Inference of host-pathogen interaction matrices from genome-wide polymorphism data

Hanna Märkle[1,b,1], Sona John[1,1], Lukas Metzger[1,1], STOP-HCV Consortium[c], M Azim Ansari[c], Vincent Pedergnana[d], Aurélien Tellier[1,1]

[a]*Population Genetics, Department of Life Science Systems, School of Life Sciences, Technical University of Munich, 85354 Freising, Germany*
[b]*Center for Genomics & Systems Biology, New York University, New York, NY 10003, USA*
[c]*Nuffield Department of Medicine, Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK*
[d]*Laboratoire MIVEGEC (UMR CNRS 5290, UR IRD 224, UM), Montpellier, France*

Abstract

Coevolution is defined as the evolutionary change in one antagonist (host) in response to changes in the other antagonist (pathogen). At the genetic level, these changes are determined by genotype x genotype (GxG) interactions. We build on a general theoretical model of a host-pathogen interaction to derive four indices to retrieve key features of GxG interactions. The four developed indices extract relevant information from polymorphism data of randomly sampled uninfected hosts as well as infected hosts and their respective pathogen strains. Using these indices as summary statistics in an Approximate Bayesian Computation method, we can show their power to discriminate between GxG interaction matrices. Second, we apply our ABC method to a SNP data set of 451 European humans and their infecting Hepatitis C Virus (HCV) strains supplemented by polymorphism data of 503 individuals from the 1,000 genomes project. As our indices encompass and extend previous natural co-GWAs we recover many of the associations previously reported for this dataset and infer their underlying interaction matrix. We reveal a new candidate gene for resistance to HCV in the human genome, and two groups of significant GxG associations exhibiting gene-for-gene interactions. We suggest that the inferred types of GxG interactions result from the recent expansion, adaptation and low prevalence of the HCV virus population in Europe.

Significance statement

Why are some host individuals susceptible/resistant to infection by certain pathogen genotypes and others not? Understanding the genetic characteristics of genes driving host-pathogen interactions is crucial to predict epidemics. We develop four indices based on a mathematical model and build a Bayesian statistical method computing these indices on full genome data of infected hosts and their infecting pathogen strains and data of non-infected hosts. We can pinpoint the genes underlying host-pathogen interactions and infer their characteristics. Applying our framework to data from European humans and the Hepatitis C virus, we discover a new potential resistance gene in humans and reveal how the virus has adapted in the last 150 years to match the genetic diversity of the European human population.

Keywords

population genomics; linkage disequilibrium; single nucleotide polymorphism; host-pathogen co-evolution; GxG interactions.

Introduction

Host-pathogen or host-parasite antagonistic interactions are pervasive in nature. Their relevance ranges from specific simple interactions underpinning devastating epidemics [4, 31, 56] up to the multi-trophic interactions defining ecosystems and microbiomes [50]. Coevolution is defined as the evolutionary change in one antagonist (host) in response to changes in the other antagonist (pathogen). At the genetic level, these changes are determined by genotype x genotype (GxG) interactions between few (up to many) host and pathogen genes. For example, host genotypes differ specifically in their resistance to pathogen strains which in turn differ in their infectivity (ability to infect and cause disease) on the given host genotypes. Host-pathogen GxG interactions are defined by their (i) genetic architecture (how many genes are involved?), (ii) specificity (how many GxG interactions can yield a resistance phenotypic outcome?) and (iii) strength (what is the phenotypic outcome, full resistance up to severe infection?). Knowing the genetic architecture, specificity and strength of GxG interactions is crucial for understanding and predicting the speed and outcome of coevolutionary dynamics [15, 28, 54] and for disease management in agriculture and medicine.

The potentially devastating effects of infection prompted a wealth of genome-wide association (GWAs) studies to identify the genetic architecture (involved genes) of GxG interactions. Single species GWAs are performed by associating genomic variants with a binary disease outcome: (i) infected versus non-infected hosts such as humans [8, 16], invertebrates [11, 12], and plants [22, 44], or (ii) infective/non-infective pathogens [3, 46]. With the growing joint availability of host and pathogen genomic data, two types of joint Genome-Wide Association studies (so-called co-GWAs) have been developed to identify significant GxG loci [9, 10, 40, 58]. Experimental co-GWAs requires a full experimental factorial design of reciprocal infections to assess the outcome of infection (phenotype) [40, 58]. However, controlling for the genetic background and running controlled infection experiments is elusive to human hosts and often difficult to achieve for non-model natural host-pathogen interactions. As an alternative, natural co-GWAs [10, 40] jointly associate genome wide polymorphism data of infected hosts with polymorphism data of their respective infecting pathogen strains [9]. Such natural co-GWAs have since been applied successfully to find associations between human genes and pathogen loci of the Hepatitis C virus (HCV) [5], *Streptococcus pneumoniae* [36] and *Plasmodium falciparum* [7] and to study interactions between *Daphnia magna* host and *Pasteuria ramosa* [23].

Yet, deciphering the specificity and strength of the GxG interactions at the loci of interest has remained empirically out of reach for most host-pathogen systems. Specificity and strength of host-pathogen GxG interactions are classically summarized within the so-called infection matrix, which captures the extent to which each pathogen genotype successfully infects each host genotype (0 meaning full host resistance, and 1 meaning full host susceptibility, Fig. 1a,b). There is a wide range of possible infection matrices which differ in their levels of symmetry, specificity and strength [1, 15, 28]. Further, previous studies suggest that the number of loci involved varies between different host-pathogen systems and there are often epistatic effects between loci [23]. Throughout the article we will focus on four matrices of interest (Fig. 1b): 1) the generalist pathogen (P) infectivity/non-infectivity matrix in which one pathogen genotype has a high infectivity on all host genotypes, 2) the generalist host (H) resistance/susceptibility matrix where one host genotype is resistant against all pathogen genotypes, 3) the specific matching-alleles (MA) matrix where each pathogen genotype is specialized to infect one host genotype as found in the *Daphnia magna* - *Pasteuria* pathosystem [38], and 4) the specific gene-for-gene (GFG) matrix characterized by an universally infective pathogen genotype and resistance being the result of recognition of a specific pathogen effector allele [55]. GFG interactions are mainly documented for plant-pathogen interactions [25, 55], while MA interactions have been long hypothesized to underlie the interactions between the human major histocompatibility complex (MHC) and most mammalian immunity genes and corresponding pathogen genes [25, 34, 49]. Deciphering the infection matrix via experimental means requires the combinatorial infection assays of many host and pathogen genotypes (clones or isogenic lines with known allelic variants) in controlled conditions, and thus, is prohibitive for most host-pathogen systems (but see [38, 43]).

We develop here a new theoretical and statistical framework using jointly genomic data of hosts and their pathogens from natural populations to find out the genes underpinning GxG interactions along with inferring the specificity and strength of the interactions. This framework explicitly takes three fundamental sampling processes occurring in any host and pathogen populations into account (Fig. 1c): (i) the (co)evolutionary sampling [25, 39], (ii) the epidemiological sampling, and (iii) the experimental sampling. The first process is a result of coevolution itself, namely that host and pathogen genotype frequencies fluctuate in space and time as a direct result of reciprocal selection (coevolution), genetic drift, mutations and gene flow [28, 54]. As a result, only a subset of all possible interactions between host and pathogen genotypes maybe present at a given point in space and time (Fig. 1c) [25, 54], so that the sampling of host and pathogen genotypes may be incomplete. This effect is a major hindrance for host (or parasite) single species GWAs as it decreases the statistical power when not accounting for the genetic heterogeneity of populations [39]. Second, host genotypes need to encounter corresponding pathogen genotypes in order to get infected as a result of a specific GxG interaction. The frequency of such encounters in natural populations is governed by the allele frequencies, the disease prevalence, the population size and the specific dynamics of host-pathogen encounters. Frankly speaking, an observer cannot know if an uninfected host in a natural population has been exposed to pathogens but is resistant, or if the host has never been in contact with pathogens (Fig. 1c, d). Third, sampling a limited number (subset) of host (infected and non-infected) and pathogen individuals (genotypes) from the entire population for experimental and genomic studies may further blur the true infection matrix (Fig. 1d), because the sampling scheme may over- (or under-) represent some interactions. Altogether, these three sampling processes render the inference of the underlying infection matrix a non-trivial task (Fig. 1). We argue that, so far, the consequences of these stochastic processes on the statistical power of single species GWAs and co-GWAs are poorly understood. Specifically, we expect these three processes to generate variability in the samples' allele frequencies and in the statistical power of GWAs and co-GWAs studies to detect GxG interactions, depending on the true (yet unknown) infection matrix, epidemiological dynamics, and sampling scheme (infected and/or non-infected hosts) in addition to the previously reported effects of the coevolutionary dynamics and incomplete sampling of hosts and pathogen genotypes [39].

