MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK
www.malariagen.net

Supported by wellcome trust, wellcome trust sanger institute, MRC Medical Research Council, Bill & Melinda Gates foundation, Foundation for the National Institutes of Health

The Wellcome Trust Centre for Human Genetics, University of Oxford

Bayesian analysis of genetic association with severe Malaria

Gavin Band, Geraldine Clarke, Matti Pirinen, Kirk Rockett, The MalariaGEN consortium, Chris Spencer.

Introduction

Malaria is an infectious disease prevalent in sub-Saharan Africa, south-east Asia, and elsewhere, which causes an estimated 600,000 deaths each year, mostly among children under five. Many human genetic associations with malaria susceptibility have been reported, but few have been successfully replicated [1]. As part of the MalariaGEN consortium we typed 55 previously reported loci in a sample of 12,000 cases and 17,000 controls from across sub-Saharan Africa, south-east Asia and Oceania. Genetic variants in five regions (sickle cell locus, *ABO* blood group, *ATP2B4*, *G6PD* and *CD40LG)* showed strong evidence for association.

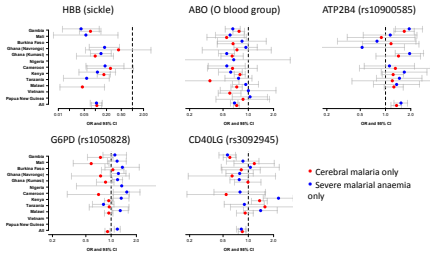

Figure 1. Effect sizes across sub-Saharan Africa for Severe Malarial Anaemia (red) and Cerebral Malaria (blue) subphenotypes at five malaria-associated SNPs.

Bayesian analysis

Several factors, including human and parasite genetic diversity, differences in environment, and differences in etiology might lead to observed patterns of effect heterogeneity. To quantify the degree of heterogeneity between populations and between subphenotypes, we compared models of effect heterogeneity in a Bayesian framework. We use multivariate normal priors to express models of effect that are fixed (i.e. equal), correlated or independent across sites, and fixed, correlated or independent across phenotypes. Explicitly, we use priors of the form

$$(\beta_1, \beta_2, \cdots)^t \sim MVN(0, \Sigma P \Sigma)$$

where

- β_i = the ith effect size parameter,
- Σ = a diagonal matrix whose ith entry σ_i specifies the prior standard deviation of β_i. (Thus Σ determines the magnitude of plausible effects under the prior.)
- P = a correlation matrix expressing the prior pattern of correlations between effect parameters.

The same framework can be used to model different modes of inheritance – e.g. additive effects, where the two alleles at a locus act independently, as well as over- or under-dominance.

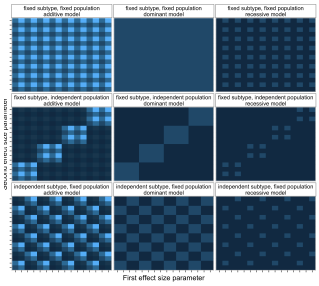

Figure 2. Illustration of prior covariance matrices ($\Sigma P \Sigma$) expressing additive, dominant or recessive modes of inheritance and effects that are fixed or independent across sites and subphenotypes. In this figure there are four effect parameters per population: one for each subphenotype, for heterozygote and homozygote genotype classes.

To apply these models, we fit a multinomial model in each population with the three possible outcomes *control*, *cerebral malaria* (CM) case or *severe malarial anaemia* (SMA) case by maximum likelihood using the nnet package [2] in R. To compute Bayes factors efficiently, we approximate the likelihood function up to a constant by the density of a multivariate normal distribution with mean equal to the combined vector of parameter estimates (denoted $\hat{\beta}$) across populations, and block-diagonal variance-covariance matrix V with i^{th} block equal to the inverse of the observed information in the i^{th} population. With this approximation and assuming independent priors for parameters that are not genetic effects, the Bayes factor can be computed simply as a ratio of normal densities,

$$BF = \frac{MVN\left(\hat{\beta}; 0, V + \Sigma P \Sigma\right)}{MVN\left(\hat{\beta}; 0, V\right)}$$

We assess the overall evidence for association by model averaging over plausible models, and assess the evidence for heterogeneity by comparing component models.

In the univariate case this *approximate Bayes factor* was used by Wakefield [3-4]. We have previously used similar methods to study effect heterogeneity between disease subtypes [5] and populations [6].

Results

Five loci (sickle cell locus, *ABO* blood group, *ATP2B4*, *G6PD* and *CD40LG*) showed strong evidence of association.

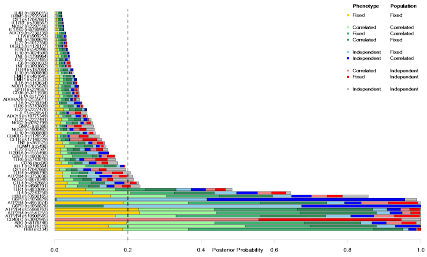

Figure 3. Posterior probability of association for each of 55 SNPs tested in this study [1], assuming that either the null model of no association or one of the models of heterogeneity holds. The dashed line indicates the prior probability of association for each SNP, here taken to be 20%. Colours represent the contribution to the posterior from each model of association (here given equal prior weight).

