BMC Evolutionary Biology

Research article

Open Access

Patterns of genetic variation in populations of infectious agents

Isabel Gordo*[1] and Paulo RA Campos[2]

Address: [1]Instituto Gulbenkian de Ciência, P-2781-901 Oeiras, Portugal and [2]Departamento de Física, Universidade Federal Rural de Pernambuco 52171-900, Dois Irmãos, Recife-PE, Brazil

Email: Isabel Gordo* - igordo@igc.gulbenkian.pt; Paulo RA Campos - paulo.campos@df.ufrpe.br

* Corresponding author

Published: 13 July 2007

BMC Evolutionary Biology 2007, **7**:116 doi:10.1186/1471-2148-7-116

Received: 25 September 2006
Accepted: 13 July 2007

This article is available from: http://www.biomedcentral.com/1471-2148/7/116

© 2007 Gordo and Campos; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: The analysis of genetic variation in populations of infectious agents may help us understand their epidemiology and evolution. Here we study a model for assessing the levels and patterns of genetic diversity in populations of infectious agents. The population is structured into many small subpopulations, which correspond to their hosts, that are connected according to a specific type of contact network. We considered different types of networks, including fully connected networks and scale free networks, which have been considered as a model that captures some properties of real contact networks. Infectious agents transmit between hosts, through migration, where they grow and mutate until elimination by the host immune system.

Results: We show how our model is closely related to the classical SIS model in epidemiology and find that: depending on the relation between the rate at which infectious agents are eliminated by the immune system and the within host effective population size, genetic diversity increases with R_0 or peaks at intermediate R_0 levels; patterns of genetic diversity in this model are in general similar to those expected under the standard neutral model, but in a scale free network and for low values of R_0 a distortion in the neutral mutation frequency spectrum can be observed; highly connected hosts (hubs in the network) show patterns of diversity different from poorly connected individuals, namely higher levels of genetic variation, lower levels of genetic differentiation and larger values of Tajima's D.

Conclusion: We have found that levels of genetic variability in the population of infectious agents can be predicted by simple analytical approximations, and exhibit two distinct scenarios which are met according to the relation between the rate of drift and the rate at which infectious agents are eliminated. In one scenario the diversity is an increasing function of the level of transmission and in a second scenario it is peaked around intermediate levels of transmission. This is independent of the type of host contact structure. Furthermore for low values of R_0, very heterogeneous host contact structures lead to lower levels of diversity.

Background

Patterns of genetic diversity in populations of infectious agents contain important information about their epidemiology and evolution. They depend on the population dynamics of the infectious agents, which involves their replication within hosts and transmission between hosts, their mutation and recombination rate. Infectious agents vary enormously in their ability to mutate and to transmit,

which will lead to large differences in levels of variability. Furthermore there can be variation within an infectious species for the ability to evade the host immune system. In fact, infectious agent genetic diversity can help in targeting genes under selection pressure created by the immune system [1]. In addition patterns of infectious agent variation can, under certain circumstances, be used to infer host population history [2], and the level of infectious agent genetic structure may reflect its evolutionary potential [3]. Importantly, the need for a continuous integration between population genetics and epidemiology has been increasingly recognized [4-7].

In population genetics the standard neutral model has a long history in DNA sequence data analysis [8], and has been extensively used as a null model for understanding genetic variation in natural populations, including that in our own species [8,9]. The standard neutral model makes several simplifying assumptions: in particular it makes the simple assumption that individuals form one single constant size population. When considering populations of infectious agents it is much more reasonable to assume, as the null model, a population composed of a collection of much smaller populations.

Here we develop population genetics models of structured populations, that incorporate epidemiological parameters explicitly, in order to study genetic variability under one of the simplest possible epidemiological models. We ask mainly two questions: 1) what do levels and patterns of sequence variation in these infectious agents look like under this model? And 2) how does host contact structure influence their diversity?

The models we will study here are very similar to the metapopulation models where each subpopulation can go extinct and be recolonized [10-12]. Generally studies of genetic diversity in such subdivided populations [13,14] assume a simple symmetric topology for the metapopulation – the most well studied is the island model of Wright. Simple as it is, this model has provided a wealth of results that have led to enormous contributions to our understanding of evolution in structured populations [15,16]. Nevertheless, there are several reasons to think that this model is too simple to be readily applicable to natural populations [14,17], especially if the goal is to understand molecular diversity of infectious agents. As we know, the underlying topology at which certain disease epidemics and spreading takes place is that of social networks [18]. Several recent investigations have demonstrated that real networks of interaction have a much more complex structure than those predicted by totally regular networks or totally random networks [19]. Most real networks of social interactions present two different topological properties: a low average pairwise distance between nodes and a high clustering degree (which measures local structuring).

The former occurs in random networks and the latter in regular networks. In such way, some models of network topologies have been recently proposed in the literature (for a review see Ref. [20]). One of the most successful models for network structure is the scale-free network [21]. In addition to the common properties of real interaction networks, in scale-free networks the distribution of connectivities obeys a power-law distribution as $P(k_i) \propto k_i^{-\gamma}$, which is observed in some actual systems ranging from World Wide Web to the network of human sexual contacts [22,23]. As initially proposed, scale-free networks are dynamical networks where growth and preferential attachment are some of the key mechanisms.

Accordingly, each newly introduced node in the network preferentially joins with an already well connected-node. As a result, it will produce a highly heterogeneous network where most nodes have a low connectivity while a few nodes display a very large connectivity. These latter ones are referred to as hubs. The understanding of the interplay between the underlying topology and the forces driving systems is of crucial relevance [24,25]. One example of this, that has received a great deal of attention, is that of network epidemiology: the study of epidemic and disease spreading [26-29], which are strictly tied to the topology of social contact networks. In this context, a striking result has arisen from the study of the classical susceptible-infected-susceptible (SIS) epidemiological model on scale free networks: scale-free networks are more prone to spreading of diseases than random graphs and regular lattices [26,27]. In this kind of model the role of microbe evolution is disregarded. Recently, we have focused on this latter feature and we have shown that although scale-free networks are more prone to infectious agent spread, the accumulation of deleterious mutations in asexual infectious agent with high mutation rates can also be accelerated in this kind of networks in comparison to random graphs [30]. This shows that not only disease dynamics but also its evolution should be considered as an important key in the investigation of epidemiological models [7]. Another very important feature that has to be considered is co-evolution between infectious agent and their hosts [31]. Modeling of these complex systems have provided us with insights into how host-parasite interactions can modulate the mode of reproduction [32], ploidy levels [33], the patterns of gene expression in hosts and parasites [34] and how different types of interspecies interactions affect genetic and phenotypic variation [35].