In this study we first derive four different indices based on host-pathogen coevolutionary theory to tackle the problem of inferring the significant GxG interactions and assess their infection matrices under the described stochastic processes. We then incorporate these indices as summary statistics (model based) into an Approximate Bayesian Computation (ABC) framework to analyse jointly genomes of non-infected hosts as well as infected hosts and their matching pathogens. We aim to (i) pinpoint genes underlying GxG interactions, and (ii) infer the underlying infection matrix. As a proof of principle, we infer the interaction matrices underpinning 535 biologically relevant GxG associations between human Single Nucleotide Polymorphism (SNPs) and Singly Amino Acid Polymorphisms (SAAPs) of HCV, using 451 infected individuals and the HCV sequences of the infecting strains [5] complemented by 503 human genomes [2].

Results

Indices capture features of the infection matrices

We develop a theoretical model of a temporal snapshot (single-time point) of the outcome of an epidemic process in a host population of large size. In short the model assumes that hosts encounter pathogens at random at a given disease encounter rate ϕ (Table S1, Supplementary Text S1). After the infection process the host population can be generally split into two compartments, namely infected hosts (frequency $\tilde{f}$) and uninfected hosts (frequency $1 - \tilde{f}$). The latter is composed of hosts which either did not encounter pathogens or resisted infection. We denote the frequency of

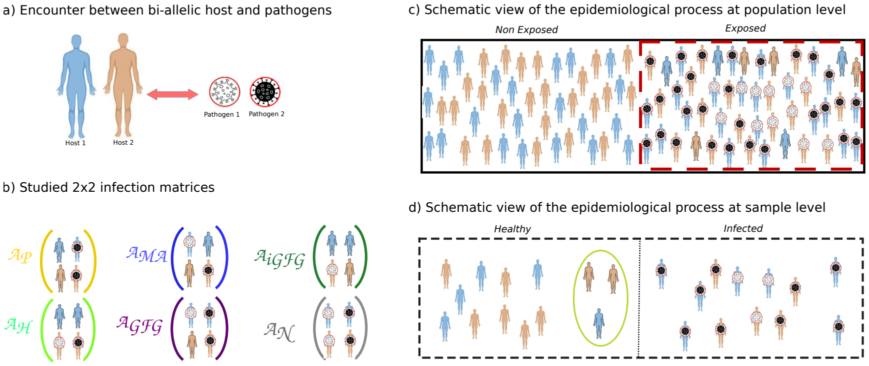
a) Encounter between bi-allelic host and pathogens
Host 1
Host 2
Pathogen 1
Pathogen 2
b) Studied 2x2 infection matrices
$\mathcal{A}_{\mathcal{P}}$
$\mathcal{A}_{\mathcal{MA}}$
$\mathcal{A}_{i\mathcal{GFG}}$
$\mathcal{A}_{\mathcal{H}}$
$\mathcal{A}_{\mathcal{GFG}}$
$\mathcal{A}_{\mathcal{N}}$
c) Schematic view of the epidemiological process at population level
Non Exposed
Exposed
d) Schematic view of the epidemiological process at sample level
Healthy
Infected

Figure 1 Schematic view of the principles of GxG interactions underlying host-pathogen coevolution and the characteristics of the sampling process. Our model captures the effects of experimental and epidemiological sampling at a single time point of the coevolutionary dynamics. (a) We assume an interaction between a bi-allelic host and a bi-allelic pathogen locus. (b) The outcome of the interaction is summarized by a 2x2 infection matrix with host genotypes as rows and pathogen genotypes as columns. Some classic examples with extreme values are schematically depicted, namely the pathogen infectivity/non-infectivity ($\mathcal{A}_{\mathcal{P}}$, yellow), matching-allele ($\mathcal{A}_{\mathcal{MA}}$, blue), inverse gene-for-gene ($\mathcal{A}_{\mathcal{GFG}}$, dark green), host resistance/susceptibility ($\mathcal{A}_{\mathcal{H}}$, light green), gene-for-gene ($\mathcal{A}_{\mathcal{GFG}}$, purple), and neutral ($\mathcal{A}_{\mathcal{N}}$, grey) matrix. (c) Schematic representation of the infection process and the host's status at the population level. In a homogeneous population, a proportion ϕ (the disease encounter rate) of hosts encounters pathogen infectious propagules at random, and a proportion $1-\phi$ does not (epidemiological sampling). Hosts with a solid outline (and circled in green) in the exposed class are resistant to the infection (appear as healthy) due to specific form of the underlying infection matrix. (d) Genomic studies are performed by taking a sample of healthy and infected hosts from the total population, generating a potential bias in sample allele frequencies compared to population alelle frequencies (experimental sampling). On a population level, the frequencies of the host and pathogen alleles are determined by the coevolutionary process (coevolutionary sampling).

uninfected hosts of type i in the entire population as f_{iz}. Assuming bi-allelic host and pathogen genotypes, there is a maximum of four possible host-pathogen associations in the infected compartment. We denote the frequency of hosts with genotype i infected by pathogens genotype j in the entire population as f_{ij} and in the infected subpopulation as $\tilde{f}_{ij}$. These frequencies depend on the frequency of hosts of type i (h_i), the initial frequencies of pathogen genotype j prior to infection, the infection matrix (α) and the pathogen encounter rate ϕ.

We develop four indices based on co-evolutionary theory to capture the characteristics of a given GxG interaction matrix α (Fig. 1b). These indices combine information of host allele frequencies from infected hosts and their associated pathogen strains (pathogen allele frequencies) as in co-GWAs [5, 9, 10] as well as additional information of allele frequencies in a sample of non-infected hosts as in host GWAs [8, 44]. Our first index, the cross-species association (CSA) index, is a cross-species analogue of linkage disequilibrium [27, 40] (also termed interlinkage [23]), which assesses the association between the genotype of infected hosts and the genotype of the respective infecting pathogen strains (thus similar to the natural co-GWAs). The host susceptibility (HS) index compares allele frequencies in the infected versus non-infected host subsamples (thus similar to host GWAs). The pathogen infectivity (PI) assesses the difference between pathogen allele frequencies (thus similar to pathogen GWAs). Finally, the host partitioning (HP) index is designed to reflect the difference of allele frequencies of one host genotype infected by one pathogen allele and when non-infected. The HP index thus contains novel information (compared to co-GWAs and GWAs) on the asymmetry, specificity and strength of the infection matrix.

The four indices are defined as:

$$\begin{aligned}
\text{CSA} &= \left|\frac{\tilde{f}_{11}\tilde{f}_{22} - \tilde{f}_{12}\tilde{f}_{21}}{\bar{f}_1}\right| \\
\text{HS} &= \left|\frac{(f_{11}+f_{12})f_{2z} - (f_{21}+f_{22})f_{1z}}{\bar{f}_2}\right|, \\
\text{PI} &= \left|\frac{f_{12}f_{22} - f_{11}f_{21}}{\bar{f}_2}\right|, \\
\text{HP} &= \left|\frac{f_{12}f_{2z} - f_{21}f_{1z}}{\bar{f}_2}\right|,
\end{aligned} \tag{1}$$

with:

$$\begin{aligned}
\bar{f}_1 &= \sqrt{(\tilde{f}_{11}+\tilde{f}_{12})(\tilde{f}_{21}+\tilde{f}_{22})(\tilde{f}_{11}+\tilde{f}_{21})(\tilde{f}_{12}+\tilde{f}_{22})}. \\
\bar{f}_2 &= (f_{11}+f_{12}+f_{1z})(f_{21}+f_{22}+f_{2z}).
\end{aligned} \tag{2}$$

Expressing these indices in terms of the population composition (Table S1) and the coefficients α_{ij} of an arbitrary 2x2 infection matrix we find (Supplementary Text S1):

$$\begin{aligned}
\text{CSA}^2 &= \left|\frac{h_1h_2p_1p_2\left(\alpha_{11}\alpha_{22}-\alpha_{12}\alpha_{21}\right)^2}{\left(\alpha_{11}p_1+\alpha_{12}p_2\right)\left(\alpha_{21}p_1+\alpha_{22}p_2\right)\left(\alpha_{11}h_1+\alpha_{21}h_2\right)\left(\alpha_{12}h_1+\alpha_{22}h_2\right)}\right|, \\
\text{HS} &= \left|\phi\Big[\left(\alpha_{11}-\alpha_{21}\right)p_1+\left(\alpha_{12}-\alpha_{22}\right)p_2\Big]\right| \\
\text{PI} &= \left|\phi^2\left(p_2^2\alpha_{12}\alpha_{22}-p_1^2\alpha_{11}\alpha_{21}\right)\right|, \\
\text{HP} &= \left|\phi\big(\alpha_{12}p_2(1-\phi\alpha_{22}p_2)-\alpha_{21}p_1(1-\phi\alpha_{11}p_1)\big)\right|
\end{aligned} \tag{3}$$

We first derive the population level values of these indices (Table 1, Table S1) for different infection matrices and host and pathogen allele frequencies. This allows us to assess their suitability to distinguish between different matrices. Note that our neutral matrix (all matrix elements 1, Fig. 1b) builds on the hypothesis that the GxG interaction for a given pair of host and pathogen alleles is not relevant for the infection status.

