LLNL-TR-420003

Bacterial Population Genetics in a Forensic Context

Developing more rigorous methods for source attribution

Stephan P. Velsko
Lawrence Livermore National Laboratory
October 30, 2009

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Executive Summary

This report addresses the recent Department of Homeland Security (DHS) call for a Phase I study to (1) assess gaps in the forensically relevant knowledge about the population genetics of eight bacterial agents of concern, (2) formulate a technical roadmap to address those gaps, and (3) identify new bioinformatics tools that would be necessary to analyze and interpret population genetic data in a forensic context. The eight organisms that were studied are *B. anthracis*, *Y. pestis*, *F. tularensis*, *Brucella spp*., *E. coli* O157/H7, *Burkholderia mallei*, *Burkholderia pseudomallei*, and *C. botulinum.*

Our study focused on the use of bacterial population genetics by forensic investigators to test hypotheses about the possible provenance of an agent that was used in a crime or act of terrorism. Just as human population genetics underpins the calculations of match probabilities for human DNA evidence, bacterial population genetics determines the level of support that microbial DNA evidence provides for or against certain well-defined hypotheses about the origins of an infecting strain.

Our key findings are:

- Bacterial population genetics is critical for answering certain types of questions in a probabilistic manner, akin (but not identical) to "match probabilities" in DNA forensics.

- A basic theoretical framework for calculating likelihood ratios or posterior probabilities for forensic hypotheses based on microbial genetic comparisons has been formulated. This "inference-on-networks" framework has deep but simple connections to the population genetics of mtDNA and Y-STRs in human DNA forensics.

- The "phylogeographic" approach to identifying microbial sources is not an adequate basis for understanding bacterial population genetics in a forensic

context, and has limited utility, even for generating "leads" with respect to strain origin.

- A collection of genotyped isolates obtained opportunistically from international locations augmented by phylogenetic representations of relatedness will not provide a useful forensic database. A more useful database for each pathogen would consist of a detailed record of human and enzootic outbreaks noted through international outbreak surveillance systems, and "representative" genetic sequences from each outbreak.

- Interpretation of genetic comparisons between an attack strain and reference strains requires a model for the network structure of maintenance foci, enzootic outbreaks, and human outbreaks of that disease, coupled with estimates of mutational rate constants. Validation of the model requires a set of sequences from exemplary outbreaks and laboratory data on mutation rates during animal passage. The necessary number of isolates in each validation set is determined by disease transmission network theory, and is based on the "network diameter" of the outbreak.

- The 8 bacteria in this study can be classified into 4 categories based on the complexity of the transmission network structure of their natural maintenance foci and their outbreaks, both enzootic and zoonotic.

- For *B. anthracis*, *Y. pestis*, *E. coli* O157, and *Brucella melitensis*, and their primary natural animal hosts, most of the fundamental parameters needed for modeling genetic change within natural host or human transmission networks have been determined or can be estimated from existing field and laboratory studies.

- For *Burkholderia mallei*, plausible approaches to transmission network models exist, but much of the fundamental parameterization does not. In addition, a

validated high-resolution typing system for characterizing genetic change within outbreaks or foci has not yet been demonstrated, although a candidate system exists.

- For *Francisella tularensis*, the increased complexity of the transmission network and unresolved questions about maintenance and transmission suggest that it will be more complex and difficult to develop useful models based on currently available data.

- For *Burkholderia pseudomallei* and *Clostridium botulinum*, the transmission and maintenance networks involve complex soil communities and metapopulations about which very little is known. It is not clear that these pathogens can be brought into the inference-on-networks framework without additional conceptual advances.

- For all 8 bacteria some combination of field studies, computational modeling, and laboratory experiments are needed to provide a useful forensic capability for bacterial genetic inference.

1. Introduction: bacterial population genetics is critical for answering certain forensic questions

President Obama is vacationing in Martha's Vineyard in the spring of 2010. Toward the end of his intended stay he becomes ill, as do several of his staff and two cabinet members who had joined him over several days. They are diagnosed with acute respiratory tularemia. Genetic sequencing and phylogenetic analysis indicate that the infecting genotype is evolutionarily close to recent isolates collected from the Squibnocket focus, which is closest to the president's compound, but surprisingly are closer to an isolate held by Tufts University that was collected from the Katama focus on Martha's Vineyard in 2003, which is more than 10 miles distant. Upon further investigation it is discovered that a former Tufts graduate student, who collected Ft isolates at the Katama site in 2003 but had since moved to the Midwest, was on Martha's Vineyard during the week prior to the President's illness. He is connected to a militant right-wing group that the FBI has had under surveillance, but he claims that he was just visiting old haunts and taking photos. Literature attributed to the group includes a pamphlet describing several ways to eliminate "enemies of Christ", which includes using plague and "rabbit fever". Several democratic senators publicly demand that the FBI prosecute the former graduate student for attempting to assassinate the president. The suspect's lawyer publicly points out that the genomic sequences of the 2003 Katama isolate and the presidential strain "do not match" – they differ by a number of mutations, and "when the DNA doesn't match, you cannot convict". He insists that the public health authorities have simply not done a thorough search to find the true source of the "tularemia virus." On the Fox news channel an expert points out that phylogenetic evidence used in HIV cases can only be used to show that two strains are "closely related", and that unlike human DNA cases, phylogenetics cannot offer a probability value for the strength of association. FBI searches include extensive sampling of the suspect's home, automobile, possessions, and places he stayed that week but no samples test positive for Ft. A northern Arizona laboratory has offered, for 25 million dollars, to sample and fully sequence "every strain of F. tularensis on Martha's Vineyard." The FBI carefully considers the offer….

- Hypothetical scenario

A recent report by the Commission on the Prevention of WMD Proliferation and Terrorism highlighted the high likelihood of an attack in the United States using a biological agent[1]. Because of the serious consequences that might follow from a decision concerning the attribution of a biological incident, and the potential danger of misconceptions among non-technical policymakers, it is important for scientists to be careful and clear about the interpretation of genotypic comparisons and any inferences drawn from them about the origin of the attack strain.

When the whole genome (consensus) sequences of any two isolates are compared, they will either be identical, or they will differ by one or more mutational changes. It is now possible through a combination of sequencing and typing protocols to ascertain with extremely high confidence that any observed mutational difference is *not* due to an error in characterization[2]. If a pathogen is used to commit a crime or act of terrorism the "source" of that pathogen is most likely to be an isolate held by a legitimate laboratory, a human or animal victim of a past outbreak of that disease, or a sample from a recognized enzootic focus. The genomic sequences of isolates from the incident and the putative source are primary evidence in a microbial forensic investigation. When the sequences are identical, it is desirable to be able to communicate how small the probability is that they are not from the same source. When they differ by one or more mutations, it may be desirable to communicate the probability that they nonetheless could be from the same source. (Alternatively, it might be desirable to state the probability that they are from different sources.) When an isolate has been associated with a source, it may be desirable to communicate how unlikely (or likely) it would be to obtain that same consensus genotype by chance in another source.

More generally, during the investigation of a bioterrorism event or biocrime, investigators use evidence to construct a narrative regarding the acquisition, production, transport and deployment of the agent. Each piece of evidence, including microbial genetic data, suggests, tests, and refines certain elements of the narrative, which can be formulated as hypotheses. Examples of hypotheses that might arise in a bioforensics investigation are:

- ☞ The bacterial strain involved in a bioterror incident was derived from one originally associated with a known past outbreak of that disease.
- ☞ The bacterial strain that infected the victims of an outbreak was derived from a known natural focus of the pathogen in a certain geographical area.
- ☞ The bacterial strain acquired certain unusual gene sequences as the result of natural lateral genetic exchange rather than by some deliberate genetic engineering process.

- ☞ The observed genetic differences between two bacterial isolates of interest are attributable to a single passage in an infected host.

In order to test the first hypothesis it is necessary to know the degree of genetic diversity associated with isolates from the outbreak in question, even when the outbreak occurred in the distant past and only a single representative reference isolate is available. Similarly, testing the second requires an estimate of the genetic diversity of the microbe in a natural maintenance population in the wild, when it is not practical to obtain more than a handful of isolates from the focus in question. The third hypothesis depends (in part) on the probability that a bacterium could find and acquire the relevant genes in its natural hosts (or in soil or water for some bacteria). The fourth depends on the probability of observing genetic change by chance or by adaptive pressures when a host becomes infected. In each case, inferences from genetic sequence data require us to know something about the structure of the population of bacteria from which the isolate in question was (hypothetically) drawn, and the probability of observing certain genetic sequences in that population. This is how the term "bacterial population genetics" is used in this report[3].

A related issue of importance is the role that databases and archives will play in future bioterror or biocrime investigations. Following an analogy with "cold hits" in human DNA forensics[4], it is tempting to consider a paradigm in which the sequence (or haplotype) of the attack strain is compared to a reference strain database hoping that observed sequence similarities will provide clues that somehow narrow the search for potential sources. When the database in question is a select agent registry whose completeness is enforced by law, such a paradigm makes sense[5]. However, its utility is restricted to excluding or including laboratories in the set of institutions compelled to provide information on strain holdings. When considering other potential sources that a terrorist or criminal might use to obtain a pathogen, this paradigm breaks down, because for practical reasons a global database of bacterial genotypes cannot be exhaustive, and will always contain a biased sampling of genetic types. Even worse, geographic associations, surmised from geographical metadata associated with strains in such a collection, will potentially confuse decision makers who give them unjustified weight.

Regardless of the structure or content of a microbial forensic database, to make any comparison of genetic sequence data useful it is necessary to have a basis for interpreting degrees of genetic similarity. Many authors have advocated a "phylogeographic" approach to interpretation[6-15]. This paradigm uses phylogenetic construction to estimate the evolutionary relationships among a set of sequences (or haplotypes) and assigns, directly or indirectly, significance to the geographic associations of clades. Whatever the merits of this approach, it provides no guidance for judging the impact of incomplete or biased sampling on source inferences made by comparisons with a database, or how one could achieve unbiased sampling even if it were practical.

Given that databases and strain collections are necessarily incomplete, bacterial population genetics plays the same role in microbial forensics that human population genetics plays in classical forensics – it is essential for computing probabilities associated with the origin of genetic material. Population genetics is only needed in human DNA forensics because databases are necessarily incomplete – if every person's DNA profile were in the database there would be no need to estimate the probability of finding a match. Similarly, *given the proper statistical framework*, microbial population genetics provides a calculable level of support for well-defined hypotheses about the relationship between the attack strain and suspect sources. In addition, this report will show how understanding bacterial population genetics helps define a useful database and suggests statistical sampling criteria.

The analysis of the forensic utility of bacterial population genetics must begin with the question: what is the relevant "population" of a given pathogen? Since the "phylogeographic" paradigm currently dominates thinking about forensic uses of genetic information, section 2 of this report summarizes the major criticisms of this approach. In effect, phylogeography ignores a major determinant of microbial population structure that is relevant to forensic inference. In section 3 we describe an alternative point of view, based on an "inference-on-networks" approach[16]. This framework was developed

recently to provide a rigorous basis for the forensic analysis of viral transmission hypotheses, and the reader is encouraged to review reference 16 for more details.

Much of the inference-on-networks framework developed for viral pathogens also applies to bacteria, but there are some crucial differences. First, the mutational spectra of bacteria are far more complex than that of viruses. Secondly, bacterial genomes still are large enough to present time and cost barriers to deep whole-genome sequencing, or the sequencing of large numbers of isolates. This places some practical limitations on the experimental validation of the network approach to bacterial genetic inference. In spite of this, the inference-on-networks framework has a salient advantage over the phylogeographic approach because it can be directly related to two well-known human forensic DNA methods – mitochondrial and Y chromosome DNA typing[17,18].

Section 4 discusses each of the 8 pathogens in the context of the inference-on-networks framework, and reviews the available data that is relevant to performing probability calculations for these pathogens. More details on estimating statistical quantities of interest are provided in section 5. Section 6 outlines a way forward and presents a roadmap towards a improved system for genetic inference that is useful for microbial forensic investigations.

A great deal of technical information is relegated to appendices. Appendix 1 derives an expression for a "match probability" applicable to microbial genetic forensics. Appendix 2 outlines the relationship between the population genetics of microbial DNA and that of mitochondrial and Y chromosome DNA in humans. Appendix 3 reviews the derivation of the "microbial paternity equation", which is the analogue of expressions used to calculate the degree of support that human DNA sequence data provides for hypotheses about parenthood. Appendix 4 discusses improved experimental designs for determining mutation rates.

2. The "phylogeographic" approach to source identification is inadequate

Within the last decade a number of articles have suggested an approach to source identification that attempts to provide statistical arguments about the significance of genetic similarity (or lack of similarity) between microbial isolates[7,19,20]. This approach involves the use of collections of geographically referenced isolates and often invokes phylogenetic construction methods to provide measures of relatedness. Some of these publications provide putative descriptions of the world-wide distribution of genotypes of several bacterial pathogens including *B. anthracis*[8] and *F. tularensis*[11], and more geographically restricted data on *Y. pestis*. The distributions are based on databases of genetic haplotypes or sequence data that have been built through extensive, but opportunistic collection efforts.

The correlation of phylogenetic information with geographical and temporal data is often referred to as *phylogeographical analysis*. It is a widely accepted mode of explaining certain features of the historical spread of pathogen genotypes over large geographical areas. The geographical association of a source strain is regarded as evidence that may guide subsequent investigation[8,19]. However, there are a number of problems and limitations associated with this approach to microbial source identification, even if it is only intended to "aid the attribution effort" rather than provide evidence that might be proffered in a courtroom or national security forum. In this section we undertake a critical analysis of the phylogeographic approach *sensu lato*, and its underlying problems.

Problem 1: The association of strain populations with geographical locations, not networks of global disease transmission

The association of pathogen genotype with geographic location is seldom rigorous. In nature, populations of pathogen variants cycle between various types of host animals and possibly the soil or water, which can migrate and spread diseases globally. The worldwide transportation network makes even long distance jumps possible. These

processes guarantee that genetic-geographic correlations will be scrambled in complicated ways. This problem is compounded by maps that use national political boundaries to describe the putative geographic locations of isolates[8,11], because non-technical decision makers may give such associations unwarranted weight.

Of course, none of this would present a problem *per se* if the databases contained isolates from *every* outbreak worldwide, and continuously sampled all new outbreaks. However, this is not practical, and all such databases realistically are incomplete.

Problem 2: Lack of authentication, completeness and accuracy of metadata associated with collected isolates

Beyond the problem of completeness, collections may also contain erroneous or misleading associations. There is generally insufficient information about the geographic coordinates and date of collection for the most isolates in pathogen databases to authenticate their origins. Moreover, important statistical characteristics such as the size of the outbreak or focus from which the isolate was collected are not available. At best, maps of "genetic diversity" are then simply annotated maps describing *the state of current collections*, whose accuracy at representing the actual worldwide distribution of the pathogen is not known. As will be explained in subsequent sections, geographic and temporal data is actually less important for testing source hypotheses than data on the size and characteristics of the outbreak or focus from which the isolate was derived.