Results and Discussion

Levels of metapopulation infection

The susceptible-infected-susceptible model (SIS model) is one of the simplest classical models in epidemiology. In this model, hosts born susceptible (S) can become infected (I) at a rate β per unit time, given contact with at least one infected host. Infected hosts become susceptible at a rate λ, such that $1/\lambda$ is the average duration of an infection. One of the most fundamental quantities to assess the equilibrium frequency of infections in the population is the R_0 of the infectious agent. The R_0 is defined as the number of secondary cases produced by an infectious individual in a totally susceptible population. At epidemiological equilibrium, the frequency of infected individuals is $i = 1 - 1/R_0$, with $R_0 = \beta/\lambda$. If $R_0 < 1$ then the infection does not spread.

To assess the patterns of variation under the SIS model, we have studied a population genetic model of a structured population that is composed of many small subpopulations, which are named demes. There is a total of D demes, which are connected according to a given network topology: corresponding to either the island model or the scale free network. These demes can go extinct and be recolonized through migrants that they received from the other demes. Each deme can contain at most N_d individuals, which reproduce and mutate within each deme (see Methods). In Table 1 we make a summary of the model's key parameters.

We now relate our metapopulation model with the SIS model and in this study we will ask what equilibrium patterns of infectious agent genetic variation look like under this model. In our model a deme corresponds to a host. An empty deme means that the host is susceptible, whereas a deme which is full corresponds to an infected host. A deme that is currently full can become empty with probability e, which means that e corresponds to λ. A deme that is currently empty can become full through the migrants it receives from nearby demes. This implies that β is proportional to m. Given that the average connectivity of a deme is K and that the number of migrants per link is $N_d m$, then β corresponds to $N_d mK$.

In order to assess the correspondence between our model and the SIS model, we have compared the average frequency of infected individuals in our metapopulation with the expectation for the deterministic SIS model, which implies that:

$$i = 1 - 1/R_0 = 1 - e/N_d mK \quad (1)$$

Equation 1 is the expected frequency when there is no variance in k_i, which is not the case in scale free networks.

In Figure 1 we show the results from our simulations, where the proportion of infected individuals in the metapopulation is measured as we increase the transmission coefficient of the infectious agent, β, through increments in m. The results for the different types of networks considered are shown, and the deterministic expectation is also plotted. In all cases $R_0 = N_d mK/e$, where $K = D - 1$ for the island model and $K = 6$ for the scale free topology. The results of the simulations show that the proportion of infected individuals observed and that predicted are quite concordant. In particular, if we assume the topology corresponding to the island model, then the level of infection is exactly that predicted by Equation 1. We notice that the prediction holds for an effectively infinite population under the mass action assumption.

One may expect deviations to be observed when these assumptions are violated [36]. Nevertheless the deviations we observe are small, unless R_0 is very low. In fact in the case of very low R_0 there is a high probability that the infection does not spread. For example in the scale free network, if the infection starts in a poorly connected host it may have very little chance of spreading. We performed simulations with the scale free topology in conditions where the infection starts in a single randomly chosen host. With the same parameters as in Figure 1 and for $R_0 = 1.5$, we observed 66% of cases where the disease could not spread. With $R_0 = 3$, the fraction of cases where the infectious agent could not invade dropped to 40%.

Levels of metapopulation diversity

We now study the level of genetic diversity in infectious agents sampled randomly from the whole population of infected hosts. We first consider a metapopulation where every host contacts every other host. This corresponds to

Table 1: Model parameterization

parameter	meaning in metapopulation genetics	meaning in epidemiology
D	number of demes	number of hosts
N_d	number of individuals within a deme	number of infectious agents within an infected host
e	probability that a deme goes extinct	probability that the immune system clears the infection
m	migration rate	transmission ability between hosts
μ	mutation rate	mutation rate of the infectious agent
k_j	number of demes connected to deme j	number of contacts of host j

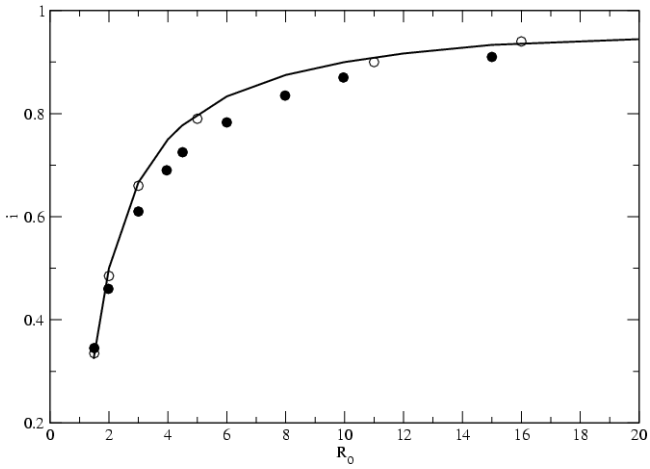

Figure 1
Infected Individuals. The proportion of infected individuals, *i*, with increasing $R_0 = N_d mK/e$. Full circles are the results for scale free networks (with $\gamma = 3$) and empty circles for the island model. $D = 900$, $N_d = 10$, $e = 0.01$ in all network topologies. The line denotes the expected value of *i* under the deterministic SIS model.