Studying the most extreme forms (all elements either 0 or 1) of these infection matrices we find that the combination of our indices shows differential behavior among infection matrices. For example the CSA index provides a clear distinction between the GFG and MA matrix from all other matrices. Our results (Table 1) further highlight dependencies of the index values on the disease encounter rate (ϕ) and/or non-linear relationships with pathogen allele frequencies prior to host exposure to pathogens (Eq. 3). When we derive expressions of the index values for more general forms of the corresponding GxG matrices (Table S1) the expressions become more cumbersome (Table S2). Yet, we still find that the combination of all four index values shows differential behaviour across the different infection matrices. Therefore, it appears that our four indices, and combinations thereof, can be suitable to discriminate between different types of infections matrices. Extending these theoretical results, it is in principle possible to directly compute the values of the coefficients of the infection matrix (α_{ij}) by simultaneously solving the set of all equations Eqs. 3. However, this approach shows only reasonable results when the disease encounter rate is known and approximately 50% and when population-level allele frequencies are known (which is in practice not the case because of the effect of the experimental sampling, Fig. 1d, Supplementary Text S1).

Table 1 Values of indices for different GxG matrices assuming host genotypes being fully susceptible to infection by pathogen genotype j when $\alpha_{ij} = 1$ or fully resistant when $\alpha_{ij} = 0$.

	HS	PI	CSA^2	HP
$\mathcal{A}_{\mathcal{N}} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	0	$\lvert\phi^2 (p_2 - p_1)\rvert$	0	$\lvert\phi (1 - \phi) (p_2 - p_1)\rvert$
$\mathcal{A}_{\mathcal{GFG}} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$	$\lvert\phi p_1\rvert$	$\lvert\phi^2 p_2^2\rvert$	$\lvert p_1 h_2\rvert$	$\lvert\phi p_2 (1 - \phi p_2)\rvert$
$\mathcal{A}_{\mathcal{MA}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\lvert\phi (2p_1 - 1)\rvert$	0	1	0
$\mathcal{A}_{\mathcal{H}} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}$	$\lvert-\phi\rvert$	0	0	$\lvert-\phi p_1\rvert$
$\mathcal{A}_{\mathcal{P}} = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$	0	$\lvert\phi^2 p_2^2\rvert$	0	$\lvert\phi(1 - \phi) + \phi p_1 (2\phi - 1 - \phi p_1)\rvert$

Indices' behaviour is robust to sampling procedures

We next quantify the behaviour of these four indices to discriminate between different infection matrices under evolutionary, experimental and epidemiological sampling procedures (Fig. 1b, c) (for more details see methods and SI Text). We explore the indices' distributions over a wide range of minor allele frequencies (h_i, p_j between 0.05 and 0.5) and allowing for random deviations of the matrix coefficients within a tolerance δ. Under evolutionary sampling the ranges of our HP, HS, and PI indices for the entire population are very small for small disease encounter rates, but still distinguishable between different matrices (Fig. 2, top row). The distributions of the indices' values become more wide spread when taking a sample from the entire population with more or less equal amounts of non-infected and infected individuals (Fig. 2, bottom row). Encouragingly, both for the population and experimental samples, the distribution of indices' values differs between the different matrices for various disease encounter rates (Fig. 2). More importantly, there is at least one combination of two or more indices (albeit not necessarily linear) for each GxG infection matrix discriminating it from the neutral infection matrix (Fig. 2).

We observe a strong dependency between the range of possible indices' values and the disease encounter rate ϕ (compare Fig. 2, S1 and S2). Consistent with our theoretical results, the ranges of values for HS, PI and HP are small for low disease encounter rates and increase with higher disease encounter rates (Fig. S1, S2). Yet, for all three disease encounter rates the different matrices are distinguished by the combination of indices under the population sample. As for $\phi = 0.05$ we observe that taking a fixed sample from the entire population changes the range of observed index values in the sample compared to the population. This effect depends on the specific combination of (host and pathogen) sample sizes and disease encounter rate (Fig. 2, S1-S4). When we consider a sampling scheme with 5% infected and 95% non-infected hosts (keeping the total number of samples to 951 hosts) for a disease encounter rate $\phi = 0.05$, the distribution of indices' values becomes more narrow and more similar to that of the population sample. This potentially decreases the extent to which different matrices can be discriminated (Fig. S4). On the other hand if we consider a sampling scheme with 95% infected and 5% non-infected hosts (keeping the total number of samples to 951 hosts) the range of indices' values further broadens and becomes less similar to the population sample. The difference in sample indices' distributions reflects the effect of experimental sampling of infected hosts on top of the epidemiological sampling for a low disease encounter rate. Our results exemplify the, so far, largely ignored effects of the epidemiological and experimental sampling in natural co-GWAs and the importance of deriving of optimal sampling schemes to overcome this interplay.

Increasing the tolerance threshold (value of δ) increases the amount of overlap between the indices' distributions. As a consequence different matrices may be confounded (Fig. S5 for δ varying between 0.1 and 0.3). In other words, choosing a low tolerance parameter generates a more stringent statistical test to disentangle between the neutral infection matrix and other matrices and should decrease the rate of false positives (association and underlying matrices appearing to be biologically relevant whereas these are in fact neutral).

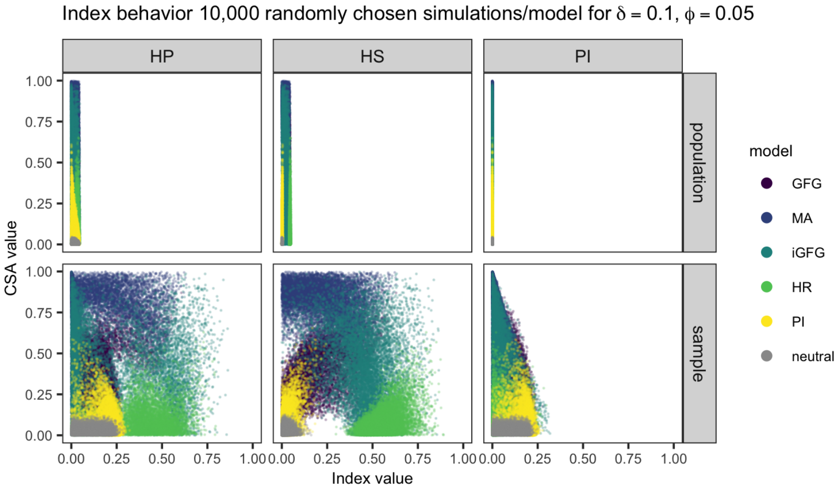

Figure 2 Distribution of values of indices' pairs comprising CSA (y-axis) and one of the other indices HS, PI or HP (x-axis) for different infection matrices ($\mathcal{A}_{\mathcal{GFG}}$, $\mathcal{A}_{\mathcal{MA}}$, $\mathcal{A}_{\rangle\mathcal{GFG}}$, $\mathcal{A}_{\mathcal{H}}$, $\mathcal{A}_{\mathcal{P}}$, $\mathcal{A}_{\mathcal{N}}$) for a low disease encounter rate ($\phi = 0.05$). The population has size $N = 100,000$ (population row) and a random sample of $n_H = 1006$ healthy and $n_I = 902$ infected haploid individuals is taken (sample row). Results are shown for 10,000 simulations where $h_1 \sim \mathcal{U}(0.05, 0.5)$, $p_1 \sim \mathcal{U}(0.05, 0.5)$ and $\delta = 0.1$. The simulations are a randomly selected subset of the 50,000 simulations used in the ABC model choice.