Problem 3: Unknown representation bias

The phylogeographic approach offers no guidance about which, or how many outbreaks or foci must be sampled to achieve a collection that is "representative" of the true diversity of the pathogen population. Clearly, not every endemic region in the world has an equal chance of being represented in the collection, let alone every outbreak or focus. Thus, representation bias is an inevitable consequence of the fact that the isolates in current collections are typically chosen by academic microbiologists using *ad hoc* criteria, such as whether the isolate is "interesting" in a scientific or medical sense, or whether isolates can be obtained through chance collaborations.

This bias can be discerned directly in the data displayed in reference 8. The WHO map for worldwide anthrax prevalence shows that central Africa (Chad, Niger, the Central African Republic, and the Sudan) is an endemic or even hyper-endemic area for this disease. Nonetheless, the map of *anthracis* genotypes provided in reference 8 has *no* genotypes associated with central Africa. Clearly there is no relationship between the number or size of outbreaks in each geographic area and the number of isolates from that area represented in the collection.

Representation bias also constrains the validity of arguments that are sometimes proffered about the "rarity" of a particular genetic variant and its implication for source attribution[7,8]. The apparent rarity of a haplotype in nature will depend very much on the level of resolution of the typing system as well as the sampling bias that is built into most collections. A simple calculation illustrating the resolution effect is provided in Appendix 1. If a large enough outbreak network is sampled, and whole genome sequences are compared, every haplotype begins to appear "rare."

Representation bias also influences the interpretation of microbial genetic "matches" based on empirical distributions of pairwise genetic differences. In references 7 and 22, an empirical distribution of pairwise differences between VNTR haplotypes was constructed from the current VNTR database of worldwide samples and an Arizona case isolate. Clearly, if the database were more heavily populated by isolates from Arizona, the resulting distribution would have been shifted to much lower difference values, greatly changing the interpretation criterion for "matching" offered in this paper. In general, the distribution of pairwise genetic differences is only accurate if the isolates are sampled in an unbiased way from a completely connected transmission network.

Problem 4: Limits to the inferential power of phylogenetic analysis

In spite of the general acceptance of phylogenetic constructions as inferential tools, all applications of phylogeny to microbial source inference suffer from one or more of the following limitations:

(1) They implicitly assume that all possible sources have been identified, and one or more sequences are available for each source so that inference is occurring on a closed set of possible sources. When this assumption is *not* true, the inferences from phylogeny are restricted to statements that a pair of isolates is genetically closer to each other than to other isolates in the compared set, that the construction provides a measure of genetic similarity between two isolates in the set (e.g. the sum of the branch lengths to the most recent common ancestor,) and that there is an inferred common ancestor sequence for any pair of sequences in the compared set.

(2) An ancestral sequence identified for two isolates cannot be associated with any particular source without some additional information or assumptions. Information that is needed includes times associated with hypothesized transmission events, or evidence excluding the possibility of additional uncharacterized sources. When sequence data from the complete set of possible sources is not available, an observed phylogenetic relationship may be consistent with many alternative transmission trees. Similarly, several phylogenetic patterns may correspond to the same transmission history with equal probability. Figure 1 illustrates alternative transmission relationships that are consistent with a given phylogeny.

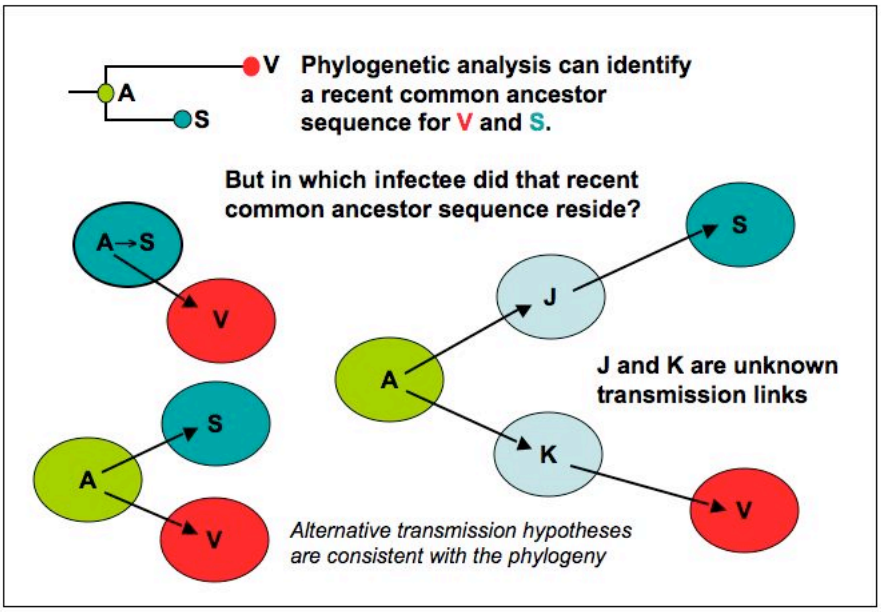

Figure 1. Alternative transmission relationships are consistent with a given phylogeny.

(3) Confidence levels are expressed as the ratio of the likelihoods of the most to the next-most probable trees[22]. The likelihoods themselves simply express the probability of a certain evolutionary path, but have no obvious relationship to the probability of a hypothesized source or transmission event. Bootstrap support levels are often cited as measures of confidence, but rigorously they provide only a relative measure of confidence that similar data would produce a similar tree, not that an inferred transmission path is probable.

A common misperception is that tests based on isolates from *known* microbial transmission trees provide support for deducing transmission relationships from phylogenetic construction. The classic paper by Leitner on human immunodeficiency virus (HIV) is often cited[23]. In fact, this perception is erroneous, being a clear case of the fallacy of exchanging the conditional. If Φ represents a phylogenetic construction, and $\mathcal{T}$ a transmission tree relating a set of genetic sequences, then comparisons such as Leitner's provide a measure of $P(\Phi|\mathcal{T})$, *not* $P(\mathcal{T}|\Phi)$. Moreover, it is not widely appreciated that several papers, including Leitner's, can be interpreted as demonstrating (perhaps inadvertently) that the probability $P(\Phi|\mathcal{T})$ of constructing the correct phylogeny given the genetic data from a known transmission tree is actually quite low. Leitner's paper showed that 14 out of 14 proffered phylogenetic constructions based on the known transmission tree were in error in at least one branch, implying that $P(\Phi|\mathcal{T}) \approx 0$.

To summarize, the current emphasis on phylogeographic analysis is questionable from three points of view:

- the *representation* of worldwide bacterial population genetic data in a form that might easily be misinterpreted by triers of fact or decision makers at the policy level
- uncertainties about (or lack of validation of) the accuracy and *representativeness* of that data and their effect on statistical inferences about "matches"

- ☞ the tendency to overestimate the *inferential power* of phylogenetic comparisons for identifying sources when all possible sources are not represented.

Even the most carefully planned campaign of worldwide collection would probably not provide an accurate picture of the global distribution of genotypes for any pathogen, given the realities of funding and worldwide cooperation. Fortunately, there is a well-defined theoretical framework that permits inferences with far more rigor, does not demand a large geographically "representative" pathogen archive, and provides explicit guidance for the number and types of isolates to be collected for reference purposes. This framework is described in the next section.

N.B. As was pointed out in section 1, these problems do *not* apply to a national pathogen registry system, in which laboratories handling select agent pathogens *must*, by law, provide genotyping information (or isolates) to a central database. The presumption of a registry system is that such a database is a complete, exhaustive, and accurate description of the "population" of genotypes held in laboratories. When an investigation like Amerithrax is considered in this context, notions such as "strain rarity" do have meaning, and a national database does have utility.

3. The transmission network theory of bacterial population genetics

Zoonotic diseases spread and evolve on host-host transmission networks. The relevant microbial "population" for answering many forensic questions regarding the origin of an attack strain (including the example hypotheses mentioned in section 1) is the set of all sub-populations of that microbe contained within the nodes of this network. The most fundamental type of node is an infected animal or human, although networks of individual hosts can be re-scaled to define more complex nodes such as outbreaks or foci. (However, very often the more complex node is itself a network of individual infected hosts.) New isolates can also be created by laboratory passage and exchange of strains between laboratories, and these may also be considered nodes in the network.

Bacterial isolates that are acquired for scientific, clinical, or nefarious use are typically samples taken from fundamental nodes – infected animals or humans – in the network for that disease. Thus, when one isolate is compared to another, we can rigorously formulate and answer questions about the probability that the nodes they came from have certain network relationships, based on genetic data. These calculations rely on certain fundamental probability distributions associated with the network and the process of genetic change on the network. As will be discussed shortly, this provides the closest analogy possible between microbial and human DNA forensics.

The bacterial population within an infected host is a mixture of genotypes differing from each other by a relatively small number of mutations. Although viral genomes are short enough to permit this distribution to be revealed by current high coverage sequencing methods, this has not been done yet for whole bacterial genomes. However, dilution plating often reveals genetic variants of bacteria if the mutations cause some noticeable phenotypic difference. Genetic sequencing of bacterial isolates currently results in a single consensus sequence, which is taken to represent the genotype of the isolate's population. The consensus genetic sequence of an isolate obtained from a host node is often regarded as representing the genotype associated with the outbreak or focus to which that node belongs, even though there is a non-zero probability that the consensus sequences derived from two independent nodes in the outbreak or focus network will differ by at least one mutation.

Changes in the consensus sequence from node to node in a disease transmission network are caused by several distinct mechanisms. First, transmission of the disease from one host to another often involves a small number of bacteria. If the genotypes of the infecting bacteria differ by chance from the consensus sequence of the population they are drawn from, then the population of bacteria formed by clonal expansion in the new host will exhibit the variant consensus genotype (this is sometimes called the "founder effect".) Second, if the immune system in a newly infected host differs from that of the infecting host, then the growth of a mutant genotype may be favored even if the infecting bolus was large. A third mechanism, random termination of lineages, can also cause a

shift of the consensus sequence in cases where a bacterial population is maintained at a low but nearly constant level in the host. (This is called Fisher-Wright drift[3].) Finally, if a pathogen colonizes an organ such as the gut, and comes in contact with other microbial communities, it may acquire new advantageous genes through horizontal transfer, causing the new genotype to dominate the population.

A disease transmission network develops over time and space. Certain parts of it may be associated with a certain historical time period and geographical region, but the entire network extends into the distant past and over large parts of the world. Because some microbes can remain dormant for long periods of time before infecting a new host, and because infected hosts can be transported to new geographical regions by various mechanisms, time and location are *not* the natural variables associated with genetic similarity. Isolates are genetically similar because they were sampled from nodes that were separated by a small number of transmission steps in the network. (Conversely, the probability that the consensus sequences of two isolates differ increases as their network distance increases.) Thus, the statistics of genetic change among nodes in disease transmission networks provide a more consistent framework than phylogeography for formulating and testing forensic hypotheses about strain origin.

For example, in reference 16 (also provided in Appendix 3) we derive an expression for the probability that two isolates are derived from nodes that are M transmission steps apart. When M = 1, then we are testing the hypothesis that there was direct transmission of the disease between two nodes, and the equation can be written:

$$P(M=1|s_1,s_2) = \left[1 + \frac{P(s_1,s_2|M>1)}{P(s_1,s_2|M=1)} X \frac{\mathcal{P}(M>1)}{\mathcal{P}(M=1)} \right]^{-1}$$

where s_1 and s_2 represent the two sequences being compared, $P(s_1,s_2|M)$ is the probability of observing s_1 and s_2 given that the nodes they are sampled from are separated by M transmission steps, and $\mathcal{P}(M)$ is the probability that two randomly chosen nodes will be

separated by M steps. This equation is analogous to the probability of paternity or maternity in human DNA forensics using Y chromosome or mitochondrial DNA[24].

It is also possible to define a "match probability" for bacterial isolates. Consider the case of an outbreak where an isolate has been obtained from one node in the outbreak network and the consensus sequence is determined. How likely is it that another randomly sampled node from the network would have yielded the same sequence? In Appendix 1 we derive an approximate form for this probability:

$$P(0|\mathcal{G}) = \frac{1 - \langle k \rangle}{1 - \langle k \rangle^{\mathcal{G}+1}} \frac{1 - (\langle k \rangle e^{-\Gamma_J \mathcal{G}_{host}})^{\mathcal{G}+1}}{1 - (\langle k \rangle e^{-\Gamma_J \mathcal{G}_{host}})}$$

where <k> is the average number of secondary cases of infection, $\mathcal{G}$ is the number of generations of transmission in the outbreak network, Γ_J is the genomic mutation rate, and $\mathcal{G}_{host}$ is the average number of generations of clonal expansion of the bacterium within a host node.

Another question that the network theory addresses in a transparent way is whether an isolate "belongs" to a given outbreak. Since the advent of disease tracking networks for food pathogens and tuberculosis, rules of thumb have been offered based on experience and intuition. Epidemiologists typically use some variant of the "Tenover criteria" to judge whether an infection can be assigned to an ongoing outbreak, or is a sporadic case[25,26]. If the typing pattern exactly matches that of other isolates from the same outbreak or if it differs by just a few markers, the case is included. Isolates differing by more than a few mutations are excluded. Essentially the same approach was taken by Lowell, et. al. who attempted to use the existing VNTR database for *Y. pestis* strains to define the cutoff for declaring "match" and "mismatch"[7]. However, it is easy to recognize the arbitrary nature of this approach, even in the face of attempts to refine it by adding other *qualitative* considerations such as the size or duration of the outbreak, or

whether there is other compelling epidemiological evidence to relate the case to the outbreak.

In the network theory of microbial population genetics these arbitrary qualitative judgments about outbreak membership are replaced by an explicit probability expression. The probability that a isolate with consensus sequence s_1 belongs to an outbreak network with consensus sequence s_2 is given by $P(M \leq D_0 | s_1, s_2)$ where D_0 is the diameter of the outbreak network. D_0 is a probabilistic graph-theoretic measure of the longest path of transmission (i.e. the largest number of host-host transfer steps between any two nodes in the network,) and can be estimated from <k> and the number of nodes in the network.

Finally, the network theory of bacterial population genetics provides a framework for understanding the probability that observed genetic changes can be attributed to natural horizontal exchange. Genetic exchange events occur in nodes that have been infected from two distinct networks (say different animal hosts), or two distinct nodes from the same network. A node in which transfer of genetic material has taken place may then be the initial node in a new, distinguishable outbreak network. The analysis of the new strain is complicated by the need to identify the sub-sequences in its genome that are likely to belong to each of the original networks. Once this has been done, the framework can be used on each genomic segment separately, to test the hypothesis that any two particular suspect reservoirs/outbreaks were the sources of the recombining strains. If the most likely suspect reservoirs are unlikely to have a natural route of contact, it might indicate an artificial (i.e. man-made) origin for the event.