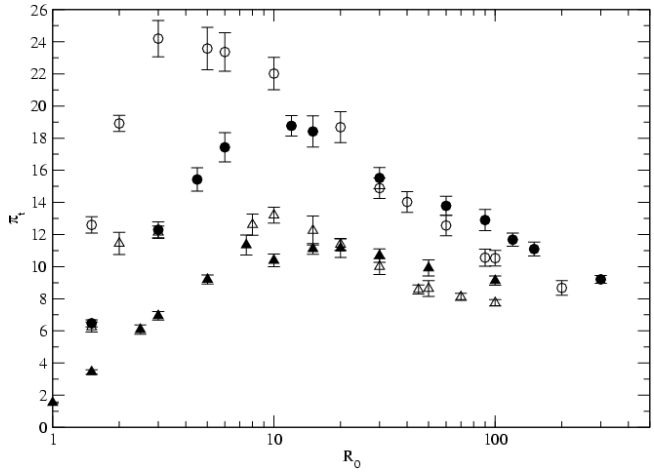

Figure 2
Diversity in the metapopulation. The level of diversity π_t as a function of $R_0 = N_d mK/e$. The parameters values are $D = 1000$, $N_d = 10$, $n_t = 50$ and $\mu = 0.0004$. The empty symbols denote the results for the island model while the full symbols correspond to scale-free networks with $\gamma = 3$. The results for $e = 0.01$ are represented by circles and $e = 0.02$ by triangles.

the island model in the populations genetics literature and mass action assumption in epidemiological models. We then assess how the level of diversity is affected by differences in the level of contact between hosts, in particular when a small number of hosts can have a very large number of contacts, such as in the scale free network. In Figure 2 we show the level of diversity in samples taken from the whole population, π_t, as we increase R_0, through increments in *m*.

We observe that, for both topologies and for the sets of parameters considered, the level of π_t is maximal for intermediate values of R_0. For instance, when $e = 0.01$ this maximum value is achieved at R_0 around 3 for the island model and around 10 for the scale free topology. Beyond these points the level of diversity starts to decrease with increasing R_0. From Figure 2 we observe the occurrence of two quite distinct regimes, according to the level of transmission. In the region of low transmission, R_0 is small, extinction is much stronger than migration ($e >> mK$), the fraction of infected hosts is small and levels of diversity are low. In fact, starting from $R_0 = 1$, where the fraction of infected individuals, *i*, is 0, as we increase R_0 (by increasing *m*), the level of infection rises and the level of diversity accompanies that increase. In this region the level of infection bounds the level of diversity in the population, since it is expected that diversity will be higher when the total number of infectious agents in the metapopulation is larger. When the level of infection achieves a value close to 0.9, increments in *m*, lead to small increments in *i* and the level of diversity stops increasing. The second regime comes about at high transmission, where R_0 is very large. In this region migration is much stronger than extinction, $mK >> e$, the level of infection is close to 1 and so it is not the limiting factor for diversity to grow. From this point, increments in migration cause a drastic reduction in the isolation between demes and lead to a reduction in diversity. In fact in the limit of extremely high levels of migration the diversity in the structured population tends to that expected in a panmitic population of size $N_t = DN_d$. So, for very high values of R_0, diversity tends toward the value $\pi_t = \pi_d = 2N_d D_\mu$, which in the case of Figure 2 is 8, for the value of the mutation rate, μ, assumed. Figure 2 also shows that in the region of low R_0, diversity in the island model is higher than in the scale free network, whereas for large values of R_0, there is little difference between the topologies. The latter is expected since the larger the value of the migration rate the less important the precise contact structure will be. The former can be understood as follows: a low value of R_0 corresponds to a small fraction of infected hosts both in the island model and in the scale free network. But whereas in the island model new infections of a susceptible host occur from contact with any of the infected hosts in the metapopulation, in the scale free network infections are more likely to come from well connected hosts, which are a small subset of the metapopulation. This then will lead to lower diversity levels in the scale free network, as compared to the island model, for the same low R_0 value.

We have compared our simulation results with some of the analytical approximations for the levels of diversity in metapopulations [13]. In the vast majority of metatopula-

tion models with extinction and recolonization, the island model of population subdivision is assumed. Furthermore, the processes of migration and recolonization are assumed to be distinct. Two different schemes of colonization are normally considered, according to where colonists come from: the migrant pool model and the propagule pool model [14,37,38]. In both models there are k colonists (where k is a fixed number independent of migration), which may constitute a random sample from the whole metapopulation (migrant pool model) or from a single deme (propagule model). Pannell and Charlesworth [13] have studied levels of within and between population diversity under these models and have provided a set of analytical approximations. We have adapted the approximations in their Table 2, which correspond to the infinite sites mutation model as we assume here, to the metapopulation model that we are studying, which is slightly different from the one they have used. In particular, besides the different types of contact structure studied here, there are two key differences in the models: 1) in our model recolonization and migration are similar processes; and 2) while in the classical model it is assumed that when one deme goes extinct it gets immediately recolonized, in our model when a deme goes extinct it will only be recolonized when it receives migrants. In this way, the equilibrium number of empty demes (susceptible individuals) decreases as m, or R_0, increases. Whereas for infectious agent populations assumption 2) is more appropriate, for some infectious agents assumption 1) may be too simple. One can imagine that when a host is infected, its ability to transmit the infectious agents to another infected host is reduced compared to its ability to transmit the infectious agent to a susceptible individual. This implies that the migration rate between subpopulations may, in some infectious agents, depend on the host history. We have taken the simplest scenario here.

We have thus compared the expected level of metapopulation genetic diversity in our simulations for the symmetric island model (where every host contacts every other host) with the following approximation:

Table 2: Mean values of Tajima's D in the scale free network with parameters

n_t	D_t 2SE
10	-0.103 0.095
25	-0.190 0.076
75	-0.328 0.110
150	-0.274 0.114
250	-0.392 0.110
300	-0.422 0.092

Parameter values: D = 1000; N_d = 10; μ = 0.0004, e = 0.01, R_0 = 1.5

$$\pi_{t_{isl}} = 2N_d D(1-1/R_0)\mu \frac{R_0 + 1/2e}{N_d + R_0} \quad (2)$$

which is adapted from the approximation for the classical case of the migrant pool model of recolonization. We can expect that in the case of the scale free topology, where a large number of hosts have few connections and a few hosts are very well connected, levels of diversity will be closer to those expected for the propagule pool recolonization model. This is because hubs in the network will contribute much more than the other nodes in the process of recolonization. We have thus compared the expected level of genetic diversity for the scale free network with the following approximation for the propagule pool model:

$$\pi_{t_{sf}} = D(1-1/R_0)\mu \frac{1-e}{e(2-e)} \quad (3)$$

which is valid only when $mK < e$ [13]. We therefore expect this expression to provide a good approximation for cases in which $R_0 < N_d$.