An ABC framework allows to infer the infection matrices

We then use these simulation results ($\phi = 0.05$ and $\delta = 0.1$) in an Approximate Bayesian Computation (ABC) framework to infer the infection (GxG) matrix (neutral, MA, GFG,...) for a given association. We use our four indices as ABC summary statistics. A GxG association is not biologically relevant (significant) for a host-pathogen interaction if the ABC model choice procedure reveals the neutral matrix as the best (or equally best) model. We assess the statistical power of ABC model (matrix) choice by running a leave-one-out cross-validation (rejection algorithm, tolerance=5%) based on randomly chosen 500 simulations per infection matrix and inferring the best model using simulations for all matrices (50,000 per matrix). We demonstrate that our ABC based on our four indices can discriminate between all matrices (Table 2, S3, S4), especially between biologically relevant GxG matrices and the neutral matrix (under the most stringent threshold, Table S3, S4). The host resistance and the MA matrix can be well discriminated from all other matrices, whereas the pathogen infectivity and iGFG matrices may still be confounded with other matrices (Table 2). Therefore, accounting for sampling effects (as introduced in priors for various parameters and fixed sample sizes) within our ABC framework allows to disentangle between the different infection matrix models (Table 2, S3, S4).

Table 2 Results of a leave-one-out ABC (rejection) cross-validation for 500 randomly chosen simulations per infection matrix under low disease encounter rate. For each model 50,000 simulations are produced for $h_1 \sim \mathcal{U}(0.05, 0.5)$, $p_1 \sim \mathcal{U}(0.05, 0.5)$, $\delta = 0.1$, $\phi = 0.05$, $N = 100,000$, $n_I = 902$ haploid and $n_H = 1006$ haploid.

	Inferred model					
True model	neutral	GFG	MA	iGFG	HR	PI
neutral	458	0	0	0	0	42
GFG	16	361	24	46	6	47
MA	0	7	471	22	0	0
iGFG	2	52	63	244	138	1
HR	0	0	0	7	492	1
PI	114	39	0	0	0	347

535 biologically relevant GxG associations between humans and HCV

We now apply our ABC framework combining a dataset of human diploid host sequences and their infecting HCV (hepatitis C virus) strains [5] (the infected sample) and additional sequences from the 1,000 genomes project [2, 52] (the non-infected sample). In order to limit confounding effects of population structure in the data, we restrict our analysis to the subset of 451 individuals of European ancestry (PCA in Fig. S6, [5]). As previously described [5], we convert the viral nucleotide sequence data into Single Amino Acid Polymorphisms (bi-allelic SAAPs) data. Our non-infected sample consists of 503 diploid individuals of European ancestry from the 1,000 genomes project (filtering SNPs corresponding to those from [5], PCA in Fig. S6). We filter for a minor allele frequency (MAF) MAF > 0.2 to maximize the power to disentangle between infection matrices. As highlighted above, below this frequency, several stochastic sampling effects decrease significantly the power to pinpoint relevant GxG associations. We compute our four indices for all possible pairwise associations between 326,520 human SNPs and 208 SAAPs. For the 800 top associations defined as exhibiting the highest values of our indices, we run the ABC model choice between the possible six infection matrices (neutral, GFG, iGFG, MA, H, P). Our model choice results in 535 interactions which differ from the neutral matrix based on a Bayes factor threshold of two (BF > 2) and for a matrix tolerance threshold $\delta = 0.1$ (Fig. 3). For each of the 535 associations, we infer the most probable infection matrix (Table S5-S8).

We summarize the estimated infection matrices and their distributions by index (Fig. 3, S7, S8). We find two main groups of associations with an estimated gene-for-gene matrix ($\mathcal{A}_{\mathcal{GFG}}$): one group includes several SNPs on the human chromosome 6 falling into the MHC region and the HCV gene NS3, and the second group has only one association under GFG between one SNP at the clathrin heavy chain linker domain containing 1 (CLHC1) gene on chromosome 2 and an SAAP on the HCV gene E2. Furthermore, we find several associations with an estimated resistance matrix ($\mathcal{A}_{\mathcal{H}}$) between one position at the Lymphocyte-specific protein 1 (LSP1) on chromosome 11 and SAAPs at various viral genes. Finally, we find also several pathogen infectivity matrices ($\mathcal{A}_{\mathcal{P}}$) between 21 SNPs in the human genome and 45 SAAPs in the HCV genome. We also highlight, that we do not find any associations which are indicative of a matching-allele (MA) infection matrix, even when lowering the detection threshold (higher δ) and considering competing best models (Tables S5-S8). Analyzing the details of these 535 biologically relevant GxG associations, we find few host sites (106) especially exhibiting GFG or resistance matrices, while pathogen AA (221) exhibit chiefly infectivity matrices. We also compare our results to a co-GWAs on the subset of 451 European human infected individuals (and their 451 pathogen strains) following the previous analysis [5] using plink. Our CSA index shares all associations with our Bonferroni corrected co-GWAs (Fig. S9). In addition, 68 candidates from our European CSA index also appear in the 104 top candidates from the Bonferroni corrected results obtained previously for the full data set (Fig. S10, [5]). We conclude that our ABC framework based on our CSA index reveals relevant GxG associations but is more stringent than co-GWAs studies. Using four indices which capture different aspects of infection matrices, we are able to reveal new associations which were not previously reported, especially potential human resistance alleles to HCV (under GFG and resistance matrices) and HCV infectivity alleles (under infectivity matrix).

Discussion

We derived four indices to tackle the problem of inferring the underlying infection matrix from host-pathogen association data. Further, we developed some general predictions on the behavior of these indices, establish their joint ability to discriminate between several infection matrices and use them successfully as summary statistics in an ABC framework to reveal infection matrices in a human host/HCV virus data set. Therefore, our study is not only the first study attempting to establish the theoretical aspects of natural co-GWAS and the effect of various type of sampling, but also lays ground for a framework to infer the infection matrix using natural co-GWAs setups. We specifically present results tuned for studying the interaction between European humans and HCV, namely we assume a sample size of 451 infected diploid humans (and their respective viral strains) and 503 non-infected European humans, and a known disease encounter rate of 0.05 (slightly higher than the disease prevalence of approximately 3% previously reported [42, 45]).

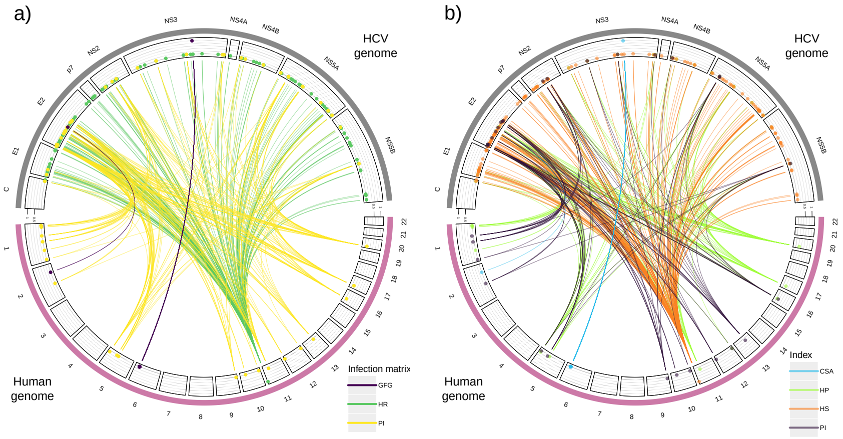

Figure 3 Genome-to-genome relevant associations from the ABC-model choice for all 535 associations ($BF > 2$ to the neutral matrix). a) The 535 associations between host SNPs and pathogen SAAPs colored by the single best infection matrix. b) The 535 associations colored by the most informative index. The human chromosomes are shown on the bottom, the virus contigs on the top. The second circle of lines indicate the number of sites sharing one association (most inner line indicates a single site up to most outer line indicating multiple closely linked sites). The color coding is as follows for the infection matrices: purple=$\mathcal{A}_{\mathcal{GFG}}$, darkgreen=$\mathcal{A}_{\mathcal{H}}$, yellow=$\mathcal{A}_{\mathcal{P}}$. For the indices blue is for CSA-index, lightgreen for HP-index, orange for HS-index, and purple for PI-index.