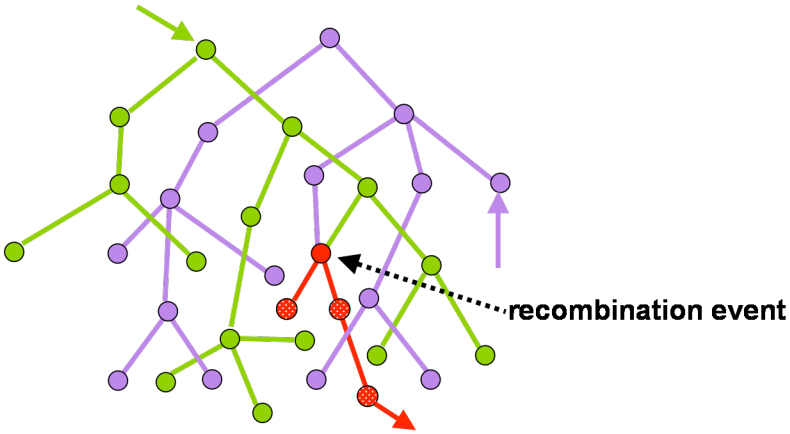

Figure 2. Two overlapping disease transmission networks that infect a common node.

To re-capitulate, the inference-on-nets approach recognizes that pathogen lineages are constrained to run along the vertices of a disease transmission network, and isolates that are the objects of forensic analysis are sampled from nodes in the network. Generally, a consensus sequence is used to represent a node, although deep sequencing raises the possibility that genetic sequence populations from isolates can be used in the future. Nodes can be any relevant isolated sub-population of the pathogen – an infected host, a herd, a flock, a focus, or an outbreak. The connectivity properties of the transmission network influence the probability of observing genetic relationships among representative sequences from different nodes.

Two types of information are needed in order to apply the inference-on-nets framework to bacterial pathogens. First, we must understand the structure of the disease transmission network in the enzootic and epizootic contexts. For 6 of the 8 pathogens in this study, the natural nodes are animal hosts, sometimes complicated systems of several mammals and insect vector species. It is not necessary to understand the detailed network in any actual case; we simply need good estimates of its statistical properties. For example, how many nodes (animals) are involved? What is the degree distribution P(k)? Is there node heterogeneity that can lead to segregation of the population into distinct sub-populations with different statistical parameters or increase the influence of adaptation on the genetic diversity of the pathogen? It should also be kept in mind that sometimes the relevant number of nodes is not simply the number of infected hosts that

exist in the present. Finally, given our current state of knowledge about many zoonoses there are probably cryptic nodes – i.e. classes of hosts we don't yet know about. The success of the approach may depend on whether cryptic nodes have a negligible influence on prediction, and can be ignored.

In infectious diseases such as SARS or FMDV, the transmission network properties can be estimated from contact tracing studies of actual outbreaks[16]. For the 8 bacterial diseases in this study such an approach is generally not feasible – the primary networks are built from transmission events involving wildlife, or unrecorded historical transport and migration events. However, it is possible to construct and validate models for disease transmission in the wild, and perform stochastic simulations of transmission networks in order to obtain the necessary statistical parameters. The next two sections of this report will provide more details on this approach.

The second type of information that must be acquired is the rate of genetic change measured in terms of transmission steps along the transmission network. The basic quantity of interest is the distribution $P(s_1,s_2|M)$, noted above. The best way to determine this is through direct empirical study of the consensus sequences of isolates drawn from pairs of hosts related by direct transmission. This is seldom possible for animals in the wild, so laboratory studies of host-host transmission using the actual animal in question is often more practical. In many cases, it will probably be necessary to use surrogate hosts or even *in vitro* passage to estimate the mutation rate. In any case, genetic typing is almost certainly inadequate for determining changes among isolates so closely related, i.e. whole genome sequencing of each isolate is required. Using laboratory determined mutation rates to infer rates of genetic change in the wild is not new, and forms the basis for inferences in the phylogeographic approach as well[6,7,22]. Nonetheless, to capture enough data to make effective, validated rate models for more than just a few types of mutational loci in bacteria is a daunting task.

4. The 8 pathogens of concern in the inference-on-nets context

In this section we review the state of knowledge about the 8 pathogens of concern in the context of the inference-on-nets approach to microbial forensics. Although each section is fairly brief, it is the result of an extensive survey of literature on the host range, ecology, geographical foci, outbreaks, transmission mechanisms, mutation rates, typing systems, and genetic sequencing related to each organism. The central issues are whether there is sufficient data to begin constructing models for deriving transmission network statistics, and what is known about the statistics of genetic change associated with the transmission of that disease.

Bacillus anthracis

The *B. anthracis* transmission network consists of the global chain of infected animals and soil areas that grew historically and geographically[27,28]. Figure 3 displays a schematic representation of a segment of this network, with cattle or other animals becoming infected from soil-borne spores, and occasionally migrating (or being transported) to a new location where they die, re-inoculating a new patch of soil. It has been suggested that healthy animals can be infected but not become clinically ill until some stress triggers acute disease, thereby permitting the transport of the animal and disease over long distances. In addition, scavengers consuming the carcasses of animals that died from anthrax and then succumbing themselves may provide an additional link in the transmission network.

Until recently, it was thought that the spores survived in soils in a dormant condition until physical processes concentrated enough spores in a grazing area that new hosts could be infected[28-30]. However, recent work by Fischetti suggests that when *anthracis* enters the soil, phage infection can suppress sporulation and permit vegetative propagation, including colonization of the earthworm gut[31]. Under certain circumstances, the phages are shed, re-activating sporulation. If this is true, then living *anthracis* communities in the soil may contribute to genetic change, including horizontal genetic exchange, as much as the infection of animal hosts does.

Anthrax outbreaks are seasonal in nature, and there is some evidence that environmental changes can trigger synchronized outbreaks from spatially distinct soil foci over extended geographic regions. Humans, infected by spores contained on hides, hair or other fomites derived from animals, are almost always terminal nodes, and do not further the transmission of *B. anthracis*.

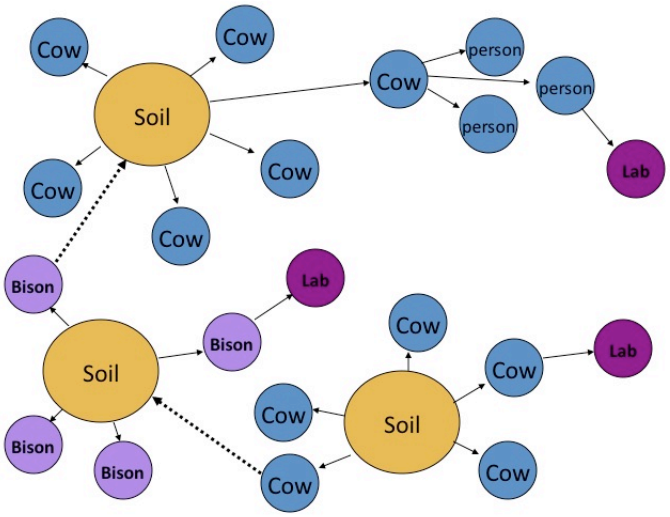

Figure 3. Simplified schematic transmission network segment for *B. anthracis*. Dashed arrows represent migration or transportation of animals to new locations.

A quantitative statistical description of the *B. anthracis* transmission network would primarily involve both natural and man-directed movement of infected cattle and other domesticated and wild animals among grazing areas. Secondarily, the transportation of animal products such as feed, bone meal, hides and hair may be taken into account. From the work of Hugh-Jones and others, soil type and other climatic factors appear to be important constraints that determine the location of soil nodes[28-30]. It is likely that the number of new soil nodes created by any given node after an outbreak is small, i.e. $\langle k \rangle \approx 1$ when measured over one "anthrax season". However, the persistence of nodes would raise the overall $\langle k \rangle$ leading to a more rapid growth in the network.

The observed genetic variability in geographic regions that have been studied often indicates re-introduction of anthrax from distant sources (i.e. distant branches of the worldwide net.)[32,33] However, the detailed modes of introduction a seldom analyzed, even though it is not uncommon to speculate about the possible geographical provenance of the source. For example, Kenefic, et. al. provide plausible phylogenetic arguments that the Western North American (WNA) strain followed human migration from Asia through the Bering straight, while the Ames strain originated in China[33]. Historical and geographical records for both the Ames and the WNA networks are quite extensive[28], and would provide a reasonable basis for testing and validating *B. anthracis* transmission network models when they are developed.

Kenefic et. al. applied high resolution typing to a set of Bacillus anthracis isolates from a single, but spatially distributed outbreak[35]. They examined "canonical" single nucleotide polymorphisms (SNPs), 15 variable number tandem repeat (VNTR) markers and four highly mutable single nucleotide repeat (SNR) markers. Only the SNR markers showed changes, 1 allele state change each, in 17 out of 47 isolates. If we were to assume that each isolate represents a single transmission step (M=1) from a common soil focus whose consensus sequence is characterized by the majority genotype, then a histogram of the fraction of isolates showing n allele changes would be an estimator of $P(\delta|M=1)$ where δ is a measure of genetic distance, in this case $\delta = n$.

Figure 4 compares the data observed by Kenefic with a binomial distribution estimator for $P(n|M=1)$. The binomial model predicts that the chance of seeing 2 or more allele changes between haplotypes is about 7%, so the lack of any 2-allele changes in Kenefic's data set could be attributed to chance. However, there are a number of reasons why both the data set and the probability model are not completely adequate for estimating the distribution of genetic changes. First, it is not clear that all the isolates are from hosts infected at the same focus. In fact, the geographical separation of many of the cases implies many presumably different foci – although they may be related to each other by a small number of transmission steps. Second, the typing system has a relatively slow rate of change, and more samples would be required to observe multiple allele changes

between haplotypes. The observation of n>1 data would greatly improve the fit with statistical models for P(n|M=1), and reduce the uncertainty in the fitting parameters. Finally, the model assumes that each locus has the same rate of change, whereas it is known that each of the 4 SNR loci has a different rate[20]. Regardless, this analysis shows how sequence data on multiple isolates whose transmission relationships are known can be used to deduce the fundamental statistical parameters for genetic change on transmission networks.

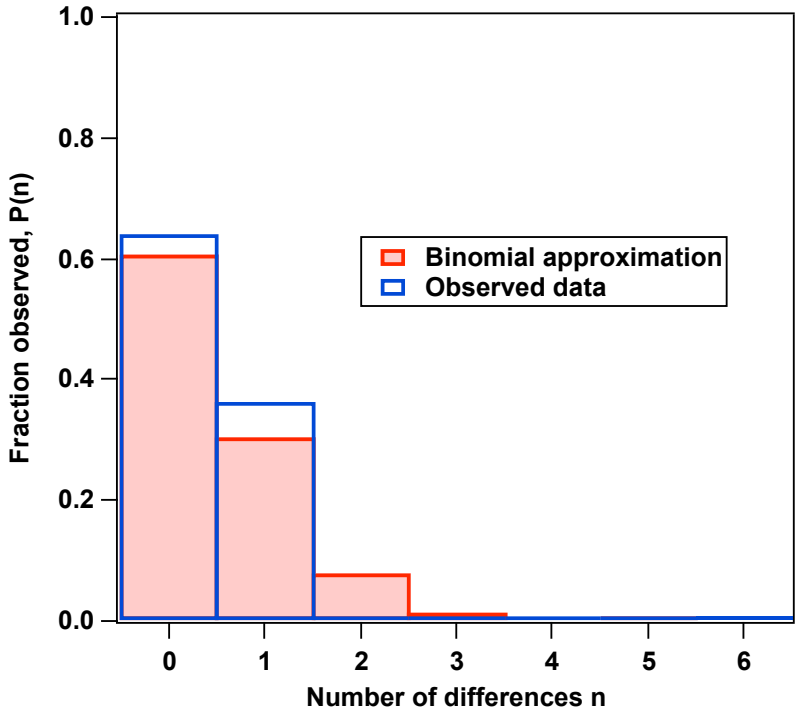

Figure 4. Estimate of the probability distribution P(n|M=1) from SNR data provided by Kenefic. The data is fit to the binomial model of Lee, et. al., assuming 4 loci, 25 generations and a per locus probability of mutation of 2.2 x 10^{-3}. This model predicts only a 7% chance of seeing isolates with n > 1 among 47 samples.

Unfortunately, aside from Kenefic's study, no other data has been collected from single anthrax outbreaks. Future studies will be most useful if they analyze multiple isolates from infections that are attributable to a single focus. In addition, isolates from outbreaks originating from a single focus over many years would provide valuable information on possible evolution of *B. anthracis* within a focus – presumably due to vegetative growth and interaction with the microbial community in that patch of soil.

Keim and co-workers have apparently used in-vitro passage experiments to study the mutation rates at VNTR and SNR marker sites in *B. anthracis*, but a comprehensive description of these experiments has not been published. Rates for VNTR markers in the range of 10^{-5} to 10^{-4} gen^{-1} and SNR markers in the range 10^{-4} to 10^{-3} gen^{-1} have been quoted, in line with similar estimates in other bacteria[20].

Yersinia pestis

The primary transmission network for *Y. pestis* is composed of populations of various rodents, with fleas as the primary vector[36]. Many different rodent hosts and flea species are often involved, and the participation of cryptic reservoir or vector species is often suggested to explain *Y. pestis* persistence in the wild[37,38]. In endemic regions, quiescent "maintenance" periods are punctuated by epizootic outbreaks, driven by changes in susceptible host populations, and possibly by mutational changes in *Y. pestis* virulence properties. Human cases of plague are usually correlated with epizootics, sometimes with intermediary roles played by household pets. Rarely a human outbreak will involve the pneumonic form, and small networks of infected human hosts with aerosol or droplet transmission linkages will form.

As long as there is no vertical (parent to offspring) transmission of *Y. pestis* within the flea population, a plague focus can be considered a simple network with fleas as a transmission linkage. (The clonal expansion of the pathogen within a flea host presents no additional complication to the network structure, or to the statistics of genetic change upon transmission.) However, each focus probably consists of linked networks of several host species roughly occupying separate roles as "maintenance" or "amplification" hosts. Thus, node heterogeneity may be an important detail in determining the genetic diversity of *Y. pestis* within a focus.

Focus-to-focus transmission is driven by animal movement and contact events that cause fleas to leave one host and acquire a new one. It has been noted that some Western US farmers have transported plague-infected rodents to their ranches from locations as far

distant as 100 km (as a rodent control measure.)[39] A well-known example of inadvertent global scale transmission was the introduction of plague into San Francisco near the beginning of the 20th century via trans-Pacific shipping.

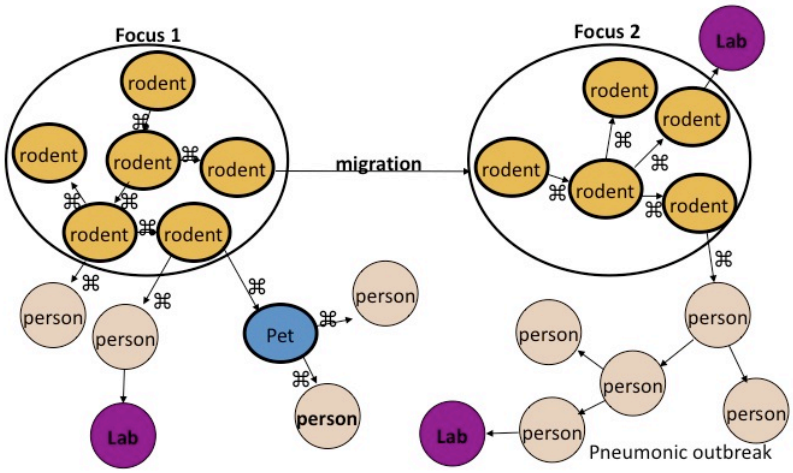

Figure 5. Schematic representation of plague transmission network elements. The ⌘ symbol indicates flea-mediated transmission.

There are already a variety of transmission models and data for plague that could provide starting points for building and validating models for the network statistical distributions required by the inference-on-nets framework[39-41]. A cellular automata model for plague transmission described by Keeling and Gilligan provides a potential approach to estimating focus-to-focus statistical distributions[41]. Adjemian et. al. have assembled a database of over 1000 historical human and animal plague cases in the United States[39]. From this they extracted 95 human cases and animal epizootics that were the first reported plague cases in a geographic location (at the city or county level) where plague was not previously confirmed. This data was fit to a diffusive type model for the spread of plague foci across the western United States, although it was acknowledged that transmission is better described by a stochastic network[42].

Transmission networks among rodents *within* foci have also been modeled in at least one case. Davis and co-workers modeled plague transmission among great gerbils, which are a major reservoir for plague in Kazakhstan[40]. This study defined nodes to be entire gerbil family groups located at a burrow system. Transmission between family groups occurs

with some probability when an infected gerbil comes in contact with a gerbil from another family group and a flea jumps from the infected to the uninfected one. Gerbil movement data, recorded during mark-and-recapture studies, was combined with data from field studies of flea dispersal in which fleas were marked using radionucleotides and their movements observed directly, to deduce network transmission probabilities. Random networks of vertices representing occupied burrow systems were generated from the real spatial arrangement of burrow systems observed on satellite images.

A plague focus consisting of a system of prairie dog colonies was studied by Auerbach, et. al., who performed GPS mapping and sampling for fleas at the burrow level[21]. However, no network properties were considered, and even the simplest connection between network connectivity properties and genetic diversity was missed. At least one traditional SIRS model for epizootic plague has been assembled[43]. Parameters governing the transmission of plague in a system of 9 rodents and 19 flea species within a single California focus are either extracted from experimental data or estimated in this paper. Again, no explicit network analysis was performed. Nonetheless, the data, and the field methodologies employed in these studies would clearly be appropriate for building actual network descriptions of plague foci in the Western US.

Finally, a number of human pneumonic plague outbreak transmission networks that were determined by contact tracing have been described and analyzed[44,45]. Gani and Leach studied a large number of outbreaks and generated a probability density P(k) for the number of secondary cases k generated by each primary case[45]. This was fitted to a geometric distribution $P(k) = p(1-p)^k$ with $p = 0.43$. In addition, distributions for latency period and infectious period were determined. A simple Markov chain model was then constructed to simulate the time-course of epidemics, but which also could easily be used to generate disease transmission network statistics.

To demonstrate how the primary data generated by Gani and Leach can be used to estimate network distributions required by the inference on nets approach, we have used methods described in reference 16 to estimate the distribution of node-node pair distances

for pneumonic plague outbreaks of various sizes. Figure 6 shows an example of $\mathcal{P}$(M) for an outbreak with 39 infected patients (about the size of pneumonic plague outbreaks that occurred in Mukden, China in 1946 or Madagascar in 1957[44].) As described in section 3, $\mathcal{P}$(M) is the probability that isolates sampled from two infected people drawn at random from the outbreak would be separated by M transmission steps. Figure 6 thus provides an estimate of the prior probability that any two samples will be connected by direct transmission, and an estimate of the maximum separation in transmission steps that can be expected. $\mathcal{P}$(M) is combined with information about the probability of genetic change after M transmission steps to infer whether two isolates could be related by direct transmission, or whether an isolate "belongs" to a given outbreak.

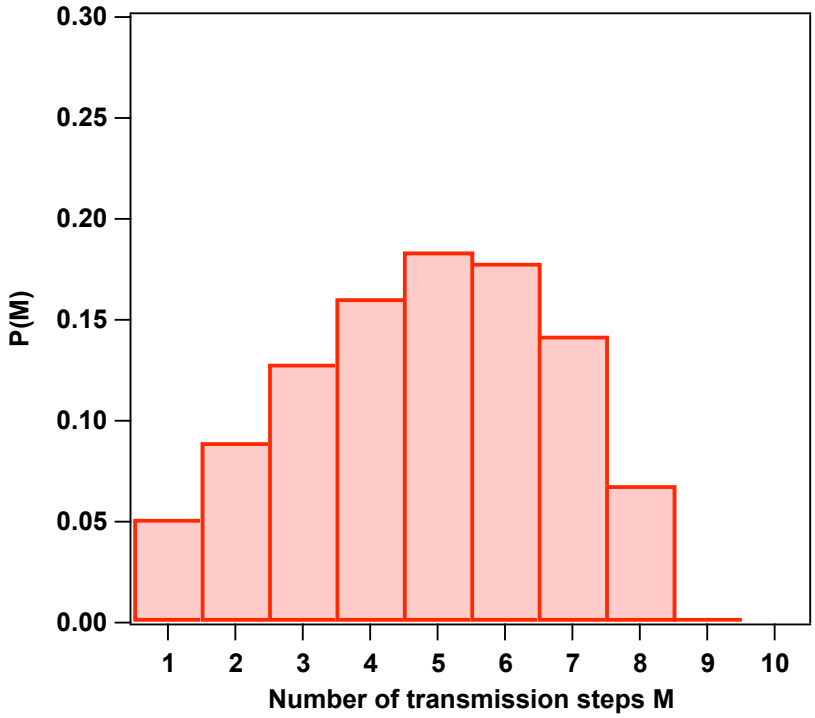

Figure 6. Pair distance distribution $\mathcal{P}$(M) calculated using the degree distribution P(k) derived by Gani and Leach for pneumonic plague. This example was generated using a Galton-Watson simulation of an outbreak with 39 patients.

Mutations observed in *Y. pestis* include large genome rearrangements, inversions, insertion, deletion and movement of transposable elements, insertion and deletion of

tandem repeat elements, and single nucleotide substitutions[46]. Vogler, et. al. reported mutation rate constants for 43 variable number tandem repeat (VNTR) loci determined by laboratory serial passage experiments[6,7,22]. They also obtained upper bounds for the deletion rate of a set of loci containing IS100 insertion sequence elements. Most importantly, by comparing VNTR mutation rates in *Y. pestis* and *E. coli*, they found essentially identical behavior, which is important support for the hypothesis that a general model of VNTR mutation rates applies to all bacteria. Unfortunately, nearly every aspect of the analysis of field data presented in this paper is incorrect. This included both an incorrect formula for estimating the most likely number of replication generations given an observed set of mutational differences, and an incorrect estimate of the maximum number of transmission generations expected in the prairie dog population studied. Nonetheless, the mutation rate data is a valuable contribution towards the construction of models of $P(s_1,s_2|M)$ for *Y. pestis*.

Escherichia coli O157 H7

The primary host species for *E. coli* O157 H7 is domestic cattle where it is maintained in the gut[47]. Cattle are generally asymptomatic, and shed the pathogen in their feces. Oral ingestion of contaminated fomites or food products is the major mode of transmission. Spillover into other animals both wild and domestic occurs regularly, including large human outbreaks through contamination of meat and other farm produce. It is estimated that in human outbreaks, about 20% of observed cases are due to secondary transmission, primarily through the oral-fecal route[48].

There are then two basic levels at which transmission networks can be considered for *E. coli*. First transmission at the individual animal level is used to describe the spread and genetic variation within a herd (or farm). Secondly, transmission among farms and feedlots can be analyzed to describe the spread and genetic variation across larger regions. Typically, when a human outbreak occurs, it involves a large number of primary cases infected from the same source, so the genetic diversity among the isolates will correspond to a single farm or even a single cow. (Note that a single farm may harbor more than one distinguishable genotype, as can ground meat products, which often

have multiple sources of contamination.) In the inference-on-nets approach, computing the strength of association of an outbreak to a farm would mean computing the genetic distance δ between the consensus sequences from the outbreak and farm, and the network diameter D_0 for that farm. A farm would be implicated if $P(M \leq D_0|\delta) > ½$. (This implies that it is more likely than not that the outbreak isolate was a directly transmitted from a node in the network of infected animals on that farm. When multiple genotypes are involved, the test is applied to each type separately.) Many of the basic elements needed to compute these quantities are already under development.

Turner et. al. developed an elaborate network model for E. coli O157 transmission within a farm[49]. They included population structuring found among typical UK dairy farms such as animal social groupings and separate management groups for unweaned, weaned, dry and lactating cows. In addition, they differentiate between direct transmission, associated by cow-cow interactions such as grooming, and indirect transmission caused by general environmental contamination (including feed and water) with feces. They observed that indirect transmission, which effectively raises the probability of one cow transmitting infection to all other cows (rather than to just close contacts), greatly influences the outbreak dynamics. This effect could clearly also have an important influence on the pathogen genetic population, by making the initially infected cattle into "superspreaders", and shortening the network diameter D_0. Since *E. coli* O157 is known to persist in hay, feed, and manure for long periods of time[50], the indirect transmission mode might be very important.

Transmission of *E. coli* O157:H7 among farms and feedlots is also well characterized. The movement of infected cattle may account for some of the observed spread, but Hancock, et. al. observed that farms separated by hundreds of kilometers often exhibit the same strains in the absence of recorded cattle movements[47]. *E. coli* O157:H7 has been reported in a variety of wildlife species, and wildlife movement has sometimes been implicated in transfer of E. coli between farms. Another possible mechanism is contamination of commercial feed at the place of manufacture. Like indirect transmission within a farm, feed contamination may be a superspreading mechanism that

shortens the inter-farm network diameter. This makes source attribution more difficult, as several farms may have $P(M \leq D_0|\delta) > ½$.

Using data from the United Kingdom's Cattle Tracing System, Woolhouse, et. al. studied the network structure of 55 Scottish farms[51]. They computed an R_0 value (a quantity similar, but not identical to <k>) and found that 20% of cattle holdings accounted for 80% of the R_0 value. This implies there may be a strong "superspreading" component to farm-farm transfer of O157 as well. Brennan, et. al. studied the contact networks among U.K. farms, markets, dealers, and slaughterhouses[52]. Heath, et.al. developed a network model for farm-farm spread and derived degree distributions P(k)[53]. Vernon and Keeling recently published a dynamic network model for cattle movements among farms in the United Kingdom[54]. Thus, there are a wide variety of studies that provide a basis for simulating disease transmission among cattle and farms, which can provide the network statistical distributions necessary to apply the inference-on-networks approach to *E. coli* population genetics.

There have been a number of studies of genetic change within populations of E. coli at the typing system level, and some of these raise certain issues regarding current models for mutation rates. Noller, et. al. demonstrated differences in VNTR haplotypes among outbreak related and non-related isolates[55]. Isolates that had the same PFGE pattern differed sometimes differed at one or two VNTR loci. However, with no sound theory to guide them, the authors of this work were not able to provide a coherent test for whether two isolates whose PFGE patterns are identical, but VNTR haplotypes differ belonged to the same outbreak.

Noller et. al's observations of epidemiologically linked isolates of O157 with different VNTR haplotypes prompted them to study the mutation rates at VNTR loci by in-vitro serial passage studies[56]. They performed two types of experiments involving different mixtures of clonal selection and serial dilution steps. In a system of 7 loci they estimated a total mutation rate of $\approx 4 \times 10^{-3}$ gen^{-1}, with the highest probability of change at one particular locus with rate $\approx 3.5 \times 10^{-3}$ gen^{-1}.

Vogler, et. al. studied a set of 28 VNTR loci in 10 different strains of O157 using a well-defined serial passage protocol involving clonal selection in parallel lineages[57]. Vogler's VNTR loci included the subset that had been studied by Noller. The highest combined mutation rate measured by Vogler for all loci was $\approx 6 \times 10^{-4}$ gen^{-1}, nearly an order of magnitude lower than was observed by Noller. The reason for this discrepency is not clear from the information provided in the two papers, but might be due to the different experimental protocols. Vogler, et. al. did find that (as for Y. pestis VNTRs) a simple statistical model for VNTR mutations similar to one used for human STRs seemed to fit the E. coli data. This paper and others by Keim and co-workers argue that VNTR mutations, like STRs are largely neutral[6,7].

In a long series of articles, Lenski and co-workers have studied the mutational spectrum of a laboratory-adopted strain of E. coli after many generations of passage[58]. For example, a recent paper by Barrick, et. al. describes a comparison of the consensus genomes of a non-O157 E. coli strain that had been continuously passaged by 1:100 serial dilution for 40,000 generations[59]. They observed a large number of mutations of different types, including substitutions, insertions, deletions, transposable elements, and inversions. Many of these mutations occurred in the same genes in independent lineages. They concluded that nearly all the observed mutations were *adaptive*, rather than neutral. This observation is consistent with their serial dilution protocol, where only adaptive sweeps are likely to change the consensus genotype of the population, and the constant growth environment for all lineages.

Thus, while some basic estimates of *E. coli* VNTR mutation rates have been published, there are still a number of apparent inconsistencies among experimental results that will have to be resolved before reliable modeling of VNTR rates for bacteria is possible. Many of these same issues apply to the VNTR rates of *B. anthracis* and *Y. pestis*, but have not come to the fore because there have been no other independent experimental measurements of rate constants for these pathogens.

Brucella melitensis

Several *Brucella* species cause disease in humans, including *melitensis, abortus*, and *suis*. In this section we focus on *melitensis* since it is the most virulent in man, and a significant cause of human disease in the US Southwest[60,61]. *Brucella melitensis* is primary maintained in goats and sheep, with spillover into cattle and humans. It can be transmitted by fomites and aerosols, but most large human outbreaks are associated with contaminated goat/sheep milk and cheese. Cattle can become infected from goats or sheep through shared pasturage. The *B. melitensis* strain Rev-1, which is commonly used to vaccinate livestock, is an attenuated strain, and is known to cause disease in humans. Like anthrax, plague and E. coli O157, *Brucella melitensis* occurs naturally in wild ruminants. Thus, at the transmission network level, there are natural parallels with these diseases, as shown in Figure 7. Since milk products are major routes for transmission to humans, factories involved in goat milk packaging or the manufacture of goat cheeses should properly be included as nodes.

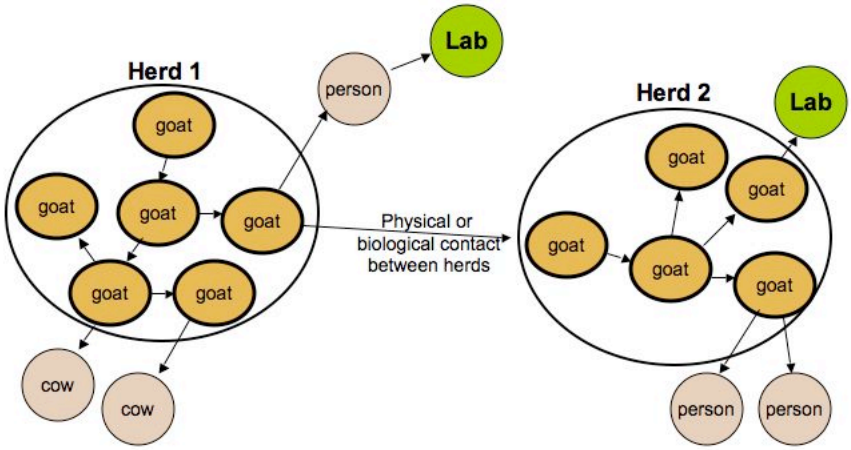

Figure 7. Schematic representation of Brucella melitensis transmission network elements. Herds of sheep and cattle, and sheep milk processing sites are also nodes in the larger scale network (not shown).

Yamamoto et. al. developed a model for the transmission of *Brucella abortus* within and between cattle farms[62]. The parameters of this model are probably not appropriate for *Brucella melitensis*, but could be modified to permit simulation based network estimates

of P(M) and other statistics of outbreaks. Zinstag modeled *Brucella melitensis* transmission among sheep, cattle, and humans using parameters appropriate for Mongolia[63]. However, this model only considers aggregated populations of animals, rather than individuals. The empirical basis for constructing realistic network models for *melitensis* in wild caprine populations is far less developed than for *Y. pestis, E. coli*, and *B. anthracis*. Similar findings obtain for the *suis* and *abortus* strains.

In a major review, Whatmore identified a number of typing systems that have been applied to characterizing *Brucella* genomic diversity[64]. In the past, the objective of most typing system development has been to differentiate the various host-species specific biovars within *Brucella* spp. More recently, there has been increased interest in developing higher resolution techniques. Whatmore identified 3 three VNTR schemes published to date[65-67]. Tiller,et. al. recently published a comparison of two VNTR typing systems for Brucella melitensis[68]. These typing systems used partly overlapping sets of 15 loci each. While it is highly plausible that additional resolution could be obtained from the examination of insertion sequence loci and genome-level nucleotide substitutions, neither system has been characterized in *Brucella*.

Mutation rate data for the VNTR systems have not been published at this time, but Whatmore has reported both in vitro and in vivo (pigs) passage experiments in which single repeat changes at one or more loci were observed[67]. An interesting observation is the presence of mixed cultures containing both the new and old alleles at comparable titers in the blood of the infected animals. This strongly implies some selection process driving the conversion of the consensus genotype within the infected host. It remains to be seen if the Brucella VNTRs can be quantitatively described by the same model that appears to work so well in *Y. pestis* and *E. coli*.

Burkholderia mallei

Burkholderia mallei is the causative agent of glanders in animals and humans[69]. The pathogen is primarily maintained in horses, which can be cryptic, or subclinical carriers of the disease[70]. Transmission mechanisms are not well understood, but are believed to

involve ingestion of bacteria shed by oral, nasal or ulcerous discharges, and less probably inhalation and skin contact with lesions. Some association of transmission with the sharing of water or food sources has been cited[69]. A schematic representation of part of a transmission network for glanders is displayed in figure 8. As in other zoonoses there is both a within-focus (in this case herd) and between focus (herd) aspect to the overall network.

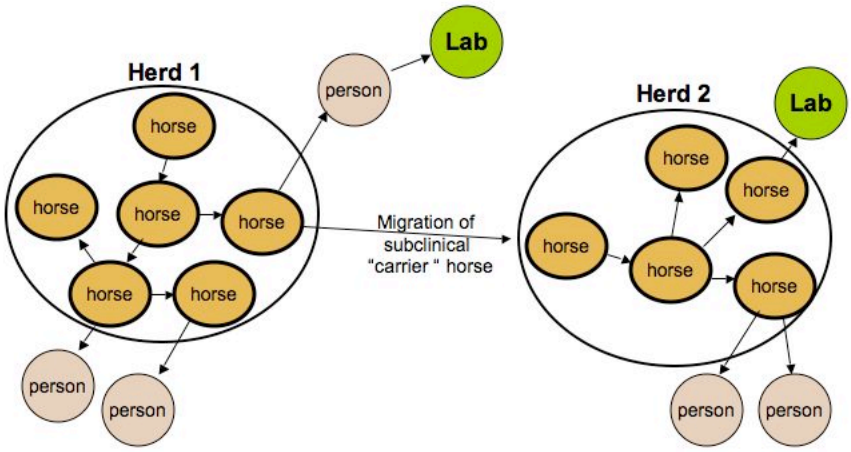

Figure 8. Schematic representation of *Burkolderia mallei* transmission network elements. In addition to horses, donkeys and other equine hosts may be involved.

Since *B. mallei* apparently does not exist outside its host, and is strongly adapted to a narrow class of host species, the transmission network description of outbreaks and foci should be relatively simple. However, we found no published work in this area. However, there is a rich literature on social interactions among horses, both wild and domestic, and it is likely that a disease transmission model could be constructed[71,72]. It is possible that many aspects of transmission models for *E. coli* and *Brucella* would be similar for *B. mallei*. In addition, there are some models available for equine influenza that might provide useful estimates of transmission parameters[73].

The genome of *B. mallei* is densely populated with repeat regions and insertion sequence elements. Romero, et. al. observed that a high level of genomic variation in clonal populations of *B. mallei* would be consistent with the fact that exposure consistently fails

to induce immunity in a variety of hosts. High levels of genetic diversity could be a consequence of the large number of repeats and IS elements. (The terminology "simple sequence repeats" or SSRs is roughly equivalent to "VNTRs" in other publications and "STRs" in human DNA.) They tested this hypothesis by determining the mutations that occur on passage of *Burkholderia mallei in vitro* and through animal and human hosts using whole genome sequencing. They report a very large number of insertion and deletion events at simple sequence repeat loci as well as other sites, even on single passages[74]. This is one of the first experiments that systematically studied genetic change in infected hosts during short-term acute infection with a bacterial agent. By comparing the initial and passaged genotypes, $P(\delta|M=1)$ could be estimated from the detailed data generated in this experiment, given a definition of genetic distance that is defined consistently for the variety of observed mutation types.

The mutable SSR and IS loci clearly provide a basis for typing systems for *Burckholderia mallei*. A recent comparison of 9 *mallei* genome sequences[75] identified a number of diverse loci, but no validated typing system derived from these observations has yet appeared. U'Ren et. al. developed a 32 locus MLVA system that can be used with *mallei* or *pseudomallei* and estimated mutational rate constants by laboratory serial passage experiments of *pseudomallei*[76]. Application of this system to sets of outbreak related isolates will be an important milestone for determining its suitability for transmission-on-nets analysis.

Francisella tularensis

Like *Brucella*, *Francisella* is characterized by a number of closely related, ecologically distinct sub-species that can cause disease in humans. In this section we will restrict most of our attention to the type A strain, since it causes the most severe disease in humans, and because it has been involved in a number of well-known and incompletely resolved outbreaks of respiratory tularemia within the United States[77]. In the standard epidemiological picture, *Francisella tularensis* type A (tularensis) in North America is maintained in populations of rabbits and hares in the wild. It can be transmitted by ingestion, inhalation, or direct contact through the skin and mucous membranes, but

insect vectors, particularly ticks and biting flies, are considered the most common mechanism. *F. tularensis* can survive for long periods on fomites in the environment, including food and water. Individual flies may carry the organism for as long as 2 weeks and ticks throughout their lifetimes. Viable bacteria can also be found for weeks to months in the carcasses and hides of infected animals, and occasionally it is transmitted to carnivores. Outbreaks in humans represent spillover events sometimes mediated by pets such as dogs.

Recent findings suggest that the pathogen is transmitted vertically in tick populations (transovarial transmission[78].) The possibility of tick-tick transmission complicates the network structure and estimates of the rate of genetic change within a *tularensis* focus, as indicated schematically in figure 9. In contrast to the role of the flea in *Y. pestis* transmission, where it is assumed that only one flea connects a pair of direct transmission related rodents, we now must consider the possibility of two or more ticks intervening in a rabbit-to-rabbit (or other host mammal) transmission linkage, and pathogen reproduction and adaptation within the tick.

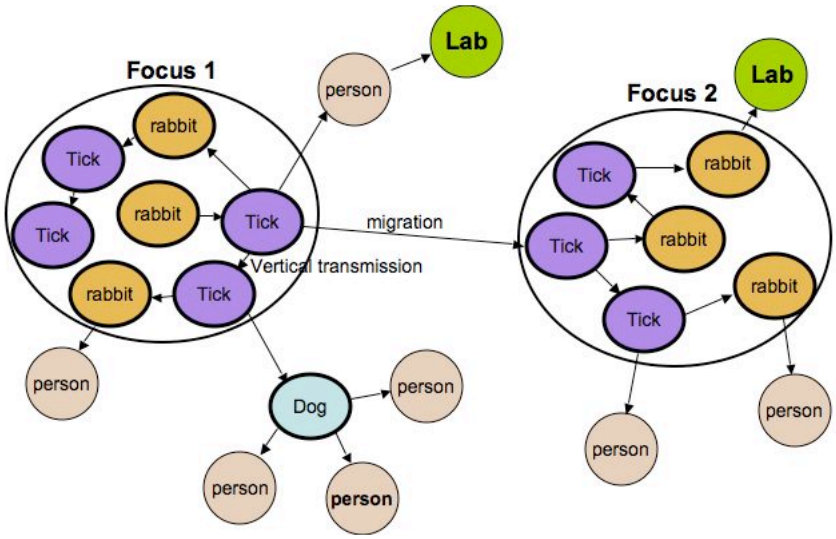

Figure 9. Schematic representation of *Francisella tularensis tularensis* transmission network elements. The importance of rabbits in maintaining the disease is not yet established, and the proper model of maintenance foci might involve only ticks, while focus-to-focus transmission may be mediated by other animals. Note also the inclusion of tick-tick transmission.

Careful study of the structure of *F. tularensis* foci has only begun recently. Goethert and Telford studied ticks on Martha's Vineyard and found a distinct "microfocus" with a diameter of only 290 meters where the probability of finding ticks infected with *F. tularensis* was 20-30 times higher than background[79]. More significantly, this area harbored the highest percentage of uncommon VNTR haplotypes, including one that matched an isolate from fatal case of human tularemia on Martha's Vineyard. They describe a plausible mechanism for transmission of *F. tularensis* to animal hosts and their participation in establishing of new foci involving ticks. Intriguingly, this picture reverses the usual picture by, in effect, giving the ticks the role of hosts and the animals the role of vectors. Thus, a possible qualitative basis for a network model has been laid, but statistical parameters from field studies are not yet available.

Many genotypes can be associated with a single focus, and hence a single outbreak. For example, Peterson, et. al. found two distinct type A PFGE patterns associated with a cluster of 5 cases of tularemia in Utah[80]. The Martha's Vineyard focus exhibited a number of different VNTR haplotypes[79]. This can be understood as a consequence of large effective network size associated with long-term maintenance of a pathogen in an insect host.

Like other bacteria, the genome of *F. tularensis* has a large number of mutable loci including VNTRs, IS elements, and single nucleotide substitutions. Many of these have been utilized to generate typing systems with high resolution. Vogler, et. al. describe an 11 locus VNTR typing system for *F.t.* which was consolidated from an earlier 25 locus typing system[81]. Pandya et.al. have developed a whole genome resequencing array that could be used to identify single nucleotide substitutions in outbreak samples[82]. In a separate publication, Vogler and co-workers also reported a whole genome SNP typing assay using a microarray. None of these systems have as yet been applied to genotyping isolates from a single outbreak or a single focus. We were not able to find any recent published work on insertion sequence based typing. Moreover, we found no published measurements of mutation rates, either in vitro or in vivo passage for VNTR or IS loci.

Goethert et. al. applied a 4 locus VNTR system to characterize two foci on Martha's Vineyard, including the microfocus described above[79]. Phylogenetic analysis indicated that the microfocus had a star-like phylogeny indicative of clonal expansion, while a "emerging" focus showed signs of multiple independent introductions of F.t. from other sources.

Johansson et. al. studied the variation of alleles at a single VNTR locus for sets of *holarctica* isolates derived from five separate tularemia outbreaks in Finland and Sweden[84]. Each outbreak set had 6 or 7 isolates, and up to 3 allelic variants each. Since the "outbreaks" occurred over relatively large geographic areas and over periods of 1 year or more, it is not clear that a single focus was the source. However, assuming that the structure of the foci for *holoarctica* are similar to those that Goethert studied, it is not surprising to find multiple allelic variants.

Gurcan, et. al. examined Turkish and Bulgarian isolates of *Ft holoarctica* using a 6 locus MLVA system, and found that one Bulgarian isolate exhibited the same haplotype as several Turkish isolates[85]. These isolates were from outbreaks in geographic regions 1000 kilometers distant from each other. They point out the many possible modes of contact between the two countries that could have resulted in direct transmission.

Thus, while many of the necessary elements for building a network-based model for *F. tularensis* population genetics are beginning to take shape, the increased complexity of the transmission network and unresolved questions about maintenance and transmission suggest that it will be more complex and difficult to validate than *B.anthracis*, *Y. pestis*, *E. coli*, *B. melitensis*, and *Burkholderia mallei*. On the other hand, continuing progress in characterizing the Martha's Vineyard focus system, especially if statistical studies of tick-animal contacts and further characterization of tick transovarial transmission become available, may permit such models to be developed.

Burkholderia pseudomallei

Burkholderia pseudomallei, the causative agent of melioidosis, is genetically closely related to *mallei* but has a completely different ecology[86,87]. It can live and reproduce in soil and water, which are regarded as its primary reservoir. Currie has suggested that it may switch from a dormant to a reproducing state in response to climatic factors[87]. Transmission to animals and humans occurs through ingestion of contaminated soil or water, infection of open cuts and sores, and possibly through aerosols. Strictly speaking, this is not a zoonotic disease, in the sense that infection of animals is apparently only incidental to its lifestyle, and the major engine of genetic change is therefore not within-host adaptation and host-host transmission.

Because the primary reservoir is soil and water, the spatial boundaries that separate definable network nodes are diffuse, and many individual clonal sub-populations may overlap in one spatial region. The bacterial population size (and hence diversity) is probably correlated with the spatial volume of the colonized area. While it is known that the distribution in soils is uneven, the factors that determine the local concentration are not known. Clonal populations of *pseudomallei* in water are often associated with biofilm structures[88], but rainfall and flooding can cause general mixing of populations so that, in general, foci will be complex metapopulations. This situation is represented in figure 10, where coupled systems of diffuse (and overlapping) soil and water populations form "supernodes" that may be related to each other by transmission events mediated by humans, animals, or transport of agricultural products.

This picture is supported by recent work in Thailand which determined the genetic diversity of *pseudomallei* sampled from soil in defined areas. Chantaratita and co-workers found that the *pseudomallei* cultured from soil samples taken from 3 locations in a 240 m^2 plot of land had 12 PFGE types and 9 MLST patterns[89]. Thus, it was possible to obtain relatively distant genotypes only a few meters apart. U'Ren et. al. determined genotypes of pseudomallei isolates from a focus in Thailand using a 26 locus MLVA typing system[90]. They cultured isolates from the soil in 19 individual 20 cm^2 areas sampled within a 50 km^2 region. Many of the areas harbored 2 or more different

genotypes, and there was little correlation between location and genotype. Finally, Pearson and co-workers found multiple VNTR genotypes associated with outbreak on a single farm in Australia[91].

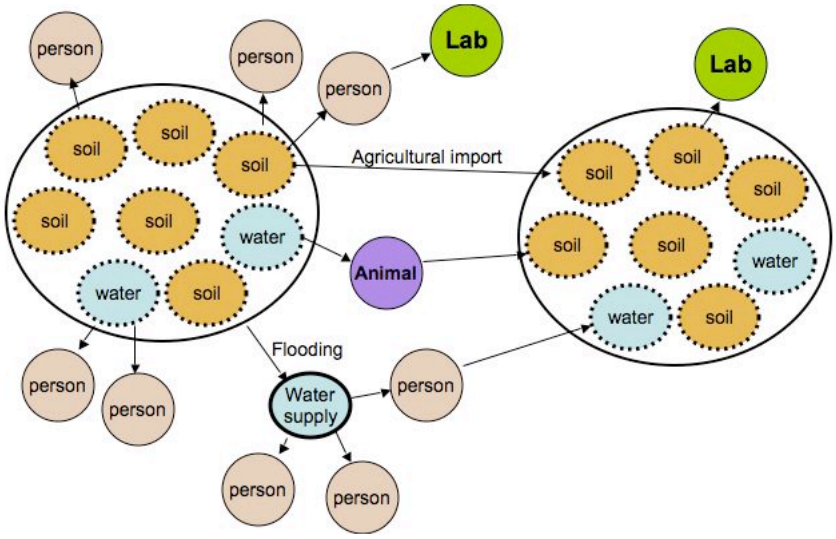

Figure 10. Network elements for *Burkholderia pseudomallei*. Partially localized, but spatially and temporally overlapping soil and water populations form a metapopulation that forms a complex node, difficult to represent by a network structure.

If we regard an entire *pseudomallei* focus as a "supernode" in a larger network, then describing its internal population genetic structure raises several new technical issues. When a zoonotic focus is modeled as a transmission network, the individual infected host or vector provides a natural "unit" that bounds generation numbers and defines aggregate statistical properties such as degree distributions and the number of nodes. Empirically, for example, the number of hosts and vectors can be estimated by counting methods in the field. Within a landscape colonized by *pseudomallei* it is not apparent what the "natural unit" for network analysis is, other than individual re-colonization events themselves. If this is true, it is difficult to see how to estimate the number of nodes associated with a focus, or estimate the average number of generations associated with a node so defined.

Thus, several basic conceptual problems must be resolved before the inference-on-networks approach can be applied to Burkholderia pseudomallei. In addition, there is a

need for basic experimental work on growth in the soil environment – growth rates, colony sizes, and mutation rates including genetic exchange.

Clostridium botulinum

Populations of *Clostridium botulinum* are maintained in soil and water under anaerobic conditions and exhibit tremendous genetic diversity and worldwide presence[92,93]. Subspecies are typically designated by the toxin serological groupings A – G with various recognized subgroupings. Outbreaks of botulism in humans are typically caused by *C. botulinum* contamination of food. In animal outbreaks, consumption of dead animals, or feed that is contaminated by feces or carcasses is usually implicated. Generally, botulism is caused by the toxin produced by anaerobic growth in the consumed material, but occasionally colonization of the gut by *C. botulinum* occurs. The most likely scenarios for deliberate use of *C. botulinum* to kill or injure people involve the introduction of botulinum toxin into foods. If the toxin is not highly refined, there is reasonable probability that genetic material from the *C. botulinum* strain could be found in the contaminated food, and subjected to typing or sequencing.

Like *Burkholderia pseudomallei*, it is likely that *C. botulinum* foci have a complex metapopulation structure that is currently difficult to model as a network. While many authors have considered the geographic distribution of the various toxin types, there are only weak correlations between location and type[94]. We have found little published literature on the ecology of this species, and studies of genetic change have focused on the toxin gene region. As might be expected for a soil dwelling bacterium, there is evidence of extensive recombination and insertion events in the toxin gene complexes of *Clostridium botulinum*[95].

5. Estimating the statistics of disease transmission networks and validating the inference-on-nets approach

In this section we want to describe in more detail how network models for bacterial pathogens can be constructed in order to determine the probability distribution functions that are used to make inferenence-on-networks calculations. For convenience, various statistical quantities that are used in the framework are listed in table 1.

Table 1. Distributions and other statistical quantities used in inference-on-networks calculations.

Statistical quantity	Name	Description
G, N	Network size	Either the number of generations or the number of infected hosts in a disease transmission network
P(k)	Degree distribution function	Probability that an infected host transmits to k other hosts
<k>	Average degree	Average number of secondary cases per infected host
$\mathcal{P}$(M)	Pathlength distribution function	Probability that two randomly chosen infected hosts are M transmission steps apart
D_0	Network diameter	Largest value of M for a transmission network with a certain number of infected hosts
P(G|N)	Network scaling distribution	Probability that a transmission network with N hosts contains at least one host G generations from the index case
$\delta(s_1,s_2)$	Genetic distance metric	A measure of how different two genetic sequences are
P(δ|M)	Sampling distribution for genetic change	Probability that two sequences will differ by δ given that they are separated by M transmission steps
P(δ|M=1)	Direct transmission sampling distribution	Probability that two sequences will differ by δ given that they are separated by 1 transmission step
P(0|G) or P(0|N)	Match probability	Given a sequence from one host, the probability that another host in the network will have the same sequence, conditional on the size (G or N) of the network
$P(M \le D_0\|\delta)$	Outbreak inclusion probability	Probability that two sequences differing by δ came from the same outbreak or focus
P(M=1|δ)	Direct transmission probability	Probability that two sequences differing by δ are related by direct transmission

To illustrate how these quantities come into play in a case investigation, we turn to the hypothetical scenario described at the beginning of this report. In this scenario, the

investigator has one or more isolates of *Francisella tularensis* that were collected from the Katama site in 2003 and had been stored at Tufts (these are the "K-03 isolates"). In addition, he has the set of clinical isolates from the president and his entourage (the "P-10 isolates".) From their genetic sequences he can calculate the genetic distance $\delta(s_1, s_2)$ between any K-03 sequence s_1 and any P-10 sequence s_2. How does he compute the probability that the *Ft* that infected the president could have originated at the Katama site[79] 7 years ago?

First, he will need to have an estimate for the function $\mathcal{P}(M)$. This can be computed if there is a model for the *Ft* transmission network that describes foci like the Katama site, and some estimate of the size of the focus (number of infected ticks and rabbits) in 2003 from field data. This also permits estimation of D_0, the network diameter for the Katama 2003 site. The network diameter is basically the value of M beyond which P(M) is negligible.

Secondly, he will need an estimate of $P(\delta|M)$, computed using data from mutation rate experiments or from carefully correlated field data comparing sequences from transmission-linked pairs of hosts. From $P(\delta|M)$ and $\mathcal{P}(M)$, the investigator can then compute $P(M \leq D_0|\delta)$ from equation (9) in Appendix 3. $P(M \leq D_0|\delta)$ is the *probability that the P-10 sequence and the K-03 sequence both originated from within the Katama transmission network*. (To account for possible laboratory growth of the agent, and passage through the victims, this calculation might use D_0+1 rather than D_0.)

Note that the investigator can (and should) compare each reference sequence from the P-10 set with each from the K-03 set; the results will vary slightly, depending on the relative location within the transmission network of the hosts from which the isolate was obtained. But ambiguous results would only result if one of the K-03 isolates fortuitously came from a host or vector very close to the "edge" of the transmission network.

The provenance can be further narrowed by computing $P(M \leq M_0|\delta)$ with $M_0 < D_0$. Ultimately, the investigator can state that the P-10 strain is less than M transmission steps away from a particular isolate obtained from the Katama focus in 2003, with an explicitly calculated probability.

The most general description of how the inference-on-nets approach can be implemented is illustrated in figure 11. The problem neatly divides into three separate technical areas: the network statistics of disease transmission, quantitative descriptions of genetic change during disease transmission and infection, and synthesis of the desired distributions. Color shading has been used in table 1 to indicate which statistical quantities belong to each area.

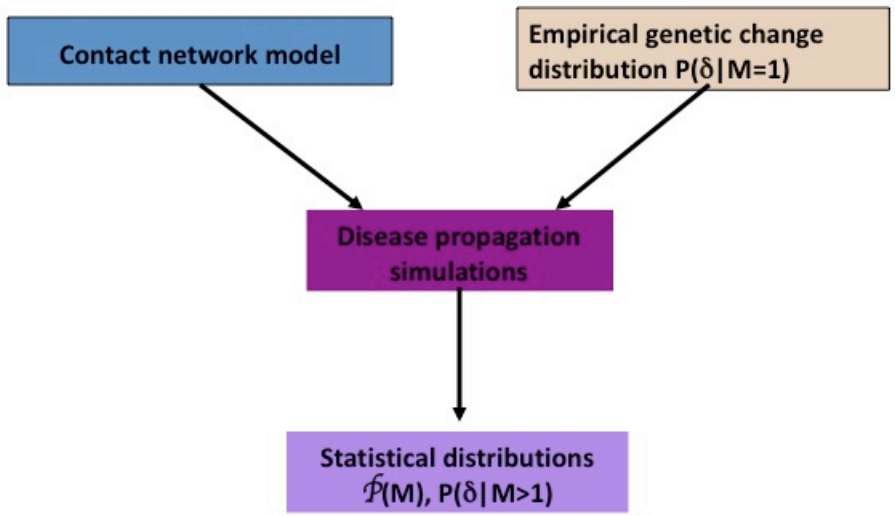

Figure 11. General scheme for implementing the inference-on-networks approach.

It is important to understand that it is *not necessary to know the actual network of an outbreak or focus in question* to make inferences. Obviously, the actual network was determined by stochastic events, so it makes little sense to attempt to reproduce it in detail. For zoonoses it is seldom possible to know even part of the actual network in any detail. For example in the case of anthrax, most of the relevant network is a historical entity, clearly not amenable to study. However, a valid model for the network should reproduce the statistical features that are determined from empirical sampling of real epizootics and foci of the same disease: How well does the model reproduce the flow of

anthrax and pestis across the US, as known from historical records? How well does it explain the distribution of *tularensis* genotypes in Martha's Vineyard? In a known outbreak of brucellosis in a herd, does it reproduce the distribution of observed pair-wise genetic distances?

As suggested in reference 16, the most general approach to estimating transmission network statistical quantities like $\mathcal{P}(M)$ would be to utilize computer simulations of disease transmission on large social contact networks. As was pointed out in section 4, elaborate disease transmission models have been constructed to investigate outbreak dynamics and the effect of control measures for several zoonotics in networks of animal hosts. Other examples of models for transmission networks at the animal-animal, farm-farm, or focus-focus scale are also available[96-98]. These are of interest not solely for the animal-pathogen systems studied, but also for their methodologies, which may be applicable to the host systems associated with our pathogens. Social contact networks are relatively stable but flexible descriptors of reservoirs and transmission nets over time and can easily be stored as reference data. Moreover, it is decidedly more practical to consider collecting field data about the underlying social net, or take advantage of field studies funded through basic epidemiological science programs, than it is to directly gather contact tracing data from an animal outbreak. A network model developed for one disease may apply in large part to others with minor changes in parameter values. Thus, this approach holds considerable promise as an operational way to determine $\mathcal{P}(M)$ for bacterial pathogens relevant to a forensics case.

The next important technical area to be addressed is quantitative descriptions of genetic change during disease transmission and infection, i.e. the sampling distribution $P(\delta|M)$. The genetic distance value δ may be defined in many ways, and a variety of sophisticated metrics have been proposed that take into account not only substitutions, but also insertions and deletions and other types of genetic change[100]. It is important that genetic differences be scored according to a realistic "biological" model of genetic change. For example, a deletion of n adjacent nucleotides is not simply equivalent to n single

nucleotide deletions. Similarly an inverted region should not be scored as a region with a large number of substitutions. Ideally, the value of δ should reflect the number of distinct mutational events that separate two sequences. This may not be unambiguous in some events because there may be more than one possible sequence of events that cause a particular change in the genome, but we can expect that improved understanding of mutation rates will help decide between alternative evolutionary paths in such cases.

A considerable simplification in the calculation of δ arises when the network size is small enough because the total number of genomic changes in any two sequences is then a modest fraction of the number of loci in the genome, and corrections for multiple mutations at a given locus are negligible. Second, for smaller networks it is reasonable to approximate the statistical processes that lead to sequence diversification as stationary[4], which is equivalent to assuming that the mutation rates did not change appreciably during the formation of the network.

Assumptions about stationary mutation rates in bacteria can only be approximately correct and should be treated with caution, because it is known that "mutator" strains can spontaneously appear[58]. These strains have mutation rates orders of magnitude higher than normal because of mutations in replication and repair genes. In any case, we can expect the stationary approximation to be most accurate for sequences that are "closely related" to the consensus sequence for the population in question. The effect of non-stationary effects on the practical accuracy of the inference-on-networks framework can only be established through future experimental studies.

Implicit in this discussion is the assumption that whole genome sequencing will be used to compare case related samples. Given the rate of progress in the capacity and cost of whole genome sequencing this is almost certainly warranted. A simple illustration of the potential for future capacity is shown in figure 12, showing the number of days required to M bacterial isolates at an average coverage of N where $NxM = 10^5$ (E.g. 1000 isolates at 100x coverage). However, error rates are still high enough that the application of

locus-specific allele analysis (e.g. SNP, VNTR, or IS typing assays) will still be required to provide high confidence that an observed mutation is real.

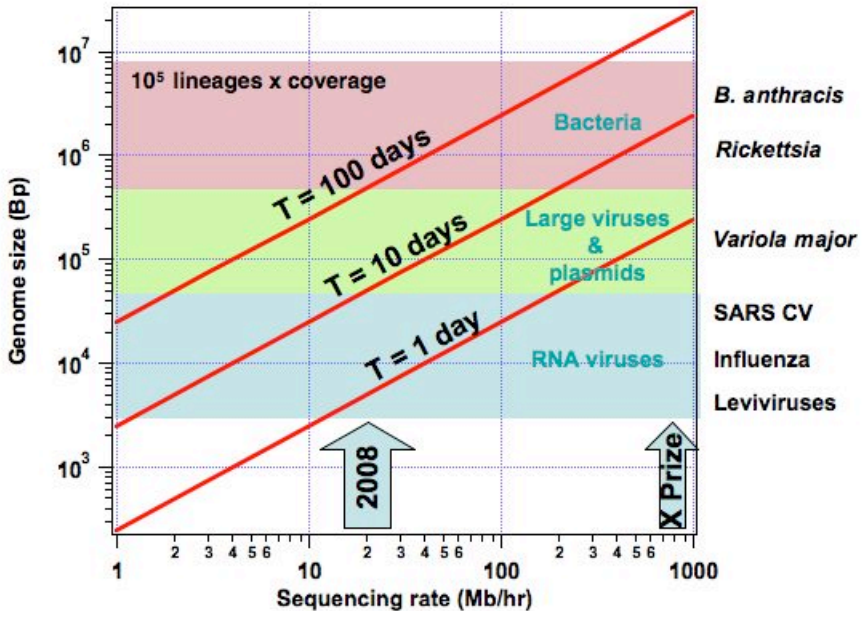

Figure 12. Time required to sequence N genomes of a certain size to a coverage of m per nucleotide where m x N = 10^5.

In addition, even when whole genome sequence data is available it might be advantageous to select a restricted (although possibly large) set of mutational loci to calculate δ because it simplifies the determination of P(δ|M). For example, we might choose a specific set of loci that exhibit neutral evolution, or some particular rate. (The composition of this set might be different for each bacterium.) However, whole genome sequencing would permit unambiguous identification of recombination and re-assortment events, which could confound simplistic distance measures.

Assuming δ is defined, there are several approaches to determining P(δ|M). First, there are purely empirical approaches for determining it from sequence data determined from pairs of nodes with known epidemiological relationships. For any pair of nodes separated by M transmission steps in a completely connected outbreak tree, the observed δ value is a sample from its parent distribution. Unfortunately, a direct approach to determining P(δ|M) by random (or exhaustive) sampling of many infected hosts in an outbreak is

generally impractical for animal outbreaks and foci, because obtaining a complete and accurate transmission tree (so that the relationship between all the nodes is known) is not possible. However, in some cases we might be able to obtain reasonable representations of $P(\delta|M=1)$ and its complement $P(\delta|M>1)$ using empirical data from known transmission pairs. For example, pairs of isolates from known secondary transmission cases in human pneumonic plague or *E. coli* O157 outbreaks could be used for this way.

If δ is a random variable distributed as $P(\delta|M=1)$ for a single transmission step, and each transmission event represents an independent sampling of the genomic distribution in the transmitting host, then δ after M transmission steps is distributed as the sum of M independent random variables each independently distributed as $P(\delta|M=1)$. Thus for larger M we may write:

$$P(\delta|M=M_0>1) = P(\delta|M=1) \otimes P(\delta|M=1) \otimes P(\delta|M=1) \otimes \ldots\ldots..P(\delta|M=1) \qquad (11)$$

Where the right hand side of equation (11) is the M_0-fold auto-convolution of $P(\delta|M=1)$. Hence, it is only necessary to obtain $P(\delta|M=1)$ in order to estimate $P(\delta|M)$ for larger values of M. A reasonably large set of transmission linked pairs of isolates from a single outbreak thus allows us to validate the self-consistency of an entire network model because the empirical $P(\delta|M>1)$ generated by pairwise comparisons among unlinked members of the set can be compared to the $P(\delta|M>1)$ synthesized by combining $P(\delta|M)$ from (11) with $\mathcal{P}(M)$ calculated from the transmission network model for that outbreak:

$$P(\delta|M>1) = \Sigma P(\delta|M)\mathcal{P}(M)/\Sigma\mathcal{P}(M) \qquad (12)$$

where the sums are taken from M = 2 to ∞.

A second option is to derive an estimate of $P(\delta|M=1)$ from laboratory animal passage experiments. Most of the 8 bacteria have laboratory animal models, as shown in table 2. In some cases it is possible to perform passage experiments on actual hosts in veterinary

facilities. However, it is important that the laboratory experiments replicate the important features of the actual host-host transmission process in nature. For example, infection through natural mechanisms ensures that the size of the infecting bolus (e.g. a population bottleneck) mimics the natural process. Mouse transmission models that reproduce the natural processes (flea bite and oral-fecal, respectively) have been demonstrated For *Y. pestis* and *E. coli*, and aerosol infection model has been demonstrated for *Brucella melitensis*, but this area of research is just beginning. Host-host variation in immune response and other selective pressures is another factor that must be considered. If host diversity effects are a selective driver of genetic variation in the natural transmission network, then it should be present in the experimental system.

Table 2. Laboratory animal models for bacterial pathogens.

Bacterium	Laboratory animal models or *hosts*	Natural infection model?
B. anthracis	Mouse Rabbit	No
Y. pestis	*Rat* Mouse	Yes
E. coli	*Cow* Mouse	Yes
Brucella melitensis	*Goat* Mouse	Yes
Burkholderia mallei	*Horse* Mouse	No
Francisella tularensis	*Rabbit* Mouse Rat	No
Burkholderia pseudomallei	Mouse	NA
Clostridium botulinum		NA

In the absence of direct experimental determinations of $P(\delta|M=1)$ a less satisfactory, but potentially useful approach is to perform in-vitro passage experiments to determine mutation rates at the relevant set of loci over which δ is defined, and then calculate $P(\delta|M=1)$ using the mutation rates as parameters. There are several approaches to performing multi-locus, multi-allele simulations of pathogen evolution on transmission

networks, and some of these have already been applied to viral systems, but would need to be extended to include VNTR, IS, and other insertion-deletion loci characteristic of bacteria. A somewhat less rigorous attempt to incorporate *in vitro* VNTR mutation rates into models for genetic inference has been pursued by Keim and co-workers, but this approach has not been subject to experimental validation[6,7].

Mutation rates are locus and sometimes strain specific, depending on genomic context, allele state, and the presence of mutations in key replication and DNA repair enzymes. This makes developing a realistic global model for mutations in even a single bacterial species a complex undertaking. On the other hand, mutation rate experiments on *E. coli* and *Y. pestis* suggest that the mutation rate matrix at VNTR loci can be described by a common model in all bacteria, and it is possible that tractable models for other restricted sets of loci, such IS elements will be feasible. Nonetheless, the practical success of this approach will depend on accumulating much more data than is now available. This, in turn, will necessitate improvements in the rate and accuracy of these experiments.

In vitro mutation rates are determined by assessment of the genetic changes that have occurred after a well-defined number of generations in a bacterial or viral lineage. In order to create a lineage that represents a significant number of generations, repeated passage or prolonged continuous growth is necessary. Therefore precise measurements require repeated passages over extended lengths of time, with even the most rapidly growing bacteria requiring several years to accumulate a significant number of generations. Under these circumstances, passage experiments to find anything but the fastest mutation rates become heroic multi-year exercises.

One consequence is that the types of loci that can be studied in practice are generally restricted to those that have rates greater than 10^{-5} per generation, and important pathogens that have very slow generation times (e.g. M. tuberculosis) are essentially excluded from laboratory study. Moreover, even those rates that have been reported are generally acknowledged to be only rough estimates. In many cases, meaningful measures

of uncertainty cannot be assigned to these estimates because the experimental designs are not statistically robust.

For example, the most comprehensive study of bacterial mutation rates to date is due to Lenski and co-workers, who have examined a set of 12 lineages of E. coli that have been propagated for more than 40,000 generations[6]. Generating this set of samples required more than five continuous years of serial transfer every 48 hours. A number of important observations about mutation and adaptation were extracted from this set of samples, including the finding that several lineages evolved into "mutators" with defective DNA mismatch repair systems resulting in mutation rates ≈ 100 times faster than normal. However, only very crude estimates of the rates of IS element mediated mutations and single nucleotide substitutions could be deduced from this data[6,7].

Sets of passaged samples have also been created for *B. anthracis* and *Y. pestis*. In both cases 4 strains were passaged over 1000 generations in 100 lineages each[8]. Estimates of VNTR mutation rates were made using these samples, but only loci with rates higher than 1×10^{-5} per generation could be characterized. Thus, while the approximate mutation rates of a limited number of loci in these and a few other bacteria have been estimated, many general questions that are fundamental to the inference-on-nets framework remain unanswered:

- How much of the adaptive evolution phenomenology observed in *E. coli* is also exhibited by other bacteria, including pathogens such as *Brucella spp.* and *F. tularensis*? For example, can DNA repair deficient mutators arise at significant rates in these species? The appearance of mutators can change $P(\delta|M)$.

- What is the effect of prolonged periods of starvation or other kinds of stress on the observed mutational spectrum and rate parameters? This may be important for *E. coli* O157, *Burkholderia pseudomallei*, and *Clostridium botulinum*, which have important non-host environmental niches.

- What are the signature mutations that underlie adaptation of wild types to various laboratory growth conditions, and what are their characteristic rates?

Therefore, there is clearly a need to develop practical methods for determining more accurate rate constants for a larger number of genetic loci in a broader range of bacteria, and for greatly reducing the time needed to generate this data.

In appendix 4 we describe time-optimized experimental designs that can improve the speed at which mutation rates are determined by orders of magnitude. In addition they lead to increased precision of the measured rate constants, and permit mutation rates slower than 10^{-5} per generation to be measured accurately. Because the optimized designs require large numbers of replicate lineages to be generated, they require massive multiplexing. This can be accomplished practically through the use of automation, micro-miniaturized bioreactors, and microfluidic sample control systems. An automated serial passage instrument would accelerate a number of fundamental experiments and systematic studies that are difficult to contemplate now because of the time and manual labor that would be required. The discussion in appendix 4 is based on the work of Messenger, Dzenitis and Velsko[101], and the reader is directed to that document for additional details.

Thus, a robust program to integrate bacterial population genetics into microbial forensics requires three major elements:

Theory and modeling - It will be necessary to create individual network models for each pathogen (although some of them might be quite similar), and to consolidate and extend a general model for genome evolution within networks to include VNTRs, IS elements, and other insertion-deletion processes as well as substitutions.

Field and laboratory data collection – While *B. anthracis*, *Y. pestis*, and *E. coli* probably have sufficient published data, additional field sampling and observation will be necessary to estimate the parameters needed to build network models for the other

pathogens. For all 8 pathogens, isolates from epidemics, epizootics and foci or from animal passage experiments with realistic transmission conditions will have to be sequenced to provide data for estimating or validating genetic sampling distributions. A more extensive campaign to determine *in vitro* mutational rate constants would be desirable.

Validation - After the model-building phase, an extensive validation campaign is necessary. One of the key elements of this will be obtaining and sequencing very dense collections of isolates from exemplar outbreaks and foci. For example, the minimum number of isolates required to validate a network genetic distance model is $N_{iso} \approx [\ln(N_{out}]^2$ where N_{out} is the number of infected hosts in the outbreak. The data from these collections can be used to refine the parameters of the models. Note that the inference-on-networks framework leads to a theory-guided experimental program that replaces random, opportunistic collections with deliberate, planned collections to determine parameters and to validate predictions.

6. The way forward

The analysis outlined above provides a very general prescription for microbial forensic source inference that is applicable to many other pathogens in addition to the 8 species considered in this report. Implementing the inference-on-networks approach across even those 8 would be a substantial undertaking, especially for *Clostridium botulinum* and *Burkholderia pseudomallei* where certain basic conceptual issues are not resolved, and the existing empirical knowledge base is still very thin. Thus, this section is concerned with the practical implementation of the inference-on-nets framework within the national microbial forensics program, including how to stage and prioritize work on various pathogens.

Based on the information presented in section 4, the 8 pathogens of interest fall into four classes, based on the complexity of the network description of their foci and outbreaks. This breakdown is summarized in table 3. Class I contains those pathogens that are transmitted host-to-host through "inert foci" that do not support growth of the microbe. We currently place *B. anthracis* in this class, although additional evidence that soil vegetative growth is a significant part of its ecology would obviously change this. Class II contains those pathogens which are maintained in foci whose structures are themselves simple networks. Class III involves foci whose network structures are more complex, for example tick transovarial transmission in *F. tularensis* foci. Finally, Class IV encompasses soil dwelling microbes whose populations are complex metapopulations.

Table 3. Class breakdown of the 8 pathogens of interest

Class	Class property	Pathogen examples*
I	Simple node-to-node network	*B. anthracis*
II	Nodes (foci) are simple networks	*Yersinia pestis* *Escherichia coli* O157/H7 *Burkholderia mallei* Brucella melitensis
III	Nodes (foci) are *complex* networks	*Francisella tularensis*
IV	Nodes (foci) are overlapping metapopulations	*Burkholderia pseudomallei* *Clostridium botulinum*

*Best approximate description for that organism

The "low hanging fruit" are clearly those pathogens in Classes I and II. It is likely that existing architectures for transmission networks will be adequate to model them, and as noted for some, such models already exist, in part. Note, for example that another class I organism is *Mycobacterium tuberculosis*. We can take advantage of the extensive network modeling of TB to guide (*mutatis mutandis*) our approach to anthrax.

Table 4. Status of the 8 pathogens of interest with respect to requirements

Pathogen	Basis for transmission net model	Basis for evolution model	Availability of outbreak isolates for T&V	U.S. foci/outbreaks
B. anthracis	Sufficient	Marginal	Yes	Yes
Y. pestis	Sufficient	Sufficient	Yes	Yes
E. coli	Sufficient	Sufficient	Yes	Yes
Br. melitensis	Marginal	Insufficient	Not known	Yes?
Burk. mallei	Marginal	Sufficient	Not known	No
F. tularensis	Inadequate	Insufficient	Yes	Yes
Burk. pseudom.	Inadequate	Marginal	Not known	Yes
C. botulinum	Inadequate	Insufficient	Not known	Yes

Table 4 summarizes the status of all 8 pathogens with respect to requirements for model building, testing, and validation. A sufficient basis for a transmission net model means that one has already been developed to study outbreak dynamics, or there are analogous existing models that may be applicable. A sufficient basis for an evolution model indicates that VNTR typing system has been developed and mutation rates for the loci have been estimated. (This is a rather narrow basis, but means that the model predictions can be tested against VNTR typing data on field samples.) In cases where the availability of isolates is not known, it is because any isolates that exist are from foreign sources or historical US outbreaks.

Based on this breakdown, a phased development project should be initiated starting with *B. anthracis*, *Y. pestis*, and *E. coli* O157 H7 with the aim of demonstrating and validating the inference-on-networks approach to source identification. Simultaneously, certain basic scientific field experiments and analysis should be initiated for all 8 pathogens to fill in the missing gaps in the modeling basis. Coordination with NIH to optimize DHS

resources applied to basic science studies will be necessary. In later phases, model building, demonstration and validation can then be extended to *Brucella melitensis*, *Burkholderia mallei* and *F. tularensis*, *Burkholderia pseudomallei*, and *Clostridium botulinum*, and other pathogens as threat priorities dictate. A summary of experiments and analytical efforts related to all 8 pathogens is provided in table 5.

To support the creation of a robust system for interpreting microbial genetic data, DHS should also initiate a pilot program in which, in collaboration with the CDC, for selected outbreaks and epizootics, state and local veterinary and/or human health departments are provided with resources to obtain isolates and to have the entire genetic sequence determined from *every* confirmed case of the disease. The genetic sequence data and appropriately source-coded metadata for each isolate is then made available to DHS for testing, validating, and updating the inference-on-networks models for those pathogens.

Table 5. Some suggested experiments

Pathogen	Question	Suggested experiment
B. anthracis	Does *anthracis* have a vegetative phase in soil?	Do successive outbreaks from the same soil focus show a change in genotype?
Y. pestis	Can in-vitro VNTR mutation rates predict P(δ|M=1)?	Compare predictions with direct determination of P(δ|M=1) in mouse model
E. coli	What is P(δ|M=1) in a natural transmission setting?	Perform veterinary lab experiment[50]
Br. melitensis	What are the VNTR mutation rates?	Perform in-vitro passage experiments
Burk. mallei	What is VNTR diversity observed in a natural, large outbreak	Perform typing on isolates from a large outbreak
F. tularensis	How does transovarial transmission affect genetic diversity?	Develop laboratory tick model for *F. tularensis* culture
Burk. pseudom.	What are growth and mutation rates in soil culture?	Develop laboratory soil culture system
C. botulinum	What are growth and mutation rates in soil culture?	Develop laboratory soil culture system

In addition to providing a basis for explicit calculations of probabilities, the inference-on-networks framework has a natural analogue in mitochondrial and Y chromosome DNA forensics. This transparent relationship to accepted DNA forensics is an advantage that should enhance its plausibility to the legal and policy communities. However,

developing the scientific infrastructure to implement this framework will require a concerted effort by many laboratories involving expertise not currently involved in the national microbial forensics program. Hence, an important first step will be developing a consensus in favor of this direction within the wider microbial forensic community. Therefore, we highly recommend that DHS promote an open forum for "socializing" the concept and methodology among scientists, legal and policy experts, and program managers, perhaps through the Banbury meetings or another symposium.

Notes and references

1. World at Risk: The Report of the Commission on the Prevention of WMD Proliferation and Terrorism, December, 2008.

2. Harismendy, et. al., "Evaluation of Next generation sequencing platforms for population targeted sequencing studies", Genome Biology 2009, **10**:R32.

3. This definition is consistent with standard usage for mammalian DNA. See Gillespie, Population Genetics, 2nd Ed. (Johns Hopkins University Press, 2004).

4. Song YS, et. al., "Average Probability that a "Cold Hit" in a DNA Database Search Results in an Erroneous Attribution", J. Forensic Sci. **54**:22-27, (2009)

5. This is, of course, one of the original rationales for the current Select Agent system. See: Public Health Security and Bioterrorism Preparedness and Response Act of 2002 H.R. Conference Report 107-481 2002.

6. Girard JM, et. al. "Differential plague-transmission dynamics determine *Yersina pestis* population genetic structure on local, regional, and global scales Proc. Nat. Acad. Sci. **101**, 8408-8413.

7. Lowell, et. al. "Identifying Sources of Human Exposure to Plague" J. Clin. Microbiol. **43**, 650-656 (2005)

8. Van Ert MN, et. al. "Global Genetic Population Structure of *Bacillus anthracis*" PLoS One May 2007 e461.

9. Tuanyok A, et. al. , "A Horizontal Gene Transfer Event Defines Two Distinct Groups within *Burkholderia pseudomallei* That Have Dissimilar Geographic Distributions", J. Bacteriol. **189**, 9044-9049 (2007)

10. Achtman M, "Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Lineages", Ann. Rev. Microbiol. 2008 **62**:53-70.

11. Vogler AJ, et. al. "Phylogeography of *Francisella tularensis*: Global Expansion of a Highly Fit Clone" J. Bacteriol. **191**: 2474-2484 (2009).

12. Foster, JT, et. al. "Whole-Genome-Based Phylogeny and Divergence of the Genus *Brucella*", J. Bacteriol. **191**:2864-2870 (2009)

13. Simonson TS, et. al., "*Bacillus anthracis* in China and its relationship to worldwide lineages" BMC Microbiology 2009, **9**:71

14. Keim PS and DM Wagner, "Humans and evolutionary and ecological forces shaped the phylogeography of recently emerged diseases", Nature Reviews *Microbiology*, Vol. **7**, 813-816, (2009).

15. Pearson T, et. al., "Phylogenetic understanding of clonal populations in an era of whole genome sequencing", Infection, Genetics and Evolution **9**:1010-1019, (2009).

16. Velsko SP, J Allen and C. Cunningham, "A Statistical Framework for Microbial Source Attribution, Part 1: Forensic Inferences on Disease Transmission Networks", Lawrence Livermore National Laboratory Report LLNL-TR-414337, April 30, 2009.

17. Wilson MR, et. al., "Validation of mitochondrial DNA seqencing for forensic casework analysis", Int. J. Legal Med. (1995) **108**:68-74.

18. Budowle, B, et. al. "Texas Population Structure and its impact on estimating the rarity of Y STR haplotypes from DNA evidence", J. Forensic Sci. 2009; **54**:1016-21.

19. Lenski RE and Keim P, "Population Genetics of Bacteria in a Forensic Context", in Microbial Forensics, R. Breeze, B. Budowle, and S. Schutzer, eds. (Elsevier Academic Press, Amsterdam, 2005).

20. Keim, P., Van Ert, M.N., Pearson, T., Vogler, A.J., Huynh, L.Y., Wagner, D.M., 2004. Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales. Infect. Genet. Evol. 4 (3), 205–213.

21. Auerbach RK, et. al., Yersina pestis Evolution on a Small Timescale: Comparison of Whole Genome Sequences from North America", PloS One August 2007 **8**:e770.

22. Vogler AJ, et. al., "Mutations, mutation rates and evolution at the hypervariable VNTR loci of *Yersinia pestis*", Mutation Research 616 (2007) 145–158

23. Leitner T, "Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis", Proc. Nat. Acad. Sci. USA, 1996; 93(20):10864-9.

24. Balding, DJ, Weight-of-evidence for Forensic DNA Profiles, (J. Wiley Statistics in Practice series, 2006).

25. Foley SL, et. al. "Molecular typing methodologies for microbial source tracking and epidemiological investigations of Gram-negative bacterial foodborne pathogens", Infection, Genetics and Evolution **9**:430-440, (2009)

26. Hyytla-Trees E, et. al. "Second Generation Subtyping: A Proposed PulseNet Protocol for Multiple-LocusVariable-Number Tandem Repeat Analysis of Shiga Toxin-Producing Escherichia coli O157:H7", Foodborne Pathogens and Disease 3:117-132 (2006).

27. US Department of Agriculture, "Epizootology and Ecology of Anthrax".

28. Blackburn JK, et. al., "Modeling and Geographic Distribution of Bacillus anthracis, the Causative Agent of Anthrax Disease, for the Contiguous United States using Predictive Ecological Niche Modeling", Am. J. Trop. Med. Hyg. 2007; **77**:1103-1110.

29. Van Ness G, and Stein CD, "Soils of the United States Favorable for Anthrax", J.Am. Vet. Med. Assoc. 1956; **128**: 7-12.

30. Van Ness GB, "Ecology of Anthrax", Science 172:1303-1307, (1971).

31. Schuch R, and Fischetti VA, "The Secret Life of the Anthrax Agent *Bacillus anthracis*: Bacteriophage-mediated ecological adaptations", PLoS One 4(8):e6532 (2009).

32. Smith KL, et. al., "Meso-scale ecology of anthrax in Southern Africa: a pilot study of diversity and clustering", J. Appl. Microbiol. 1999; 87:204-207.

33. Kenefic LJ, et. al., "Pre Columbian Origins for North American Anthrax", PLoS One 4(3):e4813 (2009).

34. Stein CD, "The History and Distribution of Anthrax in Livestock in the United States", Veterinary Medicine 40 340-350 (1945).

35. Kenefic LJ, et. al., "High resolution genotyping of *Bacillus anthracis* outbreak strains using four highly mutable single nucleotide repeat markers", Lett. Appl. Microbiol. 2008; **46**:600-603.

36. Wimsatt J and Biggins DE, "A review of plague persistence with a special emphasis on fleas", J. Vector Borne Dis. 6:85-99 (2009).

37. Webb CT, et. al. "Classic flea-borne transmission does not drive plague epizootics in prairie dogs", Proc. Nat. Acad. Sci. USA **103**:6236-6241 (2005).

38. Eisen RJ and Gage KL, "Adaptive strategies of Yersinia pestis to persist during inter-epizootic and epizootic periods", Vet. Res. 2009; **40**:01.

39. Adjemian JZ, et. al., "Initiation and spread of traveling waves of plague, *Yersinia pestis*, in the Western United States". Am. J. Trop. Med. Hyg. 2007; **76**:365-375.

40. Davis S, et. al. "The abundance threshold for plague as a critical percolation phenomenon", Nature 2008; **454**:634-637.

41. Keeling MJ and Gilligan CA, "Bubonic plague: a metapopulation model of a zoonosis", Proc. Roy. Soc. London 2000; **267**:2219-2230.

42. Adjemian et. al. recognized that network models can easily explain apparent random large translocations in disease which are difficult to explain in diffusion type models without sudden and artificial increases in the apparent diffusion rate.

43. Foley JE, et. al., "Modeling plague persistence in host-vector communities in California", J. Wildlife Dis. 2007; **43**:408-424.

44. Nishiura H, et. al., "Transmission potential of primary pneumonic plague: time inhomogeneous evaluation based on historical documents of the transmission network", J. Epidemiol. Community Health 2006; **60**:640–645.

45. Gani R and Leach S, "Epidemiologic Determinants for Modeling Pneumonic Plague Outbreaks", Emerging Infectious Diseases Vol. 10, No. 4, April 2004.

46. Chain PSG, et. al., "Complete Genome Sequence of *Yersinia pestis* Strains Antiqua and Nepal516: Evidence of Gene Reduction in an Emerging Pathogen", J. Bacteriol. 2006; **188**:4453-4463.

47. Hancock D, et. al., "Control of VTEC in the animal reservoir", International Journal of Food Microbiology 2001; **66**:71–78.

48. Snedecker KG, et. al., "Primary and secondary cases in Escherichia coli O157 outbreaks: a statistical analysis", BMC Infectious Diseases 2009, **9**:144.

49. Turner J, et. al., "A network model of E. coli O157 transmission within a typical UK dairy herd: The effect of heterogeneity and clustering on the prevalence of infection", Journal of Theoretical Biology 254 (2008) 45– 54.

50. Davis MA, et. al. "Escherichia coli O157:H7 in Environments of Culture-Positive Cattle", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Nov. 2005, p. 6816–6822.

51. Woolhouse MEJ, et. al., "Epidemiological implications of the contact network structure for cattle farms and the 80-20 rule", Biology Letters 2005; **1**:350-352.

52. Brennan ML, "Direct and indirect contacts between cattle farms in north-west England", Preventive Veterinary Medicine **84** (2008) 242–260.

53. Heath FM, et. al., "Construction of networks with intrinsic temporal structure from UK cattle movement data", BMC Veterinary Research 2008, **4**:11.

54. Vernon MC and Keeling MJ, "Representing the UK's cattle herd as static and dynamic networks", Proc. R. Soc. B (2009) 276, 469–476.

55. Noller AC, et. al., "Multilocus Variable-Number Tandem Repeat Analysis Distinguishes Outbreak and Sporadic Escherichia coli O157:H7 Isolates", JOURNAL OF CLINICAL MICROBIOLOGY, Dec. 2003; **41**:5389–5397.

56. Noller AC, et. al., "Locus-Specific Mutational Events in a Multilocus Variable-Number Tandem Repeat Analysis of Escherichia coli O157:H7", JOURNAL OF CLINICAL MICROBIOLOGY, Feb. 2006; **44**:374–377.

57. Vogler AJ, et. al., "Effect of Repeat Copy Number on Variable-Number Tandem Repeat Mutations in Escherichia coli O157:H7", JOURNAL OF BACTERIOLOGY, June 2006; **188**:4253–4263.

58. Woods R, et. al., "Tests of parallel molecular evolution in a long term experiment with *Escherichia coli*", Proc. Nat. Acad. Sci. USA 2006; 103:9107-9112.

59. Barrick, et. al., "Genome evolution and adaptation in a long-term experiment with Escherichia coli", Nature 2009; **461**:1243-1247.

60. Pappas G., "The new global map of human brucellosis", http://infection.thelancet.com **Vol 6** February 2006.

61. Boschiroli M-L, et. al., "Brucellosis: a worldwide zoonosis", Current Opinion in Microbiology 2001, **4**:58–64.

62. Yamamoto T, et. al., "Evaluation of surveillance strategies for bovine brucellosis in Japan using a simulation model", Preventive Veterinary Medicine 2008; **86**:57-74.

63. Zinstag J, et. al., "A model of animal–human brucellosis transmission in Mongolia", Preventive Veterinary Medicine **69** (2005) 77–95.

64. Whatmore AM, "Current understanding of the genetic diversity of Brucella, an expanding genus of zoonotic pathogens", Infection, Genetics and Evolution xxx (2009) xxx–xxx (In press).

65. Bricker, B.J., Ewalt, D.R., Halling, S.M.,. "*Brucella* 'HOOF-Prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs)", BMC Microbiol 2003; **3**: 15.

66. Le Fle` che, P., Jacques, I., Grayon, M., Al Dahouk, S., Bouchon, P., Denoeud, F., No¨ ckler, K., Neubauer, H., Guilloteau, L.A., Vergnaud, G., "Evaluation and

selection of tandem repeat loci for a *Brucella* MLVA typing assay. BMC Microbiol. 2006; **6**: 9.

67. Whatmore, A.M., Shankster, S.J., Perrett, L.L., Murphy, T.J., Brew, S.D., Thirlwall, R.E., Cutler, S.J., Macmillan, A.P., "Identification and characterization of variable-number tandem-repeat markers for typing of Brucella spp.", J. Clin. Microbiol. 2006; **44**:1982–1993.

68. Tiller RV, et. al., "Comparison of Two Multiple-Locus