As seen in Figure 3, these formulas provide very good approximations to the simulation results, for low values of R_0. For very large values of R_0, the level of diversity is similar in the two topologies and is very well approximated by Equation 2.

Equation 2 suggests a strong dependence of the level of metapopulation diversity with N_d, the effective population size within a host. This effective population size is

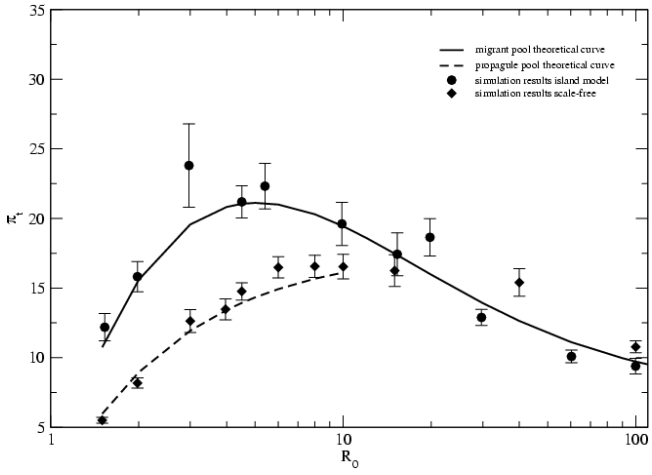

Figure 3
Theoretical approximations and the different topologies. Comparison of the level of diversity π_t between topologies and with the theoretical approximations. D = 900, N_d = 10, e = 0.01, n_t = 50 and μ = 0.0004 in all networks.

likely to vary considerably among different infectious agent species. We have therefore explored how the value of N_d affects the levels of diversity with simulations.

In Figure 4 we show the results of varying N_d, for both types of network (island model in the left panel and scale free network in the right panel). Figure 4 clearly shows that when $e < 1/N_d$ (filled symbols in both panels), levels of diversity are maximal for intermediate R_0. But for $e > 1/N_d$ diversity always increases with R_0. This occurs both in the island model and in scale free networks. This shows that, independently of the type of host contact structure, for infectious agents with large intrahost effective population size, levels of diversity increase with increasing R_0.

Furthermore, as suggested by Equation 2, for small values of R_0, increasing N_d has a very small effect on the level of diversity, but for intermediate to high R_0 values the effect is more pronounced.

When $R_0 > 10$, the level of infection is not a limiting factor in the level of diversity, because the number of infected hosts is very high. Thus for large values of R_0 infectious agent diversity will increase with N_d.

Comparing the panels in Figure 4, we can observe that when $R_0 << 10$, diversity is always smaller in the scale free network, whereas when $R_0 >> 10$ and $e > 1/N_d$ the levels of diversity are similar in both contact networks. In fact, for large values of R_0, the largest difference between the topologies can be observed when $e = 1/N_d$.

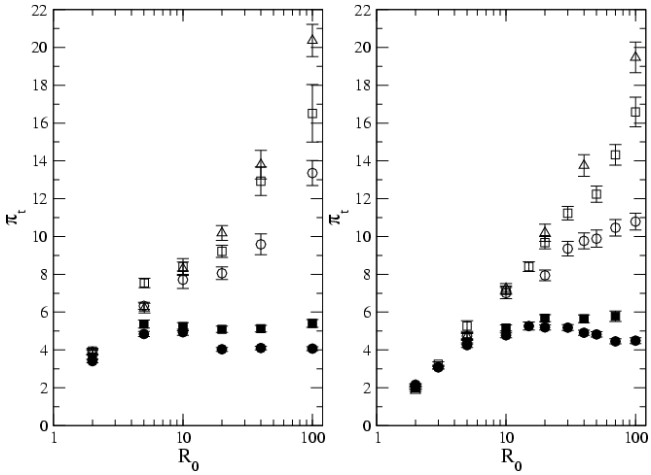

Figure 4
Effect of N_d in metapopulation diversity. The level of diversity, π_t, with R_0 and N_d in the island model (left panel) and scale free networks (right panel). N_d = 5 in full circles, N_d = 7 in full squares, N_d = 20 in open circles, N_d = 40 in open squares and N_d = 60 in open triangles. Other parameters are D = 1000, e = 0.05, n_t = 50 and μ = 0.0004 for all networks.

In this metapopulation model there are two forces which generate diversity within each host: mutation and transmission; there are also two forces that undermine diversity: extinction and genetic drift. So in general, we can expect that, when the forces that reduce diversity are stronger than those that generate it (that is low R_0, low N_d or high e), diversity levels will be low. On the contrary, high R_0, high N_d or low e, we can expect levels of diversity to be much larger.

Metapopulation mutation frequency spectrum

The spectrum of frequencies of mutations that are segregating in the population is important to understand deviations from the standard neutral model, which assumes an undivided, constant size population at equilibrium between mutation and drift [39]. In fact, the mutational spectrum of infectious agent gene sequences has been used to reject the standard neutral model suggesting that natural selection is determining the evolution of certain genes [40,41]. Tajima's D is a widely used statistic to assess distortions in the frequency spectrum [42]. If the number of mutations that appear at frequency $1/n$ in sample of size n (singletons) is higher than that expected under the standard neutral model, then Tajima's D becomes negative. On the other hand if the number of mutations at intermediate frequency is large then Tajima's D becomes positive. When a departure from the standard neutral model is observed in a given gene of a given species, several alternative hypotheses can be made. These typically involve natural selection and/or demographic factors, such as population growth or population structure. In infectious agent populations the relevant null model against which we would like to test for the molecular signature of selection is closer to a metapopulation neutral model than to the standard neutral model. From all the simulations in all the metapopulation structures we have studied, we have observed that D_t was always very close to 0. This is in agreement with the results of coalescent theory and simulations in metapopulations under the island model [43,44]. However, we have observed that in some simulations of scale free networks a slight distortion of the frequency spectrum was apparent. In cases of low R_0 mean values of the Tajima's D statistic become negative. In Table 2 we show one example where this occurs. Although the values of D_t are not very negative when the sample size is small, they become more negative with increasing sample size.

