

Data Release Policy for Genome-wide Association Data

We are working on a scientific research project that will generate a lot of genetic data from samples collected from participants in malaria-endemic countries. We have an obligation to release these data to the wider scientific community¹ and also to protect the participants and communities being sampled. This policy aims to maximise the impact our data have in driving medical research forward, while protecting those who contributed to this valuable resource.

Background

Who belongs to the MalariaGEN consortium?

MalariaGEN stands for the Malaria Genomic Epidemiology Network. We are an international consortium of medical doctors, scientists, computer programmers and ethicists based in over 20 countries (see <http://www.malariagen.net/>). Each site has a principal investigator (PI) who is responsible for the research at that site and who contributes to the overall management of the project. All of the institutions involved in the consortium are not-for-profit research organisations or universities.

Who is funding this research?

This project is funded by two charities: the Bill & Melinda Gates Foundation in the USA (via the Foundation for the National Institutes of Health, also a charity) and the Wellcome Trust in the UK. The funding for this project comes through the Grand Challenges in Global Health Initiative which has the objective to 'provide access to affordable health solutions for the benefit of people most in need within the developing world and to ensure the broad availability of data and information to the scientific community'.

What is the scientific project?

The project aims to identify how the human body resists malaria infection because this is key information needed for the development of an effective vaccine. Genetic factors are one reason why different people respond differently to malaria infection. By discovering these genetic factors we hope to be able to work out the mechanisms that the body naturally uses to defend itself against malaria. This information will be of great value to vaccine developers.

Our method of discovery is to compare the genetic make-up of people with severe malaria (known as the cases) with the genetic make-up of people from the general population (known as the controls). Because we want to identify *novel* mechanisms of resistance in the body, we are attempting to screen all of the thousands of genes in the human genome in a systematic manner. This is called genome-wide association analysis. When we have identified specific genes of interest by genome-wide association analysis, we will study these specific genes in detail in multiple populations.

To determine the mode of action of the malaria resistance genes that we identify, we are measuring a range of responses to malaria. These include measures of parasite infection, measures of malaria illness and levels of immunological factors (antibodies) in the blood. Age, gender, area of residence and ethnicity are factors that we need to take into account so they will also be recorded.

¹ The term 'scientific community' applies to researchers working for public, not-for-profit or private organisations, or any combination of these

What data are used in the scientific project?

The data fall into four categories:

- **Ethnic data.** Ethnicity is an important factor in genetics so we are recording information on ethnicity for participants with severe malaria and for controls from the general population. This information allows us to compare cases from a particular ethnic group with controls from the same ethnic group.
- **Parents.** Sometimes ethnicity is hard to define or record so one alternative is to use a participant's parents as the controls. Both control types have different advantages and disadvantages so the database includes parents as controls and general population participants as controls.
- **Clinical and epidemiological data.** We are collecting detailed information about clinical signs and symptoms of malaria, and laboratory measurements to assess parasite infection, anaemia and other clinically relevant variables. We are also collecting information about age, gender and area of residence as they are needed for interpretation of the clinical data.
- **Genetic data.** This consists of information on thousands of common forms of genetic variation across the human genome. Most of these variants are single nucleotide polymorphisms (SNPs), i.e. commonly-observed changes in a single letter of the human DNA code. Technically this is known as genome-wide SNP genotyping data.

What are these data used for?

The genetic data can only be used for research purposes. The genetic data are generated by an accredited facility and is therefore not of the quality required for diagnostic purposes. It is very useful for research into malaria, and possibly other diseases, because so many genes are being screened.

Who owns the data?

The data are owned by the institution(s) that generated them. A list of institutions and the *MalariaGEN Joint Policy on Data Sharing, Intellectual Property and Publications* are available at <http://www.malariagen.net/>.

Release of pre-publication data

The release of pre-publication data from large resource-generating scientific projects was the subject of a meeting held in January 2003, the "Fort Lauderdale meeting", sponsored by the Wellcome Trust. MalariaGEN has agreed to follow these data-release principles and as such, these data are being released as a 'community resource project' as defined in the report of the Fort Lauderdale meeting. The conclusion from the meeting was that responsible use of the data is necessary to ensure that data producers will continue to participate in such projects. "Responsible use" was defined as allowing the data producers to have the opportunity to publish the initial analyses using the data. MalariaGEN maintains a list of analyses it will be conducting on <http://www.malariagen.net/>. Researchers granted access to released data are not permitted to submit their analyses or conclusions in these areas for publication during the embargo period.

Can an owner refuse to allow access to data that they have generated?

The consortium has given an undertaking to the project's funders that the consortium's genome-wide association data will be released to the scientific community, subject to appropriate protection measures. Data that cannot be released must not be contributed to the consortium's database (or the scientific project). *If for any reason we cannot release the data generated for a particular batch of samples, that batch of samples cannot be used in the MalariaGEN project.*

Data Release Policy for Genome-wide Association Data

This policy covers genome-wide association data on individual participants. For the purposes of this policy, 'primary datasets' are sets of raw data that have passed appropriate data quality control checks. The datasets to be released **will not contain any personal identifiers or links to personal identifiers**, but will contain information about country of origin and will sometimes contain information about ethnicity.

Researchers working in the field of medical research can request access to these datasets through a data access committee. Access will only be granted once the investigator has signed a legally-binding agreement which states that:

- The data will only be used for the specified research;
- The data will not be used in any way that could reasonably be expected to lead to ethnic stigmatisation;
- No attempt will be made to identify the participants;
- The data will not be transferred to other researchers outside of the named research group;
- Primary data relating to individuals may not be published. Users may only publish aggregate or generic information derived from the data on the condition that such aggregate or generic information does not allow an individual participant to be identified with reasonable effort;
- All communications, publications and intellectual property (IP) arising from the use of the data will acknowledge the source of the data and will refer to the primary MalariaGEN publication(s) that describe the data;
- If use of the data gives rise to IP that could support health benefits in the developing world, the owner of the IP agrees to license it on a reasonable basis for use in the developing world and on a preferential basis to the countries whose citizens are the subject of the database.

Each dataset will be made available nine months after MalariaGEN investigators at the local study site first have access to that dataset. Datasets may be released earlier with the agreement of the relevant MalariaGEN investigators. MalariaGEN investigators retain the exclusive right to publish planned analyses of the released datasets for a defined period of time, as advised on the MalariaGEN website.

We are **not** proposing to charge a fee for access to the data unless this becomes necessary in the future solely for the purpose of recovering the costs of making the data available.