Our Bayesian framework identifies genes that exhibit statistically relevant associations which have some support from previous studies. All observed associations at the major histocompatibility complex (MHC) on chromosome 6 are associated with one viral site located at position 1,444 on the non-structural protein 3 (NS3) gene. All these human sites are likely linked to alleles at the HLA genes due to the high amount of linkage disequilibrium across the region [6]. The HLA genes of a patient determine which viral peptides are presented to T cells as part of the adaptive immune response. This process can drive viral evolution and result in the emergence of viral escape mutations, which have been previously identified for the NS3 gene [5, 41]. There is some empirical evidence that small interfering RNA-mediated clathrin heavy chain depletion affects endocytosis of HCV [14, 20]. Therefore, we speculate that the CLHC1 (Clathrin Heavy Chain Linker Domain Containing 1) gene may be involved in such process, possibly supporting our finding of a GFG interaction of this gene with an amino acid in the HCV gene E2. We further found a putative resistance allele at the Lymphocyte-specific protein 1 (LSP1) gene which is an F-actin binding protein. This protein is involved in the regulation of various immune system functions, including lymphocyte activation, proliferation, and migration [47]. It also has been shown to play a role in endocytosis and transendothelial migration of leukocytes, allowing these to be recruited to the sites of inflammation [37, 57]. Studies demonstrating that the depletion of LSP1 significantly reduces the rate of endocytosis of HIV particles [18, 19], could suggest that this protein may also play a role in the endocytosis of HCV.

Despite likely coevolving with humans over thousands of years in Africa, HCV has a very recent history of infection and spreading in the European human population (Fig. S11, [24, 26]). Therefore, the biologically most relevant SNP-SAAP associations should be interpreted in the light of the HCV virus adapting to existing standing genetic variation in the European population within the (approximately) last 150 years. Experimental results from a bacteria-phage coevolution interaction [33], indicate that initial coevolutionary dynamics are characterized by rapid fixation of advantageous alleles in hosts and pathogens (arms race dynamics [13]). The dynamics are then replaced by trench warfare dynamics [51] with maintenance of two or more alleles at the coevolving genes by balancing selection. Our inferred asymmetric matrices (host resistance, pathogen infectivity and GFG) likely indicate that we capture the initial dynamics of the interaction between humans and HCV in Europe. Asymmetric matrices are more likely to generate arms race dynamics especially when population sizes are small [1, 53, 54]. In this light, we interpret the finding of a resistance matrix at the LSP1 gene as an indication that resistance to HCV may be segregating in the human population. Several mutations in the virus populations have likely been selected for overcoming this resistance allele (green lines in Fig. 3). In addition, several SAAPs with inferred infectivity matrices likely indicate that strains of HCV exhibit mutations allowing them to infect and match several host genes and alleles. In other words, there are virus strains with different infectivity ranges. Finally, the inferred GFG interactions indicate that the virus has evolved to overcome host recognition alleles at several MHC genes and at one gene on chromosome 2 (CLHC1). These human alleles likely provided initial resistance to HCV at the onset of the epidemics which has been overcome by subsequent mutations in the virus.

We speculate that further extending our inference framework to data sampled from different time points can potentially help to elucidate the speed and timing of coevolution and changes in the GxG interactions at the genetic level. One key prediction from co-evolutionary theory is that due to various stochastic and selective processes, the number of genes under coevolution and the corresponding infection matrices are subject to change over time [15, 25], exhibiting various degrees of asymmetry [1, 28]. This in turn generates different coevolutionary dynamics in time (arms race and trench warfare dynamics, respectively, [29, 54]). It is possible that in the long term run coevolution between HCV and other human populations, SNPxSAAP interactions may exhibit different underlying infection matrices promoting the occurrence of trench warfare dynamics and balancing selection, namely : (i) symmetric MA interactions, or (ii) asymmetric GFG interactions with the necessary, but not sufficient, [53, 54] condition that costs of resistance and infectivity exist at these coevolving loci. Applying our inference framework to other diseases with a range of short to long term coevolutionary histories would shed light on the speed of coevolution between humans and their viruses [30] and the underlying coevolutionary dynamics.

We specifically focus on random disease transmission and low disease encounter which likely best describe HCV transmission dynamics in Europe [42, 45]. This allows us to account for the effect of the epidemiological sampling on the distribution of allele frequencies in the population and in our experimental samples, without the need to specify a corresponding prior for the disease encounter rate in the ABC. Our use of priors for host and pathogen allele frequencies takes into account that the 'true' allele frequencies prior to the infection process (h_i and p_j in our model) are unknown. The observed allele frequencies in the sample represent indeed the outcome of the joint interaction of disease encounter rate, the 'true' but hidden allele frequencies and the infection matrix. We acknowledge that disease encounter rates might be less well known for host-pathogen interactions involving non-model species. One way to tackle this limitation for non-model plant host-pathogen interactions would be to obtain estimates for the range of disease encounter rates from field data and include this range as an additional prior into the ABC simulations. However, based on our analytical results, we expect this approach to be only successful if the corresponding estimated range of the disease encounter rate is relatively narrow.

Furthermore, we follow previous GWAs and co-GWAs approaches and assume sufficiently large sample sizes (several hundred individuals) to allow the detection of significant associations. As some of our indices rely on estimating the allele frequencies in the non-infected sub-sample and the infected sub-sample with comparatively small error, it is crucial to obtain a sample that well reflects the population frequencies of genotypes/phenotypes in the entire population. Specifically, if the disease encounter rate is small, it is important to sample well enough the infected part of the population. Conversely, if the disease encounter rate is high, sufficient sampling of the non-infected part of the population becomes important. This emphasizes the importance of taking care the interaction between sampling size and disease prevalence into account when devising sampling schemes in co-GWAS studies. Especially low sample sizes, are very likely to produce biased allele and association frequencies in the sample and hence, erroneous infection matrix estimates.

An inherent difficulty for any co-GWAs and our ABC is to confidentially detect associations which involve alleles with low frequencies. Therefore, we conservatively restricted our testing to loci with a minor allele frequency (MAF) > 0.2 to avoid an excess of false positives. However, coevolutionary dynamics can transiently decrease allele frequencies or maintain alleles at low frequencies as a result of negative indirect frequency-dependent selection [39, 53, 54]. Therefore, we speculate that our ABC method can be further improved by incorporating sample allele frequencies as additional summary statistics and by using association data from several time points. We expect the later to help with better tracking the allele frequency changes over time which directly result from the coevolutionary dynamics and the underlying infection matrix. One further current limitation of our model (and all co-GWAs) is not accounting for epistatic effects and multi-locus infection matrices. Indeed, in some species such as *Daphnia*, the resistance phenotype depends on epistatic interactions between several loci [38]. By integrating such knowledge of epistatic interactions, the results of the co-GWAs could recently be improved and additional genes of interactions discovered [23]. Integrating time-sampled data, more summary statistics and the effect of epistasis between host (or pathogen) loci into our framework constitute the topic of future work.

It is well known from the GWAs literature that spatial structure in the host and pathogen samples can affect and distort the power to detect associations. Therefore, we restricted our analysis to a single population (European) without any obvious population structure. Our results align with, and are more conservative, than the previous co-GWAs [5] applied to the same data set (Fig. S9-S10). Therefore, we are confident that our framework is stringent and exhibits a low rate of false positives. In addition, recent studies demonstrate the usefulness of using local population rather than widespread sampling in GWAS setting [32]. Accounting for spatial structure covariates and kinship matrix is another topic of future work.

In conclusion, we built here an ABC integrative method based on four indices as summary statistics which combines ideas from host or pathogen GWAs as well with host-pathogen co-GWAs and additional information from non-infected hosts. Our framework is based on a widely applicable theoretical infection model, takes into account various sampling procedures defining observed host and pathogen allele frequencies, and thus allows us to define a threshold to detect biologically relevant GxG associations in a Bayesian framework. Our model and framework should be also applicable to other GxG interactions, such as between hosts and mutualistic symbionts or between chloroplasts/mitochondria x nuclear genes interactions.

Methods

Definition of indices

The CSA index is calculated based on the frequencies of host/pathogen genotype combinations in the infected subpopulation/sample. We define the frequency of host genotype i infected by pathogen genotype j among all infected individuals as $\tilde{f}_{ij}$ $(i, j \in [1, 2])$.

$$\text{CSA} = \left| \frac{\tilde{f}_{11}\tilde{f}_{22} - \tilde{f}_{12}\tilde{f}_{21}}{\bar{f}_1} \right|, \tag{4}$$

By analogy with the linkage disequilibrium measure in population genetics, we normalize the index by the square root of the product of all infected host and pathogen allele frequencies.

$$\bar{f}_1 = \sqrt{(\tilde{f}_{11} + \tilde{f}_{12})(\tilde{f}_{21} + \tilde{f}_{22})(\tilde{f}_{11} + \tilde{f}_{21})(\tilde{f}_{12} + \tilde{f}_{22})}. \tag{5}$$

We define the genotype frequencies of uninfected hosts of type i in the population/sample as f_{iz}. Individuals can be uninfected due to two reasons: (i) they have not been exposed to the pathogen f_{i0}, or (ii) they had a pathogen encounter but resisted infection f_{i3}. We lump these two frequencies into a single frequency f_{iz} as in a natural population it is usually impossible to tell apart the difference.