In Figure 5 we show an example of the mutation frequency spectrum in the scale free network for two values of transmission with a large sample size. Clearly we see that when R_0 is small the proportion of singletons in the samples is much higher than when R_0 is large. For R_0 = 15 the spectrum is similar to the one expected under the standard neutral model. These results imply that it is very

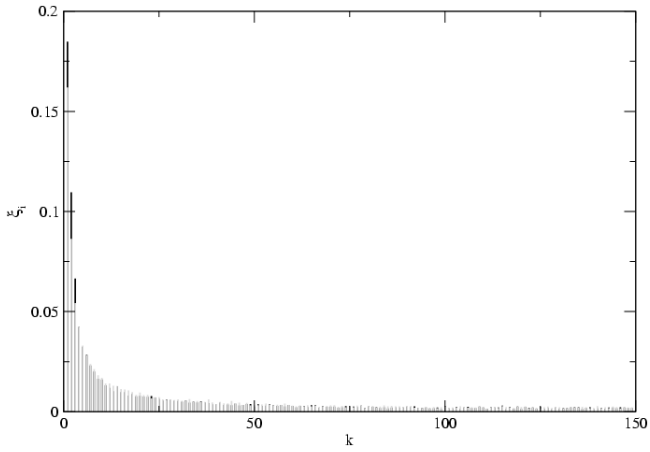

Figure 5
Frequency spectrum. The frequency spectrum of neutral mutations in scale free networks with $\gamma = 3$. In the Y-axis we plot the probability that in a sample of size $n_t = 300$ we find mutations with frequency k/n_t or with frequency $(n_t - k)/n_t$. $D = 1000$, $N_d = 10$, $e = 0.01$ and $\mu = 0.0004$. Black bars correspond to $R_0 = 1.5$ and grey bars to $R_0 = 15$.

dificult to reject an equilibrium neutral model with constant population size when using Tajima's D.

Infectious agent diversity within hosts and differentiation amongst hosts

In infectious agents with very high mutation rates, as it is the case of RNA viruses [45], one may expect some level of within host diversity to be observed. We have therefore studied the level of diversity in samples taken from each infected host. We also studied the level of genetic differentiation between hosts measured by F_{ST} [46]. The statistic of genetic differentiation we use measures the difference between the level of infectious agent diversity within an infected host and that of the entire infectious agent metapopulation. It is known that all of these statistics are important for the understanding of the relative importance of the processes governing metapopulation dynamics [13] and therefore they can be important in understanding their epidemiology. Figure 6 shows the results for the levels of genetic differentiation, as measured by F_{ST}, for different values of R_0. As can be seen from this figure, for infectious agents with low transmission, the levels of host differentiation are very high. In this case and with the value of the mutation rate considered, the levels of within host genetic variability can be very low. For example in the case of the island model with $R_0 = 2.5$, the observed mean level of intrahost infectious agent diversity was 0.46, which is only 0.04 of the level observed in the whole metapopulation. In certain RNA viruses, such as the Dengue virus, the levels of intrahost genetic diversity that have been observed are about 0.03 of that between hosts [47]. In infectious agents with high transmission rates the levels of differentiation are much smaller and are accompanied by higher levels of within host diversity. In the case of the island model every host has similar levels of diversity. And so, as the infectious agent transmission rate increases, so does the level of within host diversity. But in scale free networks, host connectivity affects infectious agent diversity within that host and the levels of differentiation between hosts. We have considered the case of two infectious agent with different transmission coeficients and have looked at the relation between host connectivity and within host diversity. Figure 7 shows that well connected hosts have much higher levels of π_d and much lower levels of F_{ST}. In this instance, F_{ST} reflects the average divergence between demes with connectivity k_i and all other demes. From the figure we see that well connected hosts also show significantly larger mean values of the Tajima's D statistic, for intermediate values of R_0.

Conclusion

One of the main goals in infectious disease research is to understand how infectious agent variation, host immunity, transmission dynamics and epidemic dynamics determine patterns of infectious agent evolution. Information about evolutionary and epidemiological processes can be extracted from studying infectious agent genetic diversity. In particular it can help us to understand the origin of disease and the selective pressures that act on certain infectious agent genes. The link between infectious agent dynamics and genetic diversity at within and

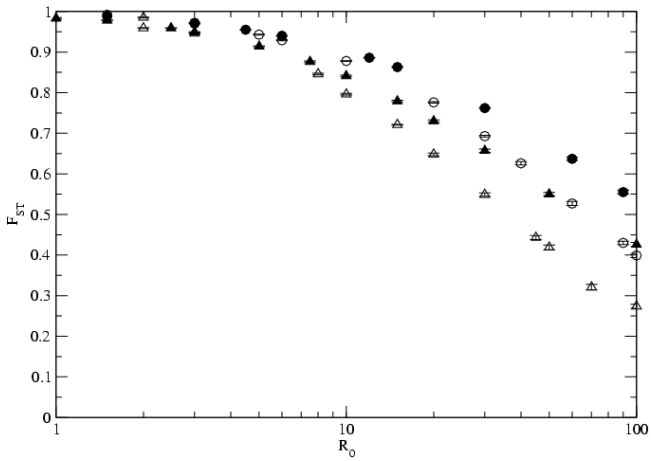

Figure 6
Level of differentiation. The level of differentiation among hosts measured as F_{ST}. The empty symbols denote the results for the island model, while the full symbols correspond to scale-free networks. The parameters are $e = 0.01$ (circles) and $e = 0.02$ (triangles), $D = 1000$, $N_d = 10$, $n_t = 50$, $n_d = 5$ and $\mu = 0.0004$.