At all five loci, power to detect the effect was greatest for models that allow some heterogeneity, while for two loci (*G6PD* and *CD40LG*) there was strong evidence of heterogeneity between populations (*CD40LG*) and phenotypes (*G6PD*). Importantly, associations at these loci would not have been detected using fixed-effect meta-analysis techniques.

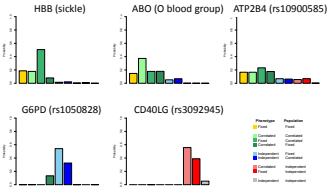

Figure 4. Posterior probability on each model of association at 5 malaria-associated SNPs, assuming one model holds. Models are given equal prior weight.

Dissecting patterns of between-population and phenotypic heterogeneity is potentially informative about the genetic etiology of disease. We advocate the approach presented here as a simple, efficient way to test for and deconstruct complex patterns of genetic effects.

[1] The MalariaGEN Consortium, "*Reappraisal of known malaria resistance loci in a large multi-centre study*", in submission.
[2] W. N. Venables and B. D. Ripley, "*Modern Applied Statistics with S, Fourth Edition*", Springer, New York (2002)
[3] J. Wakefield, "*A Bayesian measure of the probability of false discovery in genetic epidemiology studies*", Am .J. Hum. Genet. (2007)
[4] J. Wakefield, "*Bayes factors for genome-wide association studies: comparison with p-values*", Gen. Epidem. (2009)
[5] Bellenguez et al, "*Genome-wide association study identifies a variant in HDAC9 associated with large vessel ischemic stroke*", Nat Genet (2012)
[6] Band et al, "*Imputation-Based Meta-Analysis of Severe Malaria in Three African Populations*", PLoS Genetics (2013)

Scaling up across the genome

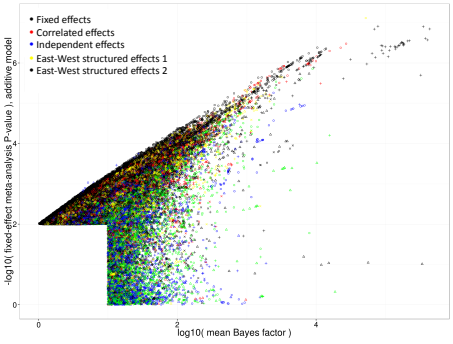

Figure 6. Genome-wide comparison of model-averaged Bayes factors for additive (○), dominant (+), recessive (x), and heterozygote (△) models (x-axis) and fixed-effect meta-analysis P-value for additive mode of inheritance (y-axis). Colours indicate the prior correlation structure with the maximum posterior probability. This plot represents all autosomal loci post-imputation, except that regions of HBB and ABO are excluded.

We applied the method presented above to a genome-wide scan of severe Malaria in Gambia, Kenya, and Malawi (a subset of the samples presented above.) Individuals were typed on the Illumina HumanOmni 2.5M array and we imputed into the 1000 Genomes reference panel to obtain around 20 million SNPs and indels of use for association testing. We developed custom software to compute Approximate Bayes Factors efficiently at millions of variants, and use this to compare models of between-population heterogeneity of genetic effect on Severe Malaria. In view of the geographic distribution of populations, we include models where the two East African populations are more similar to each other than to Gambia, and allow for both smaller and larger effect sizes (Figure 5). This approach to variant discovery leads to detection of variants that may not have been identified using traditional fixed-effect models (Figure 6). A replication experiment is in progress.

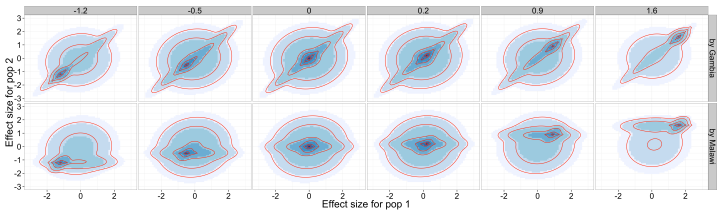

Figure 5. Two-dimensional slices through the overall prior on genetic effects determined by the bayesian model average. We use a weighted average of fixed, correlated and independent effects models, and *structured effects* models that allow for differences between East and West Africa. Top row: comparison of effect size distributions in Kenya (y-axis) and Malawi (x-axis), for constant effect size in Gambia. Bottom row: effect size distributions in Kenya (x-axis) and Gambia (y-axis) for constant effect size in Malawi. Colouring is on log10 scale. Red lines show contours of the density at powers of 10 from the maximum. While most mass is near the diagonal, reflecting a prior belief in strongly correlated effects, some weight is placed on models that allow geographically separated populations to differ.

Acknowledgements. We thank the individuals and families who contributed to this study, and the relevant funding bodies (see top of poster) for their generous support.

Contact: *gavin.band@well.ox.ac.uk*