The HS, PI and HP indices are defined as follows:

$$\text{HS} = \left| \frac{(f_{11} + f_{12})f_{2z} - (f_{21} + f_{22})f_{1z}}{\bar{f}_2} \right|, \tag{6}$$

$$\text{PI} = \left| \frac{f_{12}f_{22} - f_{11}f_{21}}{\bar{f}_2} \right|, \tag{7}$$

$$\text{HP} = \left| \frac{f_{12}f_{2z} - f_{21}f_{1z}}{\bar{f}_2} \right|, \tag{8}$$

with

$$\bar{f}_2 = (f_{11} + f_{12} + f_{1z})(f_{21} + f_{22} + f_{2z}). \tag{9}$$

We derived expressions for these indices for a single point in time given initial host genotype frequencies h_i, pathogen genotype frequencies p_j, a disease encounter rate ϕ and a given infection matrix α (see Supplementary Text S1).

Stochastic simulations

Next we assessed the effect of three types of stochastic processes on the behavior of the indices for all matrices in Table 1: (i) deviations of the matrix elements from the extreme values 0 and 1 and varying host and pathogen alleles frequencies, (ii) random sampling of a fixed number n_H healthy and n_I infected individuals from a population of size N, and (iii) a small ($\phi = 0.05$), intermediate ($\phi = 0.5$) or large disease encounter rate ($\phi = 0.95$). Therefore, we developed a simple host-pathogen interaction model/simulator for a population of size $N = 100,000$, which is subdivided into a compartment of hosts interacting with the pathogen and a compartment not interacting with the pathogen based on a disease encounter rate ϕ. Hosts encounter pathogens in a frequency dependent manner and upon encounter between a host of type i and a pathogen of type j

the host gets infected with probability α_{ij}. We run simulations for the six infection matrices in Table 1. Therefore, we first randomly chose one of the possible assignments of α values (0 or 1) to the matrix elements α_{ij} for the given matrix (two possibilities for MA, HR, PI and four possibilities for GFG, iGFG, see Supplementary Text S1). Second, after assigning 0s or 1s to the matrix elements, we replaced each element $\alpha_{ij} = 1$ by randomly drawing a value from a corresponding uniform distribution $\mathcal{U}_{[1-\delta,1]}$, and we replaced each element $\alpha_{ij} = 0$ by drawing from a uniform distribution $\mathcal{U}_{[0,\delta]}$ (Supplementary Text S1). The initial host frequencies h_1 and pathogen p_1 for each simulation are both drawn from a uniform distribution $\mathcal{U}(0.05, 0.5)$. Based on the resulting matrix and initial host and pathogen frequencies, we calculated the frequencies of all possible infected f_{ij} and healthy f_{iz} host phenotypes in the entire population and the respective sub-population ($\tilde{f}_{ij}$ for infected, $\tilde{f}_{iz}$ for healthy) (equations in Table S2). We then randomly picked a sample of $n_I = 902$ haploid infected individuals (drawn from a multinomial distribution $Mult(n_I, \tilde{f}_{11}, \tilde{f}_{12}, \tilde{f}_{21}, \tilde{f}_{22})$) and $n_H = 1006$ haploid healthy individuals (drawn from a binomial distribution $\mathcal{B}(n_H, \tilde{f}_{1z})$). Following this approach we generated 50,000 simulations for each matrix for each combination of $\phi \in \{0.05, 0.5, 0.95\}$ and $\delta \in \{0.1, 0.2, 0.3\}$.

ABC Leave-one-out cross validation for model selection

We first run a leave-one-out cross-validation using our simulated data set to test the suitability of ABC model choice with our four indices as summary statistics to distinguish between the six different matrices. Leave-one-out cross-validation was run separately for each combination of ϕ and δ for a cross-validation sample of size 500 using the function `cv4postpr` in the R-package abc (rejection algorithm, tolerance=0.05) [21].

Application to human data

In the next step we combined two existing human data sets to apply and test our framework. For the infected sample, we used human genome-wide genotype data and HCV whole-genome sequence data from [5]. This data was collected from a total of 541 patients infected by HCV genotypes 2 and 3. We only used a subset of 451 humans of European ancestry to prevent confounding effects of population structure. For the pathogen genome information, we used the viral (nucleotide and protein) data from [5] from NCBI GenBank (accessions KY620313–KY620880). Following [5], we generated whole-genome viral consensus sequences (nucleotide and protein) for each patient using MAFFT (v.7.429) [35]. Future details of how we processed the virus data for our analysis are given in Supplementary Text S1. For the non-infected sample, we used genotype data from the 1,000

Genomes Project Phase 3 [2]. We used the 503 samples from five sub-populations of European ancestry and therefore, retrieved vcf-data from 91 individuals from England and Scotland (GBR), 99 Finnish individuals (FIN), 99 Utah residents with Northern and Western European ancestry (CEU), 107 Spanish individuals (IBS) and 107 Italian individuals (TSI) for a total of 503 genomes (details in Supplementary Text S1).

Co-GWAS

We run a natural co-GWAS with PLINK2 ([48], [17]) on the data using a logistic-regression with the firth-fallback option. For each regression, we used the presence of a particular amino acid at a given position in the viral alignment as a response variable and the genotype at a given human SNP as the genotype. To account for multiple testing we calculated several p-value adjustments using the $--adjust$ option of PLINK2. We incorporate Sex, human PC1-PC3 and virus PC1-PC10 as covariates in the PLINK co-GWAs.

Index calculation with application to the HCV data

We obtained frequencies for each host-virus association from the infected human data set using PLINK2 and vcftools. We also extracted the frequencies of alleles in the non-infected human subsample. Combining these frequencies, we calculated all of our four indices using the equations 4,6,7,8 with customized R-scripts. After that, we retrieved a summary table with the top outlier associations for each index.

Model choice for the top association candidates

The authors thank Joy Bergelson, Dieter Ebert and Lars Raberg for comments on an earlier version of the manuscript. AT acknowledges support from the Deutsche Forschungsgemeinschaft project numbers 274542535 (TE809/3) and 221691301 (TE809/6) within the Priority Program SPP1819 ”Rapid adaptation”. The data aquisition was funded by a grant from the Medical Research Council (MRC) (MR/K01532X/1; to the STOP-HCV Consortium). E.B. is funded by the NIHR Oxford BRC. M.A.A. is supported by a Sir Henry Dale Fellowship jointly funded by the Royal Society and the Wellcome Trust (220171/Z/20/Z).

Data availability

All codes and pipelines developed for index computation and ABC framework, are available at https://gitlab.lrz.de/population_genetics/cogenomics_method. The genomic data can be obtained upon request to the STOP-HCV consortium (https://www.expmedndm.ox.ac.uk/stop-hcv).

Acknowledgments

The authors thank Joy Bergelson, Dieter Ebert and Lars Raberg for comments on an earlier version of the manuscript. AT acknowledges support from the Deutsche Forschungsgemeinschaft project numbers 274542535 (TE809/3) and 221691301 (TE809/6) within the Priority Program SPP1819 ”Rapid adaptation”. The data aquisition was funded by a grant from the Medical Research Council (MRC) (MR/K01532X/1; to the STOP-HCV Consortium). E.B. is funded by the NIHR Oxford BRC. M.A.A. is supported by a Sir Henry Dale Fellowship jointly funded by the Royal Society and the Wellcome Trust (220171/Z/20/Z).

Author contributions

HM, SJ and AT designed the project; HM, SJ and AT developed the model; HM, SJ and LM performed the simulations, LM performed the data analysis, HM, SJ and LM wrote the manuscript with support from AT. STOP-HCV consortium, MAA and VP obtained and pre-processed the sequence data before making them available. The authors declare no conflict of interest.

STOP-HCV Consortium: list of members and affiliations

Eleanor Barnes, Emma Hudson, Paul Klenerman, Peter Simmonds (Nuffield Department of Medicine and the NIHR Oxford BRC, Peter Medawar Building for Pathogen Research, University of Oxford, Oxford, UK); Chris Holmes (Department of Statistics, University of Oxford, Oxford, UK); Graham Cooke (Wright–Fleming Institute, Imperial College London, London, UK); Geoffrey Dusheiko (Institute of Liver Studies, King’s College Hospital NHS Foundation Trust, London, UK); John McLauchlan (MRC–University of Glasgow Centre for Virus Research, Glasgow, UK); Mark Harris (School of Molecular and Cellular Biology, Faculty of Biological Sciences and Astbury Centre for Structural Molecular Biology, University of Leeds, Leeds, UK); William Irving (University of Nottingham, Queen’s Medical Centre, Nottingham, UK); Philip Troke (Gilead Sciences Ltd., London, UK); Diana Brainard and John McHutchinson (Gilead Sciences, Foster City, CA, USA); Charles Gore and Rachel Halford (Hepatitis C Trust, London, UK); Graham R Foster (Queen Mary University of London, London, UK); Cham Herath (Gilead Sciences, Middlesex, UK).