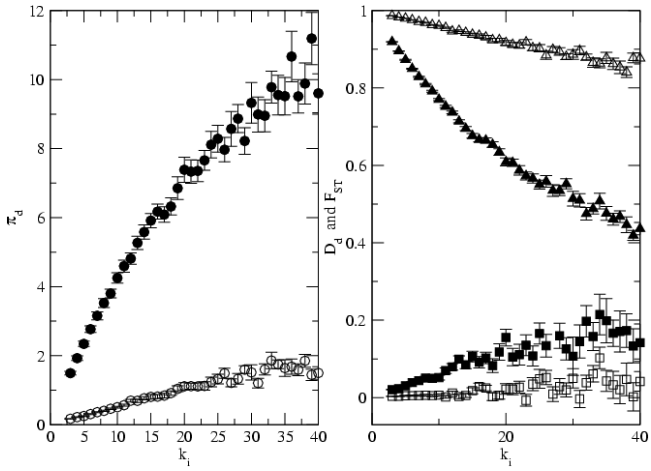

Figure 7
Within host diversity and differentiation among hosts. The level of within host diversity, π_d (in circles), and differentiation among hosts, F_{ST} (in triangles), as a function of connectivity k_i. The squares represent the mean values of Tajima's D within hosts, D_d. R_0 = 3 in open symbols and R_0 = 15 in filled symbols, other parameters are D = 1000, N_d = 10, e = 0.01, n_t = 50, n_d = 5 and μ = 0.0004.

between host level is a very important problem. The means towards its solution requires the integration of population genetics and epidemiology. This has recently been recognized as a major step for understanding infectious agent evolution [5].

Here we have studied levels and patterns of infectious agent diversity under one of the simplest classical epidemiological models: the SIS model. In this model, hosts that are susceptible can become infected at a given rate, and hosts that are infected can become susceptible by clearance of the infectious agent. We have found that, under this model and in the conditions studied, for low clearance rates and low intrahost effective population size, levels of genetic variability in samples from the whole infectious agent population are maximal for intermediate levels of transmission. This pattern of DNA sequence diversity was found to be independent of the type of host contact structure.

Although we have not performed simulations with values of N_d close to those that have been estimated for some infectious agent ($N_d \simeq 1000$ estimated for HIV-1 [48]) due to the high computational cost, from the simulations we have done we have checked that when the rate at which the immune system clears the infectious agent (e) is higher than the rate of drift ($1/Nd$) within the host, levels of infectious agent diversity in the whole metapopulation monotonically increase with R_0.

In highly transmitted infectious agents, levels of diversity are weakly dependent on the type of host contact structure. However for infectious agents with low values of R_0, levels of diversity do depend on the host contact structure: when interactions between hosts are such that every host is in contact with every other, levels of diversity are higher than when the host contact structure is such that a few hosts have a disproportionate number of contacts, whereas the majority has a small number of contacts. In this latter case levels of infectious agent diversity are expected to be low. Furthermore, in this latter case the frequency spectrum of neutral mutations can be distorted, in relation to that expected for the standard neutral model [39]. This feature is captured by negative values of the Tajima's D statistics. The observation of positive values of D_t in infections agent genes suggests that strong diversifying selection could be occurring, since even when we account for the complex contact structure in which infectious agents evolve, under a neutral model one would expect to observe values of D_t close to 0 or negative.

The results presented here can also be used to make some predictions about future adaptation in infectious agents. If we assume that new adaptive mutations in infectious agents arise from standing neutral variation [49,50], Figures 2, 3 and 4 imply that for infectious agents with low intrahost effective population size, those with intermediate R_0 will be likely to adapt more rapidly than those with larger R_0. For infectious agents in which these conditions are met, an important implication regarding public health measures can be drawn: if control programs with the aim of lowering transmission do not reduce R_0 to very low values, but instead only lead to small reductions in R_0, then this may imply an increased chance of the infectious agent escaping the immune system.

One feature of several natural populations, including infectious agent populations is the occurrence of correlations between genetic and geographical distance [14,51]. In the island model of population structure that pattern does not arise, whereas in the stepping stone model it is evident. We have explored the relation between genetic and geographical distance in the scale free contact network, which is likely to be closer to the relevant contact structure for infectious agent evolution. Although in our models we have not considered geography explicitly, we have assumed that it can be related to the shortest path length between nodes in the network. Figure 8 shows a clear correlation between these distances. One can intuitively suspect that natural selection can cause infectious agents to adapt to local conditions and that local adaptation can lead to spatial genetic structuring. But before one jumps to the conclusion that natural selection is playing a role in spatially structuring diversity one has to rule out the simpler explanation of neutral evolution in a complex

host contact network. Hopefully, the careful consideration of all diversity measures and the use of several test statistics will help us to find the molecular signature of adaptation in infectious agent gene sequences.

Methods

General model description

We consider the evolution of a haploid non-recombining population subdivided into small subpopulations-demes. There are D demes, each corresponding to a node in the network comprising all the population. Each deme has a maximum size of N_d individuals. The total maximum number of individuals in the metapopulation is $N_t = DN_d$. Each deme can go extinct with probability e and be recolonized through migration of individuals from other demes to which the deme is connected to (see below). Note that in our model recolonization occurs through migration (which is different from other metapopulation models [14]). In order to model migration we do the following. Each deme i of a given network is connected to k_i other demes according to the specific type of contact network considered. Each edge of the network connects two demes that exchange migrants at a mean rate m. We produce a new generation of individuals by taking the following steps: we draw the number of migrants going out from each deme from a Poisson distribution with mean $N_d m k_i$, if the deme is not empty. The individuals that migrate are sampled at random, without replacement, from the original deme and added to the recipient demes. The assumption of sampling without replacement, is not restrictive, since we obtain the same results in simulations where sampling with replacement is considered. The relevant parameter of the SIS model is the basic reproductive number R_0, which corresponds to $R_0 = N_d mk/e$ in our model. So in our simulations we changed the value of R_0 by changing the migration rate m while keeping constant all other quantities. After migration, reproduction and mutation occurs. N_d individuals are chosen at random to form the new population of each deme. Each individual is subject to new mutations following a Poisson distribution with mean μ. We assume the infinite sites mutational model where every new mutation occurs at a new site. At the start of each simulation run, all demes have N_d individuals, which are mutation-free and are represented by an infinitely large sequence. We then let the simulation run for an initial period, T_{eq}, to allow the metapopulation to reach an epidemiological and genetic equilibrium. The time to reach equilibrium depends on the set of parameters of the simulation. Since all the measurements are obtained after equilibrium, the results do not depend on the initial condition. Every $T = 5000$ generations, after the initial T_eq generations, we take a sample of size $n_t = 50$ from the entire population, and samples from within each deme of size $n_d = 5$, unless stated otherwise. We then calculate the average number of pairwise differences for the entire population:

$$\pi_t = \frac{\sum_{i<j} \pi_{ij}}{n_t(n_t - 1)/2} \tag{4}$$

where π_{ij} is the number of differences between two sampled sequences, and also for each deme (π_d).

