*: 129
58. Mideo N, Kennedy DA, Carlton JM, Bailey JA, Juliano JJ, Read AF., 2013. Ahead of the curve: next generation estimators of drug resistance in malaria infections. *Trends Parasitol 29*: 321–328
59. Ruybal-Pesántez S, Amaya-Romero J, Bérubé S, Brazeau NF, Diop MF, Hathaway N, Hendry J, McCann K, Murie K, Murphy M, Niaré K, Phelan J, Schaffner SF, Simkin A, Taylor AR, et al., 2025. Towards an open analysis ecosystem for Plasmodium genomic epidemiology
60. Anon., 2022. Seamless sharing and peer review of code. *Nat Comput Sci 2*: 773
61. Senn SJ., 2009. Overstating the evidence: double counting in meta-analysis and related problems. *BMC Med Res Methodol 9*: 10
62. Moodley K, Cengiz N, Domingo A, Nair G, Obasa AE, Lessells RJ, de Oliveira T., 2022. Ethics and governance challenges related to genomic data sharing in southern Africa: the case of SARS-CoV-2. *Lancet Glob Health 10*: e1855–e1859
63. Piasecki J, Cheah PY., 2022. Ownership of individual-level health data, data sharing, and data governance. *BMC Medical Ethics 23*: 1–9
64. Bull S, Bhagwandin N., 2020. The ethics of data sharing and biobanking in health research. *Wellcome Open Res 5*: 270

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.