Pathogen-driven selection and human genetic variability: the case of protozoa

Pozzoli U[1], Fumagalli M[1,2], Cagliani R[1], Comi GP[3], Bresolin N[1,3], Clerici M[4,5], Sironi M[1]

Motivation

Each year 300-500 million people develop malaria and 1.5-3 million die. Plasmodium and other protozoan genera show a widespread geographic distribution and the prevalence of protozoan infection has likely been high throughout human history. Protozoan borne-diseases are therefore considered to have exerted the strongest selective pressure in the recent history of humans. Pathogen-driven selection acts when one or more genetic variants influence the susceptibility to be infected or the severity of the resulting disease: it shapes the variability of human genes leaving signatures that can be exploited to identify that same variants. Because of the strong selective pressure imposed by protozoa, alleles that protect against these agents are expected to be at higher frequencies in heavily affected populations. Therefore, one possibility to identify susceptibility alleles for protozoa-borne diseases is to search for correlations between genetic variability and an estimate of the selective pressure exerted by the infectious agents in different human populations. Despite the relevance of protozoan-borne diseases, few susceptibility loci have been identified. In this work we intend to apply the above mentioned approach to increase their number.

Methods

To this aim we analysed genotype data of 660,832 single nucleotide polymorphisms (SNPs) genotyped in 52 human populations distributed worldwide (from the HGDP–CEPH panel). We used three different measures of selective pressure: two independent malaria prevalence estimates from the GIDEON and the WHO databases; one estimate of protozoa diversity from the GIDEON database. In order to account for environmental variables possibly correlating with protozoa diversity, latitude, temperature, short wave radiation flux, precipitation rate and relative humidity were retrieved from the NCEP/NCAR database. All measure estimates where obtained for the 21 countries where the 52 HGDP–CEPH

[1] Scientific Institute IRCCS E. Medea, Bioinformatic Lab, Via don L. Monza 20, 23842 Bosisio Parini (LC), Italy [2] Bioengineering Department, Politecnico di Milano, P.zza L. da Vinci, 32, 20133 Milan, Italy. [3] Dino Ferrari Centre, Department of Neurological Sciences, University of Milan, IRCCS Ospedale Maggiore Policlinico, Mangiagalli and Regina Elena Foundation, Via F. Sforza 35, 20100 Milan, Italy. [4] Department of Biomedical sciences and Technologies LITA Segrate, University of Milan, Via F.lli [5] Don C. Gnocchi ONLUS Foundation IRCCS, Via Capecelatro 66, 20148 Milan, Italy

populations are located. Correlations between SNPs frequencies and environmental/prevalence/diversity variables were estimated using Kendall-tau. We also applied Partial Mantel Tests to account for correlation between genomic and geographic distances and then to measure the effect of a third variable. A SNP was considered associated with a variable if it displayed a significant Kendall correlation after Bonferroni correction (at different significance levels, see results) and a Mantel's r higher than the 95th percentile of the distribution of all SNPs in the panel. A SNP was ascribed to a specific gene if it was located within the transcribed region or no farther than 500 bp upstream the transcription start site. In order to estimate the probability of obtaining n genes carrying at least one significantly associated SNP out of a group of m genes, we applied a re-sampling approach: samples of m genes were randomly extracted from a list of all genes covered by at least one SNP in the HGDP-CEPH panel (15,280) and for each sample the number of genes with at least one significant SNP were counted. The empirical probability of obtaining n genes was then calculated from the distribution of counts deriving from 10,000 random samples.

Results

Protozoa diversity resulted to be the most effective measure of malaria-driven selective pressure being able to identify 11 out of 31 malaria associated genes (Kendall, $p<0.01$). No variant were identified using the WHO and GIDEON prevalence data. We also expected SNPs in genes involved in the immune response to be more frequently associated with protozoa diversity than observed for randomly sampled loci. Results showed that among 2287 genes, 300 contained at least one associated SNP, corresponding to an empirical probability of 0.0001 and confirming our prediction. These findings indicate that protozoa diversity is a reliable estimator of the selective pressure imposed by protozoa and warrant its use for a genome-wide search of significantly associated SNP. The same procedure sketched above was therefore applied to all SNPs typed in the HGDP-CEPH panel leading to the identification of 5180 variants ($p<0.05$) associated with protozoa diversity and mapping to 1145 distinct genes. Compared to to randomly selected variants these display higher population genetic differentiation (Fst) and are more frequently mapping to genes. With a single exception no correlation was found between the identified variants and other environmental variables. In addition to providing insight into the evolutionary history of our specie, approaches as the one we have used here might complement and integrate GWA studies in identifying the genetic basis of resistance/susceptibility to disease.

Contact e-mail

uberto.pozzoli@bp.lnf.it