Analysis of Multi-Strain Pathogens – Species or Minor Genetic Variants?

Kung-Sik Chan*
Department of Statistics & Actuarial Science
University of Iowa
Iowa City, Iowa 52242-1409, U.S.A.

Michael Kosoy
Division of Vector Borne Infectious Diseases
Centers of Disease Control and Prevention
Fort Collins, Colorado 80521, U.S.A.

February 25, 2008

Abstract

Modern advances in genetic analysis have made it feasible to ascertain the variant type of a pathogen infecting a host. Classification of pathogen variant is commonly performed by clustering analysis of the observed genetic divergence between the variants. A natural question arises as to whether the genetically distinct variants are epidemiologically distinct. A broader question is whether the different variants constitute separate microbial species or represent minor variations of the same species. We developed new statistical methodologies for addressing these important issues, in the context of classifying genetically distinct variants of bartonella bacteria found in a rodent species based on marked capture-recapture trapping data of a rodent

*Author for correspondence (kung-sik-chan@uiowa.edu)

population. We developed new statistical models for grouping variants of bartonella with mixed infections caused by several genetically distinct variants of the pathogen. In particular, we carried out a frequency analysis of co-infection patterns and a Markov chain analysis of panels of successive mixed-infection time series to test whether some particular grouping of the bartonella variants within a particular rodent species is consistent with a panel of observed disease data from a rodent population. The newly developed methodologies are broadly applicable for analyzing other multi-strain pathogens, data of which are increasingly collected for diverse infectious diseases with recent advance in genetic analysis of bacteria and viruses.

Keywords: Bartonella, cross-immunity, mixed-infection, multi-strain epidemic model, pathogen population, species coexistence, statistical independence, Markov chain.

1 Introduction

With the recent advance in genetics analysis, it has been found that a disease-causing microbe may admit multiple variants, and that it is feasible to identify which variants infect a subject (Tibayrenc & Ayala 2000). The existence of multiple strains of a pathogen can alter host-microbe interactions and might have interesting implications on the epidemiological dynamics of an infectious disease (Read & Taylor 2001), Classification of pathogen variants commonly performed by clustering analysis of the observed genetic divergence between the variants (Holmes et al. 1995). However, a fundamental question arises on whether the genetically distinguished variants correspond to ecologically and epidemiologically distinct variants. A related question is whether a large genetic distance corresponds to the emergence of separate pathogen species. Knowledge of the genetical structure and relations between genetic variants is important to understand and predict the responses of pathogen populations to selective pressures imposed by host immunity (Levin et al. 1999).

The preceding questions may be studied by a detailed analysis of the epidemi-

ological structure of a multi-strain system with real data. However, the modeling will be quite involved if we desire to disentangle the epidemiological interactions of multiple strains. Some simplification is essential for reducing the modeling complexity by first using simple methods to explore the epidemiological nature of the genetically defined variants. Here, we aim at introducing some novel statistical techniques for such exploratory analysis. In particular, we consider the interesting question of whether strains from genetically distant clusters can be regarded as belonging to distinct taxonomic groups of the pathogen in the sense that there is little or no cross-immunity between strains from distant clusters. A related issue is whether strains within a cluster are minor variants of each other in the sense of existence of cross-immunity between these strains. We propose two statistical techniques for studying the above two hypotheses, and illustrate the methodologies with a set of marked-capture-recapture data collected from a multi-strain system on the prevalence of bartonella infections among cotton rats at a study site in Georgia, USA (Kosoy et al. 2004a, b).

An outline of the rest of the paper follows. In Section 2, we briefly summarize the monitoring program from which the bartonella data was collected. The blood sample of an infected host may contain a single strain of bartonella, or it may contain multiple strains. The latter case means that the host has mixed infection. Some biological hypotheses underlying mixed infection is discussed in Section 3. Mixed-infection data furnish an opportunity to assess the hypothesis of no cross-immunity between two strains of pathogen. In Section 4, we propose a frequency analysis method to assess the no-cross-immunity hypothesis. For the bartonella data, some rats were trapped repeatedly, yielding data on their mixed-infection histories. The degree of cross-immunity may also be investigated by studying the dynamical pattern of mixed infections. In Section 5, we propose a Markov-chain technique to assess the cross-immunity structure of multi-strain system, and illustrate the method with the bartonella data. Throughout the paper, we develop the new methods in the context of analyzing the bartonella data. However, the proposed methods can be equally applicable to other pathogens. We conclude briefly in Section 6

2 The Bartonella Data

Kosoy et al. (2004 a,b) studied population dynamics of diverse Bartonella infections among cotton rats in Georgia, United States and found extremely high prevalence rate of the infection. Several variants of bartonella were circulating in the cotton rate system, and co-infection of up to three variants was reported. The bartonella variants were initially classified into three genogroups A, B and C, within each of which further variants A1-A5, B1-B5, and C1-C2 were determined based on a cluster analysis of the genotypic variations among bartonella found in the cotton rat system, see Table 1. The trapping data were mainly collected from March, 1996 to July, 1997, with a small pilot study earlier done in 1995. The analysis reported herein focused on the trapping data from 1996 to 1997, altogether 483 trapping records. Details of the trapping protocol can be found in Kosoy et al. (2004a, b). First-time trapped rats were marked, blood sampled, and examined for presence of bartonella bacteremia and a genotype or multiple genotypes of the bartonella were recorded. Co-infections by two or more bartonella variants were commonly found, see Section 4 in which we carry out a relative frequency analysis. Marked and sampled rats were subsequently released.

Some of these marked rats, 117 of them, were trapped repeatedly and irregularly, thereby resulting in 117 unequally spaced time-series data of mixed infection patterns. The panel of time series of succession of mixed infections are analyzed via Markov chain in Section 5. Here, we aim at developing some exploratory methods for elucidating the epidemiological character of the genetically classified variants. Specifically, the statistical analysis aim to resolve the following two questions:

- Is the classification A1-A5, B1-B5 and C1-C2 justifiable from the epidemiological perspective?
- Is there a correlation between this classification and the temporal (successive) pattern of the mixed infections?

Our analysis reported below suggests that the answers to these questions are affirmative, thereby corroborating the usefulness of genetic clustering as a tool for identifying epidemiologically meaningful grouping of multi-strain pathogens. The new

tools developed herein are useful for studying other multi-strain disease-causing agents.

	A1	*A2*	*A4*	*A5*	*B2*	*B3*	*B4*	*B5*	*C1*	*C2*	*Pin*
A1	1.000	0.9704	0.9615	0.9970	0.9290	0.9290	0.9349	0.9349	0.8876	0.8905	0.9024
A2		1.000	0.9852	0.9704	0.9231	0.9231	0.9290	0.9290	0.8846	0.8876	0.8964
A4			1.000	0.9615	0.9231	0.9231	0.9231	0.9231	0.8817	0.8846	0.8935
A5				1.0000	0.9290	0.9290	0.9349	0.9349	0.8905	0.8935	0.9024
B2					1.0000	0.9970	0.9911	0.9941	0.9142	0.9172	0.9024
B3						1.0000	0.9911	0.9941	0.9142	0.9172	0.9024
B4							1.0000	0.9970	0.9142	0.9172	0.9112
B5								1.0000	0.9142	0.9172	0.9083
C1									1.0000	0.9970	0.8876
C2										1.0000	0.8905
Pin											1.0000

Table 1: Genetic Similarities Among the Unique Genogroups/Variants Identified by *GLT*A Gene Sequence Analysis Based on 661 Isolates of *Bartonella* From Six Species of Small Mammal. After Kosoy et al. (2004b).

3 Hypotheses for Mixed Infections

The classification of the bartonella variants into the three major genogroups, namely, A, B and C, and the further subdivision into A1-A5, B1-B5 and C1-C2 were based on cluster analysis of genetic distances between variants measured by the rate of nucleotide substitutions in selected target gene (*glt*A). Each genogroup contained from 2 to 4 variants with inter-genogroup sequence similarities ranging from 88.2% to 93.5%. Sequence similarity among variants within each genogroup ranged from 96.2% to 99.7% (Table 1). However, it is unclear whether or how the genetic differences affect the dynamics of infection between these genotypes. One hypothesis is that the three genogroups A, B and C constitute separate bacterial species and each of which admits several variants, and that there is no between-species specific immunity but some cross immunity among variants within a species. This hypothesis is partially supported by results of study of experimental infection of cotton rats with three genotypes, which suggest that cross-protection between genotypes A, B, and C may not occur (Kosoy et al., 1999). Missing informa-

tion include degree of cross-immunity between closely related variants within the same genogroup (e.g., A1 and A2). Additional study demonstrated that A, B, and C genotypes behaved differently being introduced to experimentally infected cotton rats. These differences were expressed in a wide range of variation in level of bacteremia, growth kinetics in rat blood, and minila infectious dose able to reproduce the infection (Kosoy et al., 1999, 2000). Furthermore, the mixed infection dynamics is similar to some modified version of the SIR model (Dickmann and Heesterbeek, 2000) taking into account of the competition of the bartonellae. Below, this hypothesis shall be referred to as the species-variant hypothesis.

Yet another hypothesis states that a single vector transmission may consist of a mixture of bartonella variants (e.g. A1, A2, C1) and each of them may attack a host only after a random dormant period. This hypothesis was prompted by the observation that some individual mammalian hosts were observed to have mixed infection by different bartonella variants, over time. For example, Kabeya et al. (2002) reported evidence of multiple infection of genetically different *Bartonella henselae* in the naturally infected cats. At each peak of bacteremia, genetically different variants were isolated from the blood of cats showing a relapsing bacteremia. The results obtained by Arvand et al. (2006) also suggest the populations of primary *B. henselae* isolates are commonly composed of distant genetic variants, which may disappear upon repeated passages among animals. Thus, generation of genetically independent variants may represent an escape mechanism to circumvent the host specific immune responses. This hypothesis has the implication that all variants A1-A5, B1-B5 and C1-C2 are independent variants and shall be referred below as the independent-variant hypothesis.

4 Assessing the Independent-Variant Hypothesis

If A1 and A2 infect the rats independently, then the probability of finding A1 & A2 simultaneously in a rat equals P(A1) times P(A2) where P(A1) is the probability of finding an A1 strain in the blood of a random rat, and similarly defined is

P(A2). We estimate P(A1) by the following weighted sum, namely, the sum of the relative frequency of hosts infected by A1 alone, plus 1/2 of the relative frequency of hosts infected by A1 and A2, 1/3 of that of A1A2A5, etc. This deviates from the simple scheme of estimating P(A1) by the relative frequency of hosts infected by A1 whether or not the infection is single or mixed. The use of the weighted scheme is justified by the particular protocol used for identifying the variants of the bartonella infecting a host. Specifically, if a host was found to be infected, the blood sample was repeatedly diluted in order to ascertain the bacteria load. Bacteria colonies were then cultivated from some diluted blood, which typically yielded a small number of colonies. The type of the bartonella variant(s) in the blood sample was then determined based on the morphology of the bacteria colonies. Hence, the variant identification process involved a sampling process. Consequently, the probability P(A1) should be interpreted as the probability of drawing a variant A1 bacterium in a random blood sample, in which case it should equal the sum of the probability of a single infection by A1 (because the variant A1 is then uniquely identified), plus the probability of a co-infection by A1A2 from which an A1 bacterium is drawn, plus the the probability of a co-infection A1A2A5 from which an A1 is drawn, etc., with the sum over all distinct infection types involving A1. However, the probability of finding an A1A2 infected host from which a random A1 bartonella bacterium is drawn equals the probability of finding an A1A2 co-infection times the conditional probability of drawing an A1 bacterium from such a host, which can be estimated by 1/2 times the relative frequency of the A1A2 co-infection. Similarly, the probability of finding an A1A2A5 infected host from which an A1 bacterium is drawn can be estimated by 1/3 times the relative frequency of the A1A2A5 co-infection. This completes our justification of the weighted scheme for estimating P(A1). Similarly, we can estimate the probabilities of other variants that are listed in Table 2.

variant	A1	A2	A4	A5	B2	B3	B4	B5	C1	C2
probability	0.419	0.107	0.033	0.018	0.021	0.009	0.003	0.003	0.099	0.042

Table 2: Probabilities of various bartonella strain in the blood of a host

Note that these probabilities sum to the probability that a rat is infected by some bartonella variant, and hence the sum is less than 1. We computed the theoretical probabilities of various co-infection patterns under the independent-variant hypothesis and compare them with the observed relative frequencies, see Table 3. In particular, we calculated the ratio of the theoretical probability to the observed relative frequency for various co-infection patterns. Also, we computed the bootstrap 95% intervals for the theoretical probabilities under the independent-variant hypothesis. The bootstrap was done by re-sampling the data cases with replacement, with each data case being a single trapping record.

For each co-infection pattern, we can reject the independent-variant hypothesis at 5% significance level if the 95% bootstrap confidence interval does not contain the observed relative frequency. In the case that cross immunity exists between two variants, the observed relative frequency of co-infection by the two variants is expected to be lower than the theoretical probability. On the other hand, if infection by one variant increases the chance of infection by a second variant, then the observed relative frequency of the co-infection by the two variants is expected to be higher than the theoretical probability. Thus, in the case of rejection of the independence assumption, the position of the relative frequency as compared to the theoretical probability may shed insight on the relationship between the two variants under study.

From Table 3, it can be inferred that the observed relative frequencies of co-infection by within-genogroup variants A1A2, A1A4, A1A5, A2A4 and C1C2 are smaller than the theoretical probability and lie outside the confidence interval of the theoretical probability; hence, we can reject the independence assumption for these co-infection patterns. Moreover, as the relative frequencies of these co-infection pattern are smaller than the theoretical counterparts under the independence assumption, there is some evidence that the variants within the A genogroup (and the two within the C genogroup) are subject to cross immunity. A lone counterexample to the above is co-infection B3B4 for which the independence assumption is rejected but its observed relative frequency is greater than the theoretical probability. However, owing to the rare occurrences of B3 and B4, it may be a false

co-infection	obs. freq.	theo. prob.	theo. over obs.	95% C.I. theo. prob.
A1A2	0.01863	0.04872	2.615	(0.03838, 0.05978)
A1A2A4	0.00207	0.00158	0.763	(0.00089, 0.00240)
A1A2A5	0.00207	0.00099	0.479	(0.00045, 0.00166)
A1A4	0.00414	0.01351	3.263	(0.00764, 0.01984)
A1A5	0.00207	0.00848	4.096	(0.00396, 0.01395)
A1B2	0.00414	0.00891	2.152	(0.00439, 0.01413)
A1B2C2	0.00207	0.00038	0.183	(0.00016, 0.00067)
A1B3C1	0.00207	0.00025	0.120	(0.00003, 0.00056)
A1B4	0.00207	0.00129	0.625	(0.00000, 0.00300)
A1C1	0.05176	0.04240	0.819	(0.03275, 0.05228)
A1C1C2	0.00207	0.00180	0.869	(0.00111, 0.00256)
A1C2	0.01449	0.01768	1.220	(0.01129, 0.02442)
A2A4	0.00207	0.00379	1.833	(0.00209, 0.00588)
A2A4C2	0.00207	0.00016	0.078	(0.00008, 0.00027)
A2B2	0.00207	0.00250	1.209	(0.00118, 0.00405)
A2C1	0.01242	0.01191	0.959	(0.00841, 0.01582)
A2C2	0.00828	0.00496	0.600	(0.00303, 0.00721)
A4C1	0.00621	0.00330	0.532	(0.00177, 0.00512)
B2C1	0.00207	0.00218	1.052	(0.00105, 0.00361)
B2C2	0.00414	0.00091	0.219	(0.00039, 0.00162)
B3B4	0.00207	0.00002	0.009	(0.00000, 0.00007)
B4C2	0.00207	0.00013	0.064	(0.00000, 0.00032)
B5C1	0.00207	0.00032	0.153	(0.00000, 0.00086)
C1C2	0.00207	0.00432	2.087	(0.00263, 0.00626)

Table 3: Observed relative frequencies of various co-infections and the corresponding theoretical probabilities calculated under the independence assumption. The last column lists the 95% bootstrap confidence interval of the theoretical probabilities, based on 5000 bootstrap replications. Positively (negatively) dependent mixed infection types are boldfaced (underlined).

alarm. Co-infection patterns by variants of different genogroups, including A1B4, A1C1, A1C1C2, A1C2, A2B2, A2B2, A2C1 and B2C1, are found to be consistent with the independence assumption. On the other hand, the observed relative frequencies of co-infections A2C2, A4C1, B2C2, B4C2, B5C1, A1B2C2, A1B3C1 and A2A4C2 are all greater than their theoretical counterparts under independence assumption, and furthermore lie outside the 95% confidence intervals. Thus, we reject the independence assumption for these co-infection patterns, but now for the possible reason that infection by a bartonella variant increases the chance of being infected by another variant from a different genogroup, perhaps because the immune system of the host is weakened by an infection of an independent bartonella species.

In summary, co-infections by variants within the same genogroup tend to have lower relative frequencies than the theoretical probabilities assuming these within-group variants are independent species. The smaller relative frequencies of the within-group co-infections suggests the presence of cross immunity between the within-group variants. On the other hand, co-infections by between-group variants tend to have relative frequencies similar to the theoretical probabilities or higher, with A1B2 being a lone exception, thereby suggesting that between-group variants are independent species and that infection by one group may slightly increase the chance of being infected by an independent species as the immune system of the host may be weakened by an on-going infection.

Above, the analysis is based on estimating the probability of a specific bartonella variant by the weighted scheme described in the beginning of this section. We have also repeated the above frequency analysis with such probabilities estimated by a non-weighted scheme, i.e., P(A1) is estimated by the relative frequency of hosts with a single or mixed infection by A1, etc. See Tables S1–S3 in the Appendix. Broadly speaking, the analysis based on the non-weighted estimation scheme yields less clear though generally similar conclusions as those inferred from the weighted scheme reported herein.

5 Assessing the Species-Variant Hypothesis by Markov Chain Analysis

In the previous section, we found strong evidence that the bartonella variants circulated in the rat system in Georgia are unlikely to be independent variants. Indeed, it seems to cast some support to the hypothesis of classifying the A, B and C genogroups as independent species with the variants within each group enjoying some cross immunity. But the preceding analysis is based on the frequencies of co-infections by various variants. Here, we aim to study the same problem by a temporal analysis of the mixed infections. The key idea is that the species-variant hypothesis implies some correlation structure for infections by the bartonella variants that may have some observable implications on the successive patterns of mixed bartonella infections. An example of the monthly mixed infection pattern for a rat trapped multiple times was A1, no bartonella detected, B2, A2, not trapped, no bartonella detected, see Figure 1 for other observed time series patterns. Cross immunity between variants from a species may imply that an infection is more likely to be followed by another infection from a different genogroup than from the same group, after adjusting for their epidemiological characteristics (infectivity, transmission rate and susceptibility). For example, an A1 infection may be more likely followed by a B1 than an A2, given everything else being equal.

The transition of the mixed infection pattern can be studied via a Markov chain analysis of the monthly disease status of a random host. Analysis was done using a subset of the cotton-rat data where a rat was multiply trapped, resulting in 117 time-series data on the succession pattern of mixed infections. So, only the following variants are observed: A1, A2, A4, A5, B2, B3, B4, B5, C1 and C2. However, there are two complications to this approach. First, a rat may have co-infections, e.g. A1A2, in a certain month. This necessitates enlarging the state space of the Markov chain to include all observable co-infection patterns; altogether there are 34 states for the Markov chain. Second, the trapping dataset has, naturally, many missing data, as the same rat would seldom be trapped every month. Fortunately, maximum likelihood estimation of the transition probability

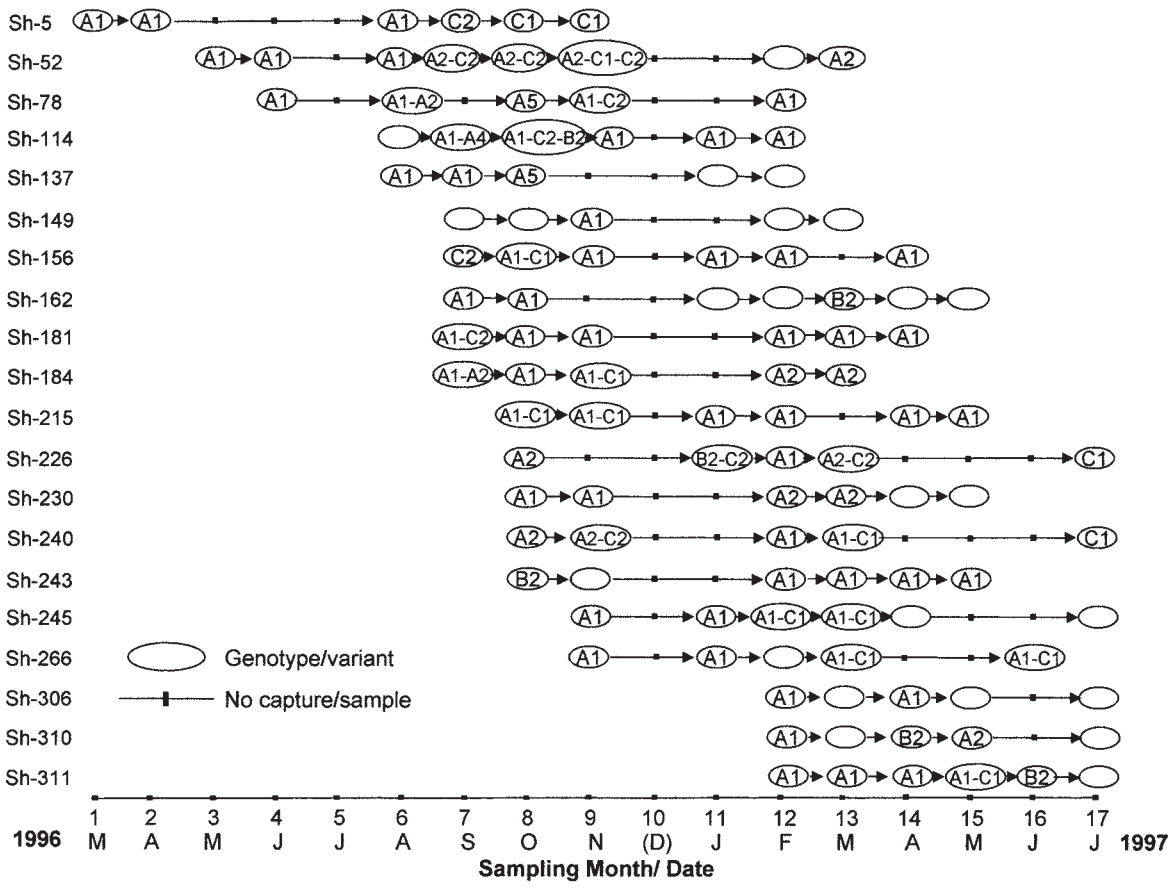

Figure 1: Re-sampling history, infection course, and genotypic characterization of sequentially recovered Bartonella isolates from 20 cotton rats captured $\geq$ 5 times during 16 months of trapping (December 1996 was not trapped). The genogroups/variants of Bartonella recovered from bacteremic rats at a sample month are shown within ovals; the first oval indicates the date an individual was first recruited into the study cohort. Blank ovals indicate no detectable bacteremia at a sample month and the tick marks indicate an individual was not recaptured during that trapping session (no trapping was performed in December 1996). After Kosoy et al (2004b).

matrix of the Markov chain with extensive missing data can be carried out by the EM algorithm (Dempster, Laird and Rubin, 1977). See Rabiner (1989) for details.

As mentioned earlier, owing to the presence of co-infections, the states of the Markov chain include also A1A2, etc. altogether 34 (observed) states! It is clearly not revealing to report the estimated 34 by 34 transition probability matrix as it is hard to comprehend such a huge matrix. But the main issue concerns the succession frequencies of infections by same genogroup or by a different genogroup. From this perspective, the estimated transition probability matrix can be used to provide such information as listed in Table 4. For example, the first row in this table gives various conditional probabilities given an infection by A1 in this month. The conditional probability that given a rat is infected by A1 bartonella variant in the current month, the same infection is maintained, i.e. no change in bartonella infection type, in the next month is 0.6254. The probability that given an A1 infection in the current month, the rat continues to be infected by A1 but also acquires another variant from the same genogroup in the next month is 0.0447. The probability that given an A1 infection in the current month, the rat continues to be infected by an A1 but also infected by a variant from another genogroup in the next month is 0.0983. The probability that given an A1 infection in the current month, the rat is no longer infected by an A1 in the next month but has an infection by another variant of the same genogroup is 0.08. The probability that given an A1 infection in the current month, the rat is no longer infected by an A1 in the next month but has an infection by another variant of a different genogroup is 0.0418. The probability that given an A1 infection in the current month, the rat has no detectable bartonella in the next month is 0.1098. Other rows give similar conditional probabilities given an infection by other bartonella variant in the current month.

The information can be further summarized in the last row of the table as the conditional probabilities given an infection in the current month. These probabilities are normalized weighted column sums with the weight of each row equal to the probability of infection by the variant, labeling that row, at the current month, and then the weighted column sums are renormalized to make them sum

to 1. The numbers in the last row of Table 4 have the following interpretations: Given that a rat is infected by some bartonella variant in the current month, the conditional probability that the rat continues to be infected by the same variant, i.e. unchanged bartonella infection type, equals 0.464, that the rat maintains the same variant and at the same time acquires another variant of the same genogroup in the next month is 0.0306, that the rat maintains the same variant and acquires another variant of a different genogroup in the next month is 0.0761. Thus, the difference of these two probabilities equals $0.0761 - 0.0306 = 0.0455$. We have also computed a 95% bootstrap confidence interval for the difference. In the bootstrap, the whole time series of bartonella infection type (or lack of infection) for each rat forms a unit, and we bootstrap these time-series units by randomly sampling the panel of time series of infection type with replacement. For each bootstrap panel of time series, we estimate the conditional probabilities corresponding to the last row of Table 4. Based on 500 bootstrap replications and Efron's percentile method (Efron and Tibshirani 1994), the bootstrap 95% confidence interval of the difference is $(0.0197, 0.944)$, suggesting that the conditional probability of acquiring another variant from a different genogroup is significantly higher than that from the same genogroup, at 5% significance level.

The conditional probability that given a bartonella infection in the current month, the rat is no longer infected by the variant but acquires another variant of the same genogroup in the next month is 0.114, and that the rat is no longer infected by the variant but acquires another variant of a different genogroup in the next month is 0.158. Note that the difference of the two probabilities equals $0.158 - 0.114 = 0.044$, with the corresponding 95% bootstrap confidence interval being $(0.00979, 0.122)$, again suggesting that the conditional probability of being replaced by a variant of a different genogroup is significantly higher than that of the same genogroup. Finally, the conditional probability that given a bartonella infection in the current month, the probability that the rat has no detectable bartonella variant in the next month is 0.157.

Altogether, these results strongly suggest that a bartonella infection is less likely to be followed by an infection by another variant of the same genogroup

	maintain	acq. same	acq. diff	repl. same	repl. diff	undetected
A1	0.6254	0.0447	0.0983	0.0800	0.0418	0.1098
A2	0.4042	0.0435	0.0000	0.2082	0.2166	0.1275
A4	0.0000	0.0000	0.1597	0.5679	0.0000	0.2723
A5	0.2097	0.0000	0.0000	0.0000	0.1875	0.6028
B2	0.0000	0.0000	0.0000	0.0000	0.3266	0.6734
B3	0.2085	0.0000	0.0000	0.0001	0.0000	0.7914
B5	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000
C1	0.4246	0.0000	0.0582	0.0012	0.3476	0.1685
C2	0.0000	0.0000	0.1092	0.2434	0.6474	0.0000
Given an infection	0.464	0.0306	0.0761	0.114	0.158	0.157

Table 4: Conditional probabilities of various successive infection patterns in the next month

than by one of a different genogroup. This finding is consistent with the species-variant hypothesis that the A, B and C genogroups are independent sub-species and variants of each genogroup enjoy cross immunity to some degree, whereas variants of different groups are largely independent.

6 Conclusion

Based on a frequency analysis of co-infections by various bartonella variants, there is some strong evidence against the hypothesis that all bartonella variants are independent species. On the other hand the results of the analysis is consistent with the hypothesis that the A, B and C genogroups function as independent species but the variants within each genogroup enjoy some cross immunity against each other. There is also some evidence that while the three genogroups are largely independent species, infection by one genogroup may weaken the host immunity which promotes infection by another genogroup.

A second analysis of the panel of time series of bartonella infection history for cotton rats that were trapped repeatedly yields results consistent with the co-

infection frequency analysis. Specifically, an infection is found to be more likely followed by another infection by another variant from a different genogroup than from the same genogroup. These analyses favor the species-variant hypothesis that the three genogroups A, B and C circulating among the rat system in Georgia (U.S.A.) are more or less independent species, which explains the high prevalence rate of bartonella infection observed in the studied rat system.

Some interesting future work consists of fitting a modified SIR model that accounts for the cross immunity between different variants, and further assessing the various hypotheses within such a framework. Another interesting direction of research is to correlate the estimated cross immunity pattern with the known genetic distance between the variants. Finally, it is of interest to assess the stability (chaoticity) of the estimated modified SIR model and the feasibility of long-run co-existence of different variants.

KSC gratefully acknowledges partial support from the US National Science Foundation (CMG-0620789).

References

[1] Arvand, M., Schubert, H. & Viezens, J. 2006 Emergence of distinct genetic variants in the population of primary *Bartonella henselae* isolates. *Microbes and Infection* **8**, 1315-1320.

[2] Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc. B* **39**, 1-38.

[3] Dickmann, O. & Heesterbeek, J. A. P. 2000 *Mathematical Epidemiology of Infectious Disease: Model Building, Analysis and Interpretation.* Chichester: John Wiley.

[4] Efron, B. & Tibshirani, R.J. 1994 *An Introduction to the Bootstrap.* Chapman and Hall.

[5] Holmes, E.C., Nee,S., Rambaut, A., Garnett, G. P., & Harvey, P. H. 1995 Revealing the history of infectious disease epidemics through phylogenetic trees. *Phil. Trans. R. Soc. Lond.* **B349**, 33-40.

[6] Kabeya, H., Maruyama, S., Irei, M., Takahashi, R., Yamashita, M. & Mikami, T. 2002 Genomic variations among Bartonella henselae isolates derived from naturally infected cats. *Vet. Microbiol.* **89**, 211-221.

[7] Kosoy, M. Y., Regnery, R. L., Kosaya, O. I. & Childs, J. E. 1999. Experimental infection of cotton rats with three naturally occurring Bartonella species. *J. Wildlife Dis.* **35**, 275-284.

[8] Kosoy, M. Y., Saito, E. K., Green, D., Marston, E. L., Jones, D. C. & Childs, J. E. 2000 Experimental evidence of host specificity of Bartonella infection in rodents. *Comp. Immunol. Microbiol. & Infect. Dis.* **23**, 221-238.

[9] Kosoy, M. Y., Mandel, E. L., Green, D. C., Marston, E. L., Jones, D. C. & Childs, J. E. 2004a Prospective studies of Bartonella of rodents. Part I. Demographic and temporal patterns in population dynamics. *Vector-Borne and Zoonotic Dis.* **4**, 285-295.

[10] Kosoy, M. Y., Mandel, E. L., Green, D. C., Marston, E. L., Jones, D. C. & Childs, J. E. 2004b Prospective studies of Bartonella of rodents. Part II. Diverse infections in a single rodent community. *Vector-Borne and Zoonotic Dis.* **4**, 296-305.

[11] Levin, B. R., Lipsitch, M., & Bonhoeffer, S. 1999 Population biology, evolution, and infectious disease: convergence and synthesis. *Science* **283**, 806-809.

[12] Rabiner, L. R. 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257-286.

[13] Read, A. F. & Taylor, L. H. 2001 The ecology of genetically diverse infections. *Science* **292**, 1099-1102.

[14] Tibayrenc, M. & Ayala, F. 2000 Molecular epidemiology and evolutionary genetics of pathogenic microorganisms: analysis and interpretation of data.

In: *Molecular Epidemiology of Infectious Diseases.* Ed. R.C.A. Thompson, Arnold, New York, 20-29.

Appendix

Here, we report the frequency analysis of the co-infection patterns, with the probability of a specific bartonella variant estimated by a non-weighted scheme. For example, P(A1) is estimated by the relative frequency of hosts with single or mixed infection by A1. Table S1 lists the estimated probabilities of the various bartonella strains which are, of course, all greater than their counterparts in Table 2. We argued in the main text that the non-weighted estimation scheme is incorrect. Nonetheless, it is interesting to repeat the analysis and contrast the findings from the two schemes. Table S2 reports the frequency analysis of the co-infection patterns with the theoretical probabilities computed based on the non-weighted scheme. The differences in the testing results between the weighted and the non-weighted scheme are summarized in Table S3. Note that the non-weighted scheme leads to changing A1A2A5, A2C2 and A4C1 from being classified as positively dependent to independent, as well as changing A1C1C2, A1C2 and B2C1 from independent to negatively dependent; otherwise the two schemes yield identical test results. Thus, we conclude that, consistent with our earlier argument for the use of the weighted scheme, the non-weighted scheme yields a less clear inference although with broadly similar conclusions as those of the wighted scheme.

variant	A1	A2	A4	A5	B2	B3	B4	B5	C1	C2
probability	0.470	0.130	0.041	0.020	0.028	0.011	0.006	0.004	0.135	0.061

Table S1: Probabilities of various bartonella strain in the blood of a host, computed by the non-weighted scheme.

co-infection	obs. freq.	theo. prob.	theo. over obs.	95% C.I. theo. prob.
A1A2	0.01863	0.06744	3.619	(0.05353, 0.08194)
A1A2A4	0.00207	0.00279	1.349	(0.00160, 0.00419)
A1A2A5	0.00207	0.00154	0.742	(0.00070, 0.00256)
A1A4	0.00414	0.01955	4.720	(0.01146, 0.02809)
A1A5	0.00207	0.01075	5.193	(0.00497, 0.01744)
A1B2	0.00414	0.01368	3.304	(0.00711, 0.02120)
A1B2C2	0.00207	0.00085	0.410	(0.00037, 0.00150)
A1B3C1	0.00207	0.00056	0.270	(0.00012, 0.00121)
A1B4	0.00207	0.00293	1.416	(0.00000, 0.00681)
A1C1	0.05176	0.06744	1.303	(0.05229, 0.08330)
A1C1C2	0.00207	0.00419	2.023	(0.00264, 0.00599)
A1C2	0.01449	0.02932	2.023	(0.01952, 0.03995)
A2A4	0.00207	0.00592	2.857	(0.00334, 0.00903)
A2A4C2	0.00207	0.00037	0.177	(0.00018, 0.00062)
A2B2	0.00207	0.00414	2.000	(0.00206, 0.00660)
A2C1	0.01242	0.02041	1.643	(0.01479, 0.02704)
A2C2	0.00828	0.00887	1.071	(0.00554, 0.01271)
A4C1	0.00621	0.00592	0.952	(0.00330, 0.00903)
B2C1	0.00207	0.00414	2.000	(0.00210, 0.00669)
B2C2	0.00414	0.00180	0.435	(0.00079, 0.00318)
B3B4	0.00207	0.00005	0.025	(0.00000, 0.00018)
B4C2	0.00207	0.00039	0.186	(0.00000, 0.00096)
B5C1	0.00207	0.00059	0.286	(0.00000, 0.00159)
C1C2	0.00207	0.00887	4.286	(0.00558, 0.01253)

Table S2: Observed relative frequencies of various co-infections and the corresponding theoretical probabilities calculated under the independence assumption, with the probability of a specific variant estimated by the non-weighted scheme. The last column lists the 95% bootstrap confidence interval of the theoretical probabilities, based on 5000 bootstrap replications. Positively (negatively) dependent mixed infection types are boldfaced (underlined).

weighted scheme	non-weighted scheme
A1A2	A1A2
A1A2A4	A1A2A4
A1A2A5	A1A2A5
A1A4	A1A4
A1A5	A1A5
A1B2	A1B2
A1B2C2	**A1B2C2**
A1B3C1	**A1B3C1**
A1B4	A1B4
A1C1	A1C1
A1C1C2	A1C1C2
A1C2	A1C2
A2A4	A2A4
A2A4C2	**A2A4C2**
A2B2	A2B2
A2C1	A2C1
A2C2	A2C2
A4C1	A4C1
B2C1	B2C1
B2C2	**B2C2**
B3B4	**B3B4**
B4C2	**B4C2**
B5C1	**B5C1**
C1C2	C1C2

Table S3: Comparison of the testing results between the weighted and the non-weighted schemes. Positively (negatively) dependent mixed infection types are boldfaced (underlined).