We also calculate the number of segregating sites in each sample (S_t and S_d) and the test statistic Tajima's D [42] which for samples of the entire population is given by:

$$D_t = \frac{\pi_t - S_t / a_{n_t}}{b_{n_t}} \tag{5}$$

where $a_n = \sum_{i=1,n} \frac{1}{i}$, $b_n = e_1 S + e_2 S(S - 1)$ and e_1 e_2 as defined by Tajima [42].

One other quantity of interest that we have studied is F_{ST}, a measure of genetic differentiation amongst demes. This measure is defined as [46]:

$$F_{ST} = \frac{\pi_t - \pi_d}{\pi_t} \tag{6}$$

A well studied topology in the population genetics literature is the island model, introduced by Wright, which corresponds to a fully connected network where every deme

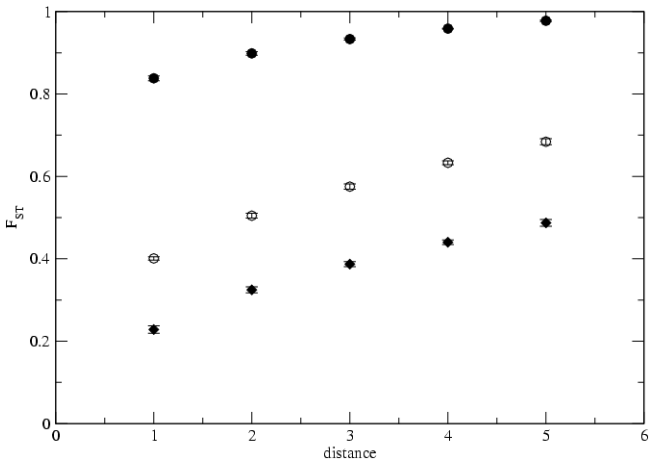

Figure 8

Level of differentiation between hosts. The level of differentiation between pairs of hosts, F_{ST}, as a function of their topological distance (which is estimated as the minimum number of links which separates two distinct demes). A scale-free network, with $\gamma = 3$, is considered with $D = 500$, $N_d = 10$, $\mu = 0.0008$, $e = 0.01$. $R_0 = 3$ in filled circle symbols, $R_0 = 30$ in empty circle symbols and $R_0 = 60$ in diamonds.

is connected to the others, so $k_i = D - 1$. A commonly studied topology in epidemiology is the scale-free network, where the distribution of connectivities obeys a power-law: $P(k_i) \propto k_i^{-\gamma}$. In real systems the exponent γ is in the range between 2 and 3. Nodes of low connectivity are predominant in the network, whereas well-connected nodes are rare. One of the mechanisms that can lead to the occurrence of a network with a power-law degree distribution is growth with preferential attachment, where nodes newly introduced to the network are preferentially attached to those nodes which are already well connected. We use the standard algorithm by Albert and Barabasi to build up the scale-free networks [21], and so we generate networks with exponent $\gamma = 3$. Scale free networks, that are extremely heterogeneous, may be appropriate descriptions for studying sexually transmitted diseases [18]. Our results for scale-free networks were compared to the island model. For every network and every parameter set we have run 30 independent simulations.

Authors' contributions

The authors contributed equally to this work.

Acknowledgements

We thank Gabriela Gomes, David Conway and Gareth Weedall for helpful suggestions. This work was supported by project POCTI/BSE/46856/2002 through Fund. para a Ciência e Tecnologia (FCT). I.G. is supported by FCT/FEDER fellowship. PRAC is partially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