References

[1] A. Agrawal and C. M. Lively. Infection genetics: gene-for-gene versus matching-alleles models and all points in between. *Evolutionary Ecology Research*, 4(1):91–107, 2002.

[2] D. Altshuler, C. Albers, G. Abecasis, and et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015. ISSN 0028-0836.

[3] J. P. Andras, P. D. Fields, L. Du Pasquier, M. Fredericksen, and D. Ebert. Genome-wide association analysis identifies a genetic basis of infectivity in a model bacterial pathogen. *Molecular biology and evolution*, 37(12):3439–3452, 2020.

[4] E. Andreakos, L. Abel, D. C. Vinh, E. Kaja, B. A. Drolet, Q. Zhang, C. O'farrelly, G. Novelli, C. Rodríguez-Gallego, F. Haerynck, et al. A global effort to dissect the human genetic basis of resistance to sars-cov-2 infection. *Nature immunology*, 23(2):159–164, 2022.

[5] M. A. Ansari, V. Pedergnana, C. LC Ip, A. Magri, A. Von Delft, D. Bonsall, N. Chaturvedi, I. Bartha, D. Smith, G. Nicholson, et al. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis c virus. *Nature genetics*, 49 (5):666–673, 2017.

[6] P. Bakker, G. McVean, P. Sabeti, M. Miretti, T. Green, J. Marchini, X. Ke, A. Wijmenga-Monsuur, P. Whittaker, M. Delgado, J. Morrison, A. Richardson, E. Walsh, X. Gao, L. Galver, J. Hart, D. Hafler, M. Pericak-Vance, J. Todd, and J. Rioux. A high-resolution hla and snp haplotype map for disease association studies in the extended human mhc. *Nature genetics*, 38:1166–72, 11 2006. doi: 10.1038/ng1885.

[7] G. Band, E. M. Leffler, M. Jallow, F. Sisay-Joof, C. M. Ndila, A. W. Macharia, C. Hubbart, A. E. Jeffreys, K. Rowlands, T. Nguyen, et al. Malaria protection due to sickle haemoglobin depends on parasite genotype. *Nature*, 602(7895):106–111, 2022.

[8] L. B. Barreiro and L. Quintana-Murci. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, 11(1):17–30, 2010.

[9] I. Bartha, J. M. Carlson, C. J. Brumme, P. J. McLaren, Z. L. Brumme, M. John, D. W. Haas, J. Martinez-Picado, J. Dalmau, C. López-Galíndez, et al. A genome-to-genome analysis of associations between human genetic variation, hiv-1 sequence diversity, and viral control. *elife*, 2:e01123, 2013.

[10] C. Bartoli and F. Roux. Genome-wide association studies in plant pathosystems: toward an ecological genomics approach. *Frontiers in plant science*, 8:763, 2017.

[11] G. Bento, J. Routtu, P. D. Fields, Y. Bourgeois, L. Du Pasquier, and D. Ebert. The genetic basis of resistance and matching-allele interactions of a host-parasite system: The daphnia magna-pasteuria ramosa model. *PLoS Genetics*, 13(2):e1006596, 2017.

[12] G. Bento, P. D. Fields, D. Duneau, and D. Ebert. An alternative route of bacterial infection associated with a novel resistance locus in the daphnia–pasteuria host–parasite system. *Heredity*, 125(4):173–183, 2020.

[13] J. Bergelson, M. Kreitman, E. A. Stahl, and D. Tian. Evolutionary dynamics of plant r-genes. *Science*, 292(5525):2281–2285, 2001.

[14] E. Blanchard, S. Belouzard, L. Goueslain, T. Wakita, J. Dubuisson, C. Wychowski, and Y. Rouillé. Hepatitis c virus entry depends on clathrin-mediated endocytosis. *Journal of Virology*, 80(14):6964–6972, 2006. doi: 10.1128/JVI.00024-06. URL https://journals.asm.org/doi/abs/10.1128/JVI.00024-06.

[15] M. Boots, A. White, A. Best, and R. Bowers. How specificity and epidemiology drive the coevolution of static trait diversity in hosts and parasites. *Evolution*, 68(6):1594–1606, 2014.

[16] J.-L. Casanova and L. Abel. Lethal infectious diseases as inborn errors of immunity: Toward a synthesis of the germ and genetic theories. *Annual Review of Pathology: Mechanisms of Disease*, 16(1):23–50, 2021. doi: 10.1146/annurev-pathol-031920-101429. URL https://doi.org/10.1146/annurev-pathol-031920-101429. PMID: 32289233.

[17] C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 02 2015. ISSN 2047-217X. doi: 10.1186/s13742-015-0047-8. URL https://doi.org/10.1186/s13742-015-0047-8. s13742-015-0047-8.

[18] A. Chauhan and M. Khandkar. Endocytosis of human immunodeficiency virus 1 (hiv-1) in astrocytes: A fiery path to its destination. *Microbial Pathogenesis*, 78:1–6, 2015. ISSN 0882-4010. doi: https://doi.org/10.1016/j.micpath.2014.11.003. URL https://www.sciencedirect.com/science/article/pii/S0882401014001673.

[19] A. Chauhan, R. Mehla, T. S. Vijayakumar, and I. Handy. Endocytosis-mediated hiv-1 entry and its significance in the elusive behavior of the virus in astrocytes. *Virology*, 456-457:1–19, 2014. ISSN 0042-6822. doi: https://doi.org/10.1016/j.virol.2014.03.002. URL https://www.sciencedirect.com/science/article/pii/S0042682214000841.

[20] K. E. Coller, K. L. Berger, N. S. Heaton, J. D. Cooper, R. Yoon, and G. Randall. Rna interference and single particle tracking analysis of hepatitis c virus endocytosis. *PLOS Pathogens*, 5(12):1–

14, 12 2009. doi: 10.1371/journal.ppat.1000702. URL `https://doi.org/10.1371/journal.ppat.1000702`.

[21] K. Csillery, O. Francois, and M. G. B. Blum. abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, 2012. doi: 10.1111/j.2041-210X.2011.00179.x.

[22] C. Demirjian, F. Vailleau, R. Berthomé, and F. Roux. Genome-wide association studies in plant pathosystems: success or failure? *Trends in Plant Science*, 28(4), 2023. doi: 10.1016/j.tplants.2022.11.006.

[23] E. Dexter, P. Fields, and D. Ebert. Uncovering the genomic basis of infection through co-genomic sequencing of hosts and parasites. *bioRxiv*, pages 2022–12, 2022.

[24] A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Molecular Biology and Evolution*, 22 (5):1185–1192, 02 2005. doi: 10.1093/molbev/msi103. URL `https://doi.org/10.1093/molbev/msi103`.

[25] M. F. Dybdahl, C. E. Jenkins, and S. L. Nuismer. Identifying the molecular basis of host-parasite coevolution: merging models and mechanisms. *The American Naturalist*, 184(1):1–13, 2014.

[26] E. Ebranati, A. Mancon, M. Airoldi, S. Renica, R. Shkjezi, P. Dragusha, C. Della Ventura, A. R. Ciccaglione, M. Ciccozzi, S. Bino, E. Tanzi, V. Micheli, E. Riva, M. Galli, and G. Zehender. Time and mode of epidemic hcv-2 subtypes spreading in europe: Phylodynamics in italy and albania. *Diagnostics*, 11(2), 2021. ISSN 2075-4418. doi: 10.3390/diagnostics11020327. URL `https://www.mdpi.com/2075-4418/11/2/327`.

[27] A. Fenton, J. Antonovics, and M. A. Brockhurst. Inverse-gene-for-gene infection genetics and coevolutionary dynamics. *The American Naturalist*, 174(6):E230–E242, 2009.

[28] S. Gandon and Y. Michalakis. Local adaptation, evolutionary potential and host–parasite coevolution: interactions between migration, mutation, population size and generation time. *Journal of Evolutionary Biology*, 15(3):451–462, 2002.

