Concept Paper

Towards Setting Minimum and Optimal Data to Report for Malaria Molecular Surveillance (MMS) 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 , Jessica Lin , Gina Cuomo-Dannenburg , Shazia Ruybal-Pesántez , Oliver Watson , Robert Verity , Jeffrey Bailey [*]

Concept Paper

Towards Setting Minimum and Optimal Data to Report for Malaria Molecular Surveillance (MMS) 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 Connelly** [7], **Jonathan B. Parr** [1,2,4], **Melissa D. Conrad** [8], **Jacob M. Sadler** [1], **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], **Jessica T. Lin** [1,2], **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: 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 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 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 allele or microhaplotype data is central to maximizing 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 is essential to advance research and strengthen global public health efforts.

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 threats. 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 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, however, including 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 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 use of pathogen genomics to inform public health decision-making.

Molecular surveillance has now become integral to understand 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. 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

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 essentially "contaminated" by the host DNA. Low parasitemia samples are often challenging to sequence and require enrichment techniques to capture enough parasite genomic material. [16] Mixed infections of different species or genotypes can occur at low relative abundances, which limits their detection relative to the sequencing error rate. [17] Additionally, eukaryotic recombination continually reassorts parts of the parasite genome, obfuscating relationships between parasites and making tracking transmission chains or the parasite origin difficult. Beyond complex biology, 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 malaria molecular surveillance (MMS) and instead less-expensive targeted sequencing approaches, such as multiplexed amplicon deep sequencing or molecular inversion probes (MIPs), [18–24] are often employed. Targeted sequencing methods are also more sensitive and better for low parasitemia samples. [25] To date, compared to WGS, bioinformatic tools have been relatively ad hoc and optimized for specific questions of interest. [17,26–28] Thus, a major challenge with targeted MMS is the heterogeneity in assays and analyses, and the need to ensure that data can be combined and analyzed in a rigorous way.

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. [29–31] These mutations have reached a prevalence of over 20% in many of these areas. [29–31] K13 polymorphisms have also been sporadically found across Africa in other locations and are cropping up now in Southern Africa. [32] 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 (ACT) partner drugs; 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) by determining if parasites found before and after treatment are the same strain. [28,33] 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. [34–37]

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 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. **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..

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 and allowing the most extensive use of data for other scientific questions. [38–40] 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 variant data with publication would allow the community to take full advantage of the previous analysis of sequencing data to integrate across data sets. 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 (all variation from a single amplicon) represents lossless data in terms of variation and is the optimal format for polymerase chain reaction (PCR)-based targeted sequencing. Haplotypes in genes or proteins are important, as combinations of polymorphisms impact antimalarial resistance, in particular for *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 [41–45]]. 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 assessed. It is essential that variant reporting also meets FAIR requirements. [40] 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 is 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. [46]

 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 be made available. First, geographic location of sample collections should be as precise as possible to allow for accurate mapping, intervention deployment, and understanding of impacts. [23,47] 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. [47–49] 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. [50,51] Fourth, human mobility is a critical aspect for studying malaria importation and migration of parasites, hence travel data should be included where available. [52] 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 background frequency of mutations due to the selective pressure within individuals making resistance parasitaemia 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.

 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. [53,54] 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. [23,24] 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. [55]

Specific technology and kits used for sequencing, as well as details of filtering and data processing, are needed. This requires deposition of code in publicly available repositories, such as GitHub or Code Ocean "compute capsules", which allows code to be run in a standalone environment with proper version control. [56] Issues around bioinformatics code and genomics data analysis tool usability and availability have been recently reviewed using defined software standards criteria. [40]

Finally, data is 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. [57]. 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.

While the goal of sharing individual-level data is 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. [58] 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. [59] 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. [58,60] 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 "how?" 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 VCFs are an excellent choice. Targeted sequencing also often employs these same formats but data is lost for non-variant regions and haplotypic linkage in mixed infections. Better formats for reuse include GVCFs (or more compressed ReblockGVCF) 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://www.plasmogenepi.org/DataStandards), is an attractive lossless intermediate under active development. Longer-term, development of robust metadata formats that leverage well-defined epidemiological vocabulary will be useful, particularly if formats can jointly hold sequence metadata (e.g. haplotypes) to promote harmonization of analysis pipelines with negligible raw data loss. 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 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 health of those affected by this terrible disease.

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 is 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 and analysis with metadata.	Fully reproducible coding pipeline that takes data and produces all results and figures from the main

		analysis.
	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. [56]	
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 attainable.*	Highest resolution data possible (GPS location of household, clinic of collection, town/city of collection) at individual level data (jittered if considered potentially 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 be delineated -- 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, travel in the last 28 days.	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.*	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.*
Read or UMI depth	Number of reads or unique molecular identifiers (UMI) informing each genotype (SNP or combination of SNPs) reported at each loci by individual. Total number of reads per loci 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. *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 580Y 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.*	Within-sample allele frequencies for each participant; these can be calculated directly or from read depth/UMI counts.

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. 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

17. 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

18. 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

19. 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

20. 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

21. 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

22. 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

23. 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

24. 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

25. 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

26. 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

27. 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

28. 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

29. 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

30. Rosenthal PJ, Asua V, Conrad., 2024. Emergence, transmission dynamics and mechanisms of artemisinin partial resistance in malaria parasites in Africa. *Nature reviews Microbiology 22*

31. 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*

32. 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

33. Holzschuh A, Lerch A, Nsanzabana C., 2024. Multiplexed nanopore amplicon sequencing to distinguish recrudescence from new infection in antimalarial drug trials

34. 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*

35. 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

36. 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

37. 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

38. Mathur MB, Fox MP., 2023. Toward Open and Reproducible Epidemiology. *Am J Epidemiol 192*: 658–664

39. Peng RD, Dominici F, Zeger SL., 2006. Reproducible epidemiologic research. *Am J Epidemiol 163*: 783–789

40. 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

41. Blasco B, Leroy D, Fidock DA., 2017. Antimalarial drug resistance: linking Plasmodium falciparum parasite biology to the clinic. *Nat Med 23*: 917–928

42. Fidock DA, Eastman RT, Ward SA, Meshnick SR., 2008. Recent highlights in antimalarial drug resistance and chemotherapy research. *Trends Parasitol 24*: 537–544

43. 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

44. Conrad MD, Rosenthal PJ., 2019. Antimalarial drug resistance in Africa: the calm before the storm? *Lancet Infect Dis 19*: e338–e351

45. 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

46. 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

47. 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

48. 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

49. 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

50. 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

51. 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

52. 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*

53. 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

54. 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

55. 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

56. Anon., 2022. Seamless sharing and peer review of code. *Nat Comput Sci 2*: 773

57. Senn SJ., 2009. Overstating the evidence: double counting in meta-analysis and related problems. *BMC Med Res Methodol 9*: 10

58. 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

59. Piasecki J, Cheah PY., 2022. Ownership of individual-level health data, data sharing, and data governance. *BMC Medical Ethics 23*: 1–9

60. Bull S, Bhagwandin N., 2020. The ethics of data sharing and biobanking in health research. *Wellcome Open Res 5*: 270