1. Connway D, Cavanagh D, Tanabe K, Roper C, Mikes Z, Sakihama N, Bojang K, Oduola A, Kremsner P, Arnot D, Greenwood B, McBride J: **A principal target of human immunity to malaria identified by molecular population genetics and immunological analyses.** *Nature Medicine* 2000, **6:**689-692.
2. Falush D, Wirth T, Linz B, Pritchard J, Stephens M, Kidd M, Blaser M, Graham D, Vacher S, Perez-Perez G, Yamaoka Y, Megraud F, Otto K, Reichard U, Katzowitsch E, Wang X, Achtman M, Suerbaum S: **Traces of human migrations in Helicobacter pylori populations.** *Science* 2003, **299:**1582-1585.
3. McDonald B, Linde C: **Pathogen population genetics, evolutionary potential, and durable resistance.** *Annu Rev Phytopathol* 2002, **40:**349-379.
4. Paterson S, Viney M: **The interface between epidemiology and population genetics.** *Parasitology Today* 2000, **16:**528-532.
5. Grenfell B, Pybus O, Gog J, Wood J, Daly J, Mumford J, Holmes E: **Unifying the epidemiological and evolutionary dynamics of pathogens.** *Science* 2004, **303:**327-331.
6. Wilson D, Falush D, McVean G: **Germs, genomes and genealogies.** *Trends in Ecology and Evolution* 2005, **420(1):**39-45.
7. Galvani A: **Epidemiology meets evolutionary ecology.** *Trends in Ecology and Evolution* 2003, **18:**132-139.
8. Kreitman M: **Methods to detect selection in populations with applications to the human.** *Annu Rev Genomics Hum Genet* 2000, **1:**539-559.
9. Przeworski M, Hudson R, Di Rienzo A: **Adjusting the focus on human variation.** *Trends in Genetics* 2000, **16:**296-302.
10. Levins R: *Evolution in changing environments* Princeton, NJ: Princeton University Press; 1968.
11. Levins R: **Some demographic and genetic consequences of environmental heterogeneity for biological control.** *Bull Entomol Soc Am* 1969, **15:**237-240.
12. Hansky I: **Metapopulation Dynamics.** *Nature* 1998, **396:**41-49.
13. Pannell J, Charlesworth B: **Neutral genetics diversity in a metapopulation with recurrent local extinction and recolonization.** *Evolution* 1999, **53:**664-676.
14. Pannell J, Charlesworth B: **Effects of metapopulation processes on measures of genetic diversity.** *Phil Trans R Soc Lond B* 2000, **355:**1851-1864.
15. Rousset F: *Genetic Structure and Selection in Subdivided Populations* Princeton, NJ: Princeton University Press; 2004.
16. Charlesworth B, Charlesworth D, Barton N: **The effects of genetic and geographic structure on neutral variation.** *Annu Rev Ecol Evol Syst* 2003, **23:**99-125.
17. Whitlock M, McCauley D: **Indirect measures of gene flow and migration: F_ST not equal $1/(4Nm + 1)$.** *Heredity* 1999, **82:**117-125.
18. Lloyd A, May R: **How viruses apread among computers and people.** *Science* 2001, **292:**1316-1317.
19. Erdös P, Rényi A: **On the evolution of random graphs.** *Inst Hung Acad Sci* 1960, **5:**17-61.
20. Newman M: **The structure and function of complex networks.** *SIAM* 2003, **45:**167-256.
21. Albert R, Barabási AL: **Statistical mechanics of complex networks.** *Rev Mod Phys* 2002, **74:**47-97.
22. Albert R, Jeong H, Barabási AL: **Diameter of the world-wide web.** *Nature* 1999, **401:**130-131.
23. Liljeros F, Edling C, Amaral L, Stanley H, Aberg Y: **The web of human sexual contact.** *Nature* 2001, **401:**907-908.
24. Camazine S, Deneubourg JL, Franks N, Sneyd J, Theraulaz G, Bonabeau E: *Self-Organizations in Biological Systems* Princeton, NJ: Princeton University Press; 2001.
25. Fewell J: **Social insect networks.** *Science* 2003, **301:**1867-1870.
26. Pastor-Satorras R, Vespignani A: **Epidemic spreading in scale-free networks.** *Phys Rev Lett* 2001, **86:**3200-3204.
27. Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani A: **Velocity and Hierarchical Spread of Epidemic Outbreaks in Scale-Free Networks.** *Phys Rev Lett* 2004, **92(17):**178701.
28. Keeling M, Eames K: **Networks and epidemic models.** *J R Soc Interface* 2005, **22(2):**295-307.
29. May R: **Network structure and the biology of populations.** *Trends in ecology and Evolution* 2006, **21:**394-399.
30. Campos P, Combadao J, Dionisio F, Gordo I: **Muller's ratchet in random graphs and scale-free networks.** *Phys Rev E* 2006, **74(4 Pt 1):**042901.
31. Woolhouse M, Webster J, Domingo E, Charlesworth B, Levin B: **Biological and biomedical implications of the co-evolution of pathogens and their hosts.** *Nature Genetics* 2002, **32:**569-77.
32. Otto S, Nuismer S: **Species interactions and the evolution of sex.** *Science* 2004, **304:**1018-20.
33. Nuismer S, Otto S: **Host-parasite interactions and the evolution of ploidy.** *Proc Natl Acad Sci USA* 2004, **101:**11036-9.
34. Nuismer S, Otto S: **Host-parasite interactions and the evolution of gene expression.** *PLoS Biol* 2005, **3:**1283-8.
35. Kopp M, Gavrilets S: **Multilocus genetics and the coevolution of quantitative traits.** *Evolution* 2006, **60:**1321-36.
36. May R, Anderson R: **The transmission dynamics of human immunodeficiency virus (HIV).** *Philos Trans R Soc London B* 1988, **321:**565-607.
37. Slatkin M: **Gene flow and genetic drift in a species subject to frequent local extinction.** *Theor Popul Biol* 1997, **12:**253-262.
38. Whitlock M, Barton N: **The effective size of a subdivided population.** *Genetics* 1997, **146:**427-441.
39. Kimura M: *The neutral theory of molecular evolution* Princeton, NJ: Princeton University Press; 1983.
40. Shriner D: **Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection.** *Genetics* 2004, **166:**1155-1164.
41. Polley S, Chokejindachai W, Conway D: **Allele frequency based analyses robustly identify sites under balancing selection in a malaria vaccine candidate antigen.** *Genetics* 2003, **165:**555-561.
42. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585-595.
43. Wakeley J, Aliacar N: **Gene genealogies in a metapopulation.** *Genetics* 2001, **159:**893-905.

44. Pannell J: **Coalescence in a metapopulation with recurrent local extinction and recolonization.** *Evolution* 2003, **57:**949-961.
45. Drake J: **The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes.** *Ann N Y Acad Sci* 1999, **870:**100-107.
46. Charlesworth B: **Measures of divergence between populations and the effect of forces that reduce variability.** *Mol Biol Evol* 1998, **15:**538-43.
47. Holmes E: **Patterns of intra and interhost nonsynonymous variation reveal strong purifying selection in dengue virus.** *J Virol* 2003, **77:**11296-8.
48. Shriner D, Liu Y, Nickle D, Mullins J: **Evolution of intrahost HIV-1 genetic diversity during chronic infection.** *Evolution* 2006, **60(6):**1165-1176.
49. Przeworski M, Coop G, JD W: **The signature of positive selection on standing genetic variation.** *Evolution* 2005, **59:**. 2312-232355-561
50. Hermisson J, Pennings P: **Soft sweeps: molecular population genetics of adaptation from standing genetic variation.** *Genetics* 2005, **169:**2335-2352.
51. Real L, Henderson J, Biek R, Snaman J, Jack T, Childs J, Stahl E, Waller L, Tinline R, Nadin-Davis S: **Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus.** *Proc Natl Acad Sci USA* 2005, **102:**12107-12111.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral