MalariaGEN Consortium
Internal Data Management and Access Policy

MalariaGEN provides structure and funding which enables large-scale genotype research and gives local investigators across Africa and Asia access to technologies and data that might otherwise be out of reach. This empowers local researchers in malaria-endemic countries to analyse genotype and phenotype data using their own parameters and builds local research capacity. The orders-of-magnitude increase in the scale of genetic data collection that this implies, and the number of partners involved, has created challenges for the MalariaGEN consortium. Recent developments in genomic research mean that data can now readily be generated at levels of detail unique to individuals, with the potential that this data could be linked back to specific individuals. This has led to concerns in the relevant literature about the risks of 'erroneous or malicious identity exposure and consequent embarrassment; legal or financial ramifications; stigmatisation and/or discrimination'.[1]

Given these complexities, it is important for the MalariaGEN consortium to agree on an appropriate model for data management which minimally impedes access by qualified investigators but keeps the risks of identifying individuals very low. The three guiding principles for this model are:

1. The privacy and confidentiality of research participants are protected.
2. Large-scale genotype research is undertaken which generates genetic discoveries that translate into an increased understanding of malaria.
3. Contributing principal investigators (PIs) are enabled to carry out analysis of the genotype and phenotype data produced from the samples they collected, and, where appropriate, link these with locally-held phenotypic databases. This recognises that:
   - Sovereignty of physical samples and clinical data contributed to the consortium experiments remains with the contributing PI[2] and his/her institution.
   - Due to the on-going and iterative nature of genomic epidemiological research there might be a requirement for further research to be carried out on the sample or for further information to be collected from participants. This necessitates having 'coded' rather than 'unidentified' or 'unlinked' samples.[3]

Putting these principles into practice can present important ethical issues. For example, providing the sites that collected samples with genetic and clinical data relating to those samples in such a way that it can be linked to locally-held phenotypic databases may be perceived as breaching the anonymity of the genetic data. However, PIs are already entrusted with protecting the privacy of their research participants in other studies and, in many cases, are already carrying out research which joins up genomic and clinical data. This local research is governed by each institution's regulations and by the permissions of local ethics review boards.

In addition, even if samples are 'coded' it also needs to be recognised that, despite best intentions and efforts in ensuring confidentiality, modern DNA identification techniques can link a sample with an individual given enough time and effort. This model therefore aims to ensure that the privacy of participants is protected by maintaining a high threshold for re-linking a genotyped sample to the participant; that this privacy be guarded by all investigators in line with the relevant guidance, regulation and ethical approvals; and that the

[1] W. Lowrence and F Collins (2007). *Identifiability in Genomic Research*, Policy Forum, Science, Vol 317., p. 600
[2] As set out in MalariaGEN's Joint Policy on Data Sharing, IP and Publications
[3] See the glossary of terms at the end of this document.

contributing PI and the genotyping centre have joint responsibility to ensure certain conditions are met before re-linking takes place (see MalariaGEN's **Standard operating procedure for re-linking genomic data with local clinical databases**).

Model for data management and access

It is important to be clear about the sovereignty of data and levels of confidentiality in genetic databases. Within this model the degree of identifiability of the sample and the responsibilities of participating institutions change as the sample collection, processing and analysis of data progresses. The model takes a practical approach to confidentiality by coding samples both during sample collection and again during processing of samples received at the Oxford Coordinating Centre. This seeks to achieve a high level of confidentiality. Making the link all the way back to the name of the participant will require the willing participation of the PIs of the Oxford Coordinating Centre and also of the investigators at the site from which the sample originated. This structure can be classified as 'coded' confidentiality. The process for managing data within the MalariaGEN consortium (as summarised in Figure 1) is:

Contributing PIs' sovereignty over samples and data:

- *Sample/data collection:* blood samples and clinical/demographic data are collected. These are coded by assigning a typically alphanumeric, serial local identifier using a system defined locally (source code).
- *DNA extraction and transfer:* following DNA extraction, a portion of each sample, labelled with its source code, is sent to Oxford for consortial genotyping/sequencing experiments. The sample manifest contains a list of the source codes with no other personal identifying information. In the case of familial samples or replicate samples, an extra piece of information is provided (typically within the source code) to identify the relationships between samples (the identity of maternal and paternal samples if any for a given sample, or the identity of an existing sample from a given individual) between samples. This information is retained in the MalariaGEN database as it is essential for various types of data analyses.
- *Plasma extraction and transfer:* in some cases plasma/serum samples, with their corresponding source codes, are extracted and sent directly to an authorised contractor for immunoassay.
- *Clinical and demographic data transfer:* data files are supplied to Oxford separately with the corresponding source codes so that they can be joined to the DNA/plasma samples. No datasets that contain personal identifiers are accepted by Oxford.

Stewardship of samples and data by the Oxford Coordinating Centre:

- *Sample processing:* In Oxford the sample is then recoded with new standardised identifiers (lab code) for consistent handling and analysis. The source code is securely stored in a central database and plays no further part in sample processing or handling.
- *High-throughput genotyping:* DNA samples are genotyped either in Oxford, the Sanger Institute (where the sample is further re-coded), or by authorised contractors.
- *Plasma immunoassay:* Like DNA, when plasma/serum samples arrive, they are re-coded by the authorised handling laboratories. The source code then plays no further part in sample processing or handling.
- *Phenotype processing:* Where appropriate, clinical/epidemiological data is processed into phenotypes according to MalariaGEN consortial standards and user-specific needs.

Co-sovereignty of, and responsibility for, site-specific merged genotype and phenotype data by contributing PIs and Oxford Coordinating Centre:

In order to ensure that each contributing PI, the Oxford Coordinating Centre and the relevant ethics review boards are confident that the three guiding principles are met, we propose a model of co-sovereignty for site-specific merged genotype and phenotype data. At no point does the Oxford Coordinating Centre have sole sovereignty of the samples or data. The model is:

- *Genotype/phenotype data analysis:* merged genotype and phenotype data is produced and analysed by the consortial analysis team. Access to this data outside the consortia is defined by MalariaGEN's **Data Release Policy for Genome-wide Association Data**.
- *Sharing coded site-specific data:* Each site which contributes samples currently has access to a merged genotype and phenotype dataset with no site-specific identifiers but instead an assigned public code (the MalariaGEN released key – see glossary). Local sites can access this information for the samples they contributed through the secure website (www.malariagen.net). This website also provides an added level of security by only allowing authorised users access to specific data/sample sets (typically the samples that they have contributed). The format of this display is sufficient to analyse data for the association of any given genotype with a phenotype.
- *Providing means for sites to link data for local analysis:* Sites may require the means to link the genotype/phenotype data back to the original clinical/epidemiological data they hold. Reasons for this can include: 1) analyses using clinical parameters not sent to/held by MalariaGEN, 2) identification of samples for further investigation and/or 3) identification of related sample materials (e.g. plasma, tissue) for further research. If such a need arises, contributing PIs will request a re-linking file from the Oxford Coordinating Centre. When requesting a re-linking file, PIs will undertake to keep the re-linked data confidential (see MalariaGEN's **Standard operating procedure for re-linking genomic data with local clinical databases**).

Glossary of Terms

- 'Coded' samples are unidentified for research purposes, but can be linked to their sources through the use of a code. Decoding is the responsibility of the PI or another designated researcher.
- 'Identified' samples are those that allow the researcher to link the biological information derived from research directly to the individual from whom the material was obtained.[4]
- 'Personal Identifiers' are data such as names and addresses which clearly link samples back to identifiable individuals.
- 'MalariaGEN released key' is used to identify each sample when data is released (both to sites and to external researchers). This key is designed to contain no information about the source of samples or any other sample attributes.
- 'Re-linking file' is the file that contains the set of source codes and corresponding MalariaGEN released key sequences.
- 'Sovereignty' refers to having authority over the samples and data.
- 'Stewardship' refers to having responsibility and a duty of care but not sovereignty.
- 'Unidentified' samples are originally collected without identifiers and cannot be linked to their sources.
- 'Unlinked' samples are those that were originally identified, but have been irreversibly stripped of all identifiers and cannot be linked to their sources.

[4] See D. Chokshi and D. Kwiatkowski (2005). *Ethical Challenges of Genomic Epidemiology in Developing Countries.* Genomics, Society and Policy, Vol. 1, No. 1, pp. 1-15.

Figure 1: MalariaGEN's internal sample and data management system for Consortial Project 1