MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

Plasmodium falciparum Community Project

About this document

This document describes the purpose and structure of the MalariaGEN *Plasmodium falciparum* Community Project, as well as outlining means of releasing data and how that data may be used. As the project evolves, this document will be revised to reflect any changes and updated versions will be posted on the MalariaGEN website (http://www.malariagen.net/projects/parasite/pf).

Community Project goals

Enable malaria researchers to analyse genome variation in the parasite samples that they collect

Advances in sequencing technology have made it possible to analyse parasite genome variation in samples collected during clinical and epidemiological studies of malaria but relatively few groups have the statistical and computational resources needed to deal with raw sequence data. To address this need, MalariaGEN has established a sequence data analysis pipeline that enables researchers who have collected parasite samples to generate sequence read data and translate this data into standardised genotype calls that are suitable for epidemiological and population genetic studies, e.g. of drug resistance or antigenic variation.

Provide the malaria research community with a collaborative framework to investigate common forms of genetic variation in *P. falciparum*

Data on the natural diversity of *P. falciparum* in different geographical regions is crucial for many aspects of malaria biology and disease control. This depends on clinical and epidemiological studies that often require a long-term investment to achieve their scientific objectives. The Community Project enables different research groups to share and release data on parasite polymorphism without compromising their ability to analyse and publish their own findings. Working together in this way, the Community Project is creating an integrated, global view of parasite diversity that can help underpin malaria research and control efforts.

Specific objectives

1. Build a catalogue of variation and allele frequencies

The Community Project provides a catalogue of variation in the *P. falciparum* genome based on sequence data analysis, together with estimated allele frequencies in different geographical regions based on aggregated analysis of all the samples and data that have been contributed. It is not possible with current methods to obtain a complete and accurate genome sequence assembly in individual field samples, but by aggregating data across many samples it is possible to identify thousands of variations with a high level of confidence and to estimate their frequencies in different groups of samples. As of August 2015, the catalogue includes more than 900,000 single nucleotide polymorphisms (SNPs) in exonic regions of the genome, based on an analysis of 3,488 samples from 43 separate locations in 23 countries.

This approach tends to be over-conservative, i.e. true variants may be omitted from the catalogue if they cannot be genotyped with confidence using the current version of the analytical pipeline. As sequencing technologies and analytical methods improve, future versions of the catalogue will progressively extend to other regions of the genome that are more difficult to analyse, and will include other forms of polymorphism such as indels and structural variants.

2. Provide researchers with standardised genotype data on their samples
While it is becoming increasingly easy to generate large amounts of sequence data on parasite field samples using next-generation technologies, it remains statistically and computationally challenging to translate raw sequence data into high-quality genotype calls that are required by most researchers for their own investigations, e.g. of phenotype-genotype associations or parasite population genetics. Moreover the accuracy of genotype calls on individual samples can be improved if they are analysed jointly with many other samples, which helps to distinguish true variants from sequencing errors. The Community Project provides a sequence data analysis pipeline to provide researchers with high-quality genotype calls on their samples in a standardised format which enables reliable comparisons to be made with data from other studies and geographical locations.

3. Publish global analyses of genome variation, population genetics and evolutionary selection
The Community Project will report on the major geographical divisions of parasite population structure and use this to calculate allele frequency data. The Community Project will also analyse other aspects of genome variation, population genetics and evolutionary selection that can best be achieved using the aggregated dataset rather than by individual research groups. These global analyses will be reported through peer-reviewed publications subject to the agreement of the groups who have contributed samples and data. Where appropriate, this will be done through mechanisms such as the Pf3k Project (www.malariagen.net/apps/pf3k) that bring together expert working groups and integrate multiple sources of data to produce specific analytical outputs.

4. Create online tools to maximise the reach and impact of Community Project data and findings
The Community Project has developed a set of web tools to enable contributing researchers to explore and analyse sequence read data and genotype calls on their own samples. We have also partnered with the MRC Centre of Genomics and Global Health to build a comprehensive web application to provide the wider research community with user-friendly tools to browse and query the substantial data resources generated by the Community Project (www.malariagen.net/apps/pf). An important feature of this public web application is to provide information about the community of researchers working on this collective endeavour, and about the samples they have contributed and the studies they are undertaking.

Community Project structure

The Community Project is coordinated by the MalariaGEN Resource Centre and comprised of partner studies, independent studies undertaken in malaria endemic areas.

Partner studies

Each partner study is unique with their own research objectives. They have agreed to contribute samples to the Community Project on the understanding that this will not interfere with their research objectives.

Prior to submitting samples, all partner studies complete a Partner Study Information Form that captures information about their study, and confirms that all relevant ethical and regulatory requirements have been met and that all stakeholders have agreed to contribute samples and data to the Community Project. We expect that partner studies have collected their samples and generated their data according to good research practice. Each partner study is represented on the Community Project website with a brief description of the study, and details of the study contact person, key associates and their affiliations.

Contact person

Each partner study designates a single person who is responsible for the partner study as it pertains to the Community Project. In many cases the contact person will be the lead investigator of the partner study, but for large or complex studies it may be someone who serves a coordinating role within a collaborating group of investigators.

The contact person will be listed publicly on the website and/or the web application, and is expected to field any questions about the samples and data contributed to the Community Project. Questions may range from technical queries to internal or external requests for access to data or for collaboration. The contact person is responsible for ensuring that the individuals and institutions that contributed the samples and generated the sequence data are informed and consulted as appropriate.

MalariaGEN Resource Centre

The MalariaGEN Resource Centre is responsible for sample and data management, large-scale sequence analysis and genotype calling, coordinating global analyses and data outputs of the Community Project, and building web applications for internal and external data access. The MalariaGEN Resource Centre is currently funded by a Wellcome Trust Strategic Award to Oxford University and the Wellcome Trust Sanger Institute, with additional support through the MRC Centre for Genomics and Global Health. For more information about the MalariaGEN Resource Centre, see http://www.malariagen.net/community/resource-centre.

Community Project workflow

- Partner studies contribute parasite DNA samples, each collected from an infected patient, along with sample information like where and when each sample was collected, and a description of why the samples were collected and any planned use of the genomic data generated by the Community Project.
- The samples are sequenced in collaboration with the Wellcome Trust Sanger Institute.
- The MalariaGEN Resource Centre provides each partner study with standardised genotype calls on their samples, based on aggregated analysis of all the samples contributed to the Community Project, for their own analyses.
- The MalariaGEN Resource Centre periodically produces a catalogue of variation and allele frequencies based on aggregated analysis of all samples contributed.
- The Community Project publishes aggregate analyses on global patterns of parasite population structure, as well as other aspects of genome variation and evolutionary selection.

Data release

Potential data users are asked to respect the legitimate interests of the Community Project and its contributing investigators and abide by any restrictions on the use of the data as described in the Terms of Use (see below).

Sequence data and sample information

Sequence read data generated by the Wellcome Trust Sanger Institute are routinely deposited in the European Nucleotide Archive and is publicly available through the ENA website (http://www.ebi.ac.uk/ena/home) unless there is a strong legal or regulatory impediment to doing so.

Sample information such as the ENA accession ID, location of collection, and the contact person for the partner study that contributed the data, can be downloaded from the MalariaGEN website. (https://www.malariagen.net/data/pf-sample-info). This sample information is released under Terms of Use.

Variant catalogue and allele frequency data

The most recent public release of the catalogue of variants and allele frequency data is available in the Community Project's web application (www.malariagen.net/apps/pf), which has been developed to enable users to browse and query the data in detail, and to view information about partner studies and sampling locations.

The web application will be updated whenever a new public catalogue of variants and allele frequency data becomes available. Previous catalogues used in Community Project publications are available for download as VCF files on the MalariaGEN website (www.malariagen.net/data).

Providing genotype data to partner studies

The MalariaGEN Resource Centre produces quality-controlled genotype calls on samples contributed to the Community Project and periodically updates these data as analytical methods improve. Genotype data are provided to partner studies in major and minor data releases. For major releases, sequence data for all samples in the Community Project is pooled and used for *de novo* SNP calling. The resulting catalogue of variation is then used to generate genotype calls for every sample in the Community Project. Minor releases provide genotype calls for recently sequenced samples, and are made using the SNP call set from the last major release.

With each new release of genotype data, an email notification is sent to the contact person for each study. The contact person can access the genotype data via a password-protected website which provides user-friendly tools to explore and download these data.

Where appropriate, the contact person can request for additional members of the partner study to receive the email notifications and be granted data access by emailing support@malariagen.net.

Terms of Use

The Community Project is a collaboration designed to support independent research studies to perform genomic analyses on their own samples, while at the same time enabling global analyses of *P. falciparum* population genetics and evolutionary selection. The Terms of Use are intended to respect the legitimate interests of all collaborators who have contributed samples and data to the

Community Project. Terms of Use are based on those established by the human 1000 Genomes Project (www.1000genomes.org) and guided by the Fort Lauderdale Agreement for Sharing Data from Large-scale Biological Research Projects (http://www.genome.gov/pages/research/wellcomereport0303.pdf). In brief, we encourage others to use Community Project data in their own work, but expect that they will allow the Community Project and its contributing investigators to make the first presentations and publications reporting on their intended scientific analyses. For details, see the full Terms of Use: https://www.malariagen.net/projects/parasite/pf/use-p-falciparum-community-project-data.

Any use of Community Project data implies acceptance of the Terms of Use.

Publications

All publications are subject to the Terms of Use (see above).

Project publications

The Community Project will report on the major geographical divisions of parasite population structure and use this to calculate allele frequency data. The Community Project will also analyse other aspects of genome variation, population genetics and evolutionary selection that can best be achieved using the aggregated dataset rather than by individual research groups. These global analyses will be reported through peer-reviewed publications subject to the agreement of the groups who have contributed samples and data.

Depending on the nature of the work and data used, publications will either have traditional named authorship or banner authorship. In all cases, all those who satisfy the conventional conditions of authorship will be listed in the main publication while other relevant contributions will be listed in the Supplementary Material. The MalariaGEN Resource Centre is responsible for the production of Community Project publications and will work with others to lead particular areas of analysis where appropriate.

Partner study publications

Partner studies are free to use and publish analyses based on their own sample data, provided that they do not conflict with intended Community Project analyses (see Terms of Use). This can include research collaborations with other groups. The partner study contact person is responsible for ensuring that all those who have provided samples and sequence data are kept informed and consulted as appropriate about work that might lead to partner study publications (see Contact Person above).

External publications

The Community Project data have many potential uses that fall outside the scope of Community Project publications and partner study publications (see Terms of Use). Researchers using Community Project data are asked to inform the MalariaGEN Resource Centre about manuscripts in preparation and to provide appropriate acknowledgement and attribution.

Attribution

All publications using Community Project data should acknowledge and cite the source of the data using the following format: "This publication uses data from the MalariaGEN *Plasmodium falciparum* Community Project (www.malariagen.net/projects/parasite/pf) as described in [cite the relevant Community Project publications]. Genome sequencing was performed by the Wellcome Trust Sanger Institute and the Community Project is coordinated by the MalariaGEN Resource Centre with funding from the Wellcome Trust (098051, 090770).

Contact Us

To explore opportunities to collaborate or for any other queries, please contact support@malariagen.net.