[29] S. Gandon, A. Buckling, E. Decaestecker, and T. Day. Host–parasite coevolution and patterns of adaptation across time and space. *Journal of evolutionary biology*, 21(6):1861–1866, 2008.

[30] M. Ghafari, P. Simmonds, O. G. Pybus, and A. Katzourakis. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. *Current Biology*, 31(21): 4689–4696, 2021.

[31] C. A. Gilligan. Sustainable agriculture and plant diseases: an epidemiological perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1492):741–759, 2008.

[32] A. D. Gloss, A. Vergnol, T. C. Morton, P. J. Laurin, F. Roux, and J. Bergelson. Genome-wide association mapping within a local arabidopsis thaliana population more fully reveals the genetic architecture for defensive metabolite diversity. *Philosophical Transactions of the Royal Society B*, 377(1855):20200512, 2022.

[33] A. R. Hall, P. D. Scanlan, A. D. Morgan, and A. Buckling. Host–parasite coevolutionary arms races give way to fluctuating selection. *Ecology letters*, 14(7):635–642, 2011.

[34] A. V. Hill, A. Jepson, M. Plebanski, and S. C. Gilbert. Genetic analysis of host–parasite coevolution in human malaria. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1359):1317–1325, 1997.

[35] K. Katoh, K. Misawa, K. Kuma, and T. Miyata. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066, 07 2002. ISSN 0305-1048. doi: 10.1093/nar/gkf436. URL `https://doi.org/10.1093/nar/gkf436`.

[36] J. A. Lees, B. Ferwerda, P. H. Kremer, N. E. Wheeler, M. V. Serón, N. J. Croucher, R. A. Gladstone, H. J. Bootsma, N. Y. Rots, A. J. Wijmega-Monsuur, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. *Nature communications*, 10(1):2176, 2019.

[37] L. Liu, D. C. Cara, J. Kaur, E. Raharjo, S. C. Mullaly, J. Jongstra-Bilen, J. Jongstra, and P. Kubes. LSP1 is an endothelial gatekeeper of leukocyte transendothelial migration . *Journal of Experimental Medicine*, 201(3):409–418, 01 2005. ISSN 0022-1007. doi: 10.1084/jem.20040830. URL `https://doi.org/10.1084/jem.20040830`.

[38] P. Luijckx, H. Fienberg, D. Duneau, and D. Ebert. A matching-allele model explains host resistance to parasites. *Current Biology*, 23(12):1085–1088, 2013.

[39] A. MacPherson, S. P. Otto, and S. L. Nuismer. Keeping pace with the red queen: Identifying the genetic basis of susceptibility to infectious disease. *Genetics*, 208(2):779–789, 2018.

[40] H. Märkle, S. John, A. Cornille, P. D. Fields, and A. Tellier. Novel genomic approaches to study antagonistic coevolution between hosts and parasites. *Molecular Ecology*, 30(15):3660–3676, 2021.

[41] S. Merani, D. Petrovic, I. James, A. Chopra, D. Cooper, E. Freitas, A. Rauch, J. di Iulio, M. John, M. Lucas, K. Fitzmaurice, S. Mckiernan, S. Norris, D. Kelleher, P. Klenerman, and S. Gaudieri.

Effect of immune pressure on hepatitis c virus evolution: Insights from a single-source outbreak. *Hepatology (Baltimore, Md.)*, 53:396–405, 02 2011. doi: 10.1002/hep.24076.

[42] K. Mohd Hanafiah, J. Groeger, A. D. Flaxman, and S. T. Wiersma. Global epidemiology of hepatitis c virus infection: New estimates of age-specific antibody to hcv seroprevalence. *Hepatology*, 57(4):1333–1342, 2013. doi: https://doi.org/10.1002/hep.26141. URL `https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.26141`.

[43] B. Moury, J.-M. Audergon, S. Baudracco-Arnas, S. Ben Krima, F. Bertrand, N. Boissot, M. Buisson, V. Caffier, M. Cantet, S. Chanéac, et al. The quasi-universality of nestedness in the structure of quantitative plant-parasite interactions. *Peer Community Journal*, 1, 2021.

[44] A. Nemri, S. Atwell, A. M. Tarone, Y. S. Huang, K. Zhao, D. J. Studholme, M. Nordborg, and J. D. Jones. Genome-wide survey of arabidopsis natural variation in downy mildew resistance using combined association and linkage mapping. *Proceedings of the National Academy of Sciences*, 107(22):10302–10307, 2010.

[45] A. Petruzziello, M. Samantha, G. Loquercio, A. Cozzolino, and C. Cacciapuoti. Global epidemiology of hepatitis c virus infection: An up-date of the distribution and circulation of hepatitis c virus genotypes. *World Journal of Gastroenterology*, 22:7824, 09 2016. doi: 10.3748/wjg.v22.i34.7824.

[46] M. Pogoda, F. Liu, D. Douchkov, A. Djamei, J. C. Reif, P. Schweizer, and A. W. Schulthess. Identification of novel genetic factors underlying the host-pathogen interaction between barley (hordeum vulgare l.) and powdery mildew (blumeria graminis f. sp. hordei). *PLoS One*, 15(7): e0235565, 2020.

[47] K. Pulford, M. Jones, A. Banham, E. Haralambieva, and D. Mason. Lymphocyte-specific protein 1: a specific marker of human leucocytes. *Immunology*, 96(2):262—271, February 1999. ISSN 0019-2805. doi: 10.1046/j.1365-2567.1999.00677.x. URL `https://doi.org/10.1046/j.1365-2567.1999.00677.x`.

[48] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. Ferreira, D. Bender, J. Maller, P. Sklar, P. Bakker, M. Daly, and P. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81:559–75, 10 2007. doi: 10.1086/519795.

[49] L. Råberg. Human and pathogen genotype-by-genotype interactions in the light of coevolution theory. *PLOS Genetics*, 19(4):1–17, 04 2023. doi: 10.1371/journal.pgen.1010685. URL `https://doi.org/10.1371/journal.pgen.1010685`.

[50] P. D. Scanlan. Bacteria–bacteriophage coevolution in the human gut: implications for microbial diversity and functionality. *Trends in microbiology*, 25(8):614–623, 2017.

[51] E. A. Stahl, G. Dwyer, R. Mauricio, M. Kreitman, and J. Bergelson. Dynamics of disease resistance polymorphism at the rpm1 locus of arabidopsis. *Nature*, 400(6745):667–671, 1999.

[52] P. Sudmant, T. Rausch, E. Gardner, R. Handsaker, A. Abyzov, J. Huddleston, Y. Zhang, K. Ye, G. Jun, M. Fritz, M. Konkel, A. Malhotra, A. Stütz, X. Shi, F. Casale, J. Chen, F. Hormozdiari, G. Dayama, K. Chen, M. Malig, M. Chaisson, K. Walter, S. Meiers, S. Kashin, E. Garrison, A. Auton, H. Lam, X. Mu, C. Alkan, D. Antaki, T. Bae, E. Cerveira, P. Chines, Z. Chong, L. Clarke, E. Dal, L. Ding, S. Emery, X. Fan, M. Gujral, F. Kahveci, J. Kidd, Y. Kong, E. Lameijer, S. McCarthy, P. Flicek, R. Gibbs, G. Marth, C. Mason, and A. Menelaou. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, Sept. 2015. ISSN 0028-0836. doi: 10.1038/nature15394.

[53] A. Tellier and J. K. Brown. Stability of genetic polymorphism in host–parasite interactions. *Proceedings of the Royal Society B: Biological Sciences*, 274(1611):809–817, 2007.

[54] A. Tellier, S. Moreno-Gámez, and W. Stephan. Speed of adaptation and genomic footprints of host–parasite coevolution under arms race and trench warfare dynamics. *Evolution*, 68(8): 2211–2224, 2014.

[55] J. N. Thompson and J. J. Burdon. Gene-for-gene coevolution between plants and parasites. *Nature*, 360(6400):121–125, 1992.

[56] F. M. Tomley and M. W. Shirley. Livestock infectious diseases and zoonoses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1530):2637–2642, 2009. doi: 10.1098/rstb.2009.0133. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2009.0133`.

[57] T. Walther, J. Brickner, P. Aguilar, S. Bernales, C. Pantoja, and P. Walter. Eisosomes mark static sites of endocytosis. *Nature*, 439:998–1003, 03 2006. doi: 10.1038/nature04472.

[58] M. Wang, F. Roux, C. Bartoli, C. Huard-Chauveau, C. Meyer, H. Lee, D. Roby, M. S. McPeek, and J. Bergelson. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the National Academy of Sciences*, 115(24):E5440–E5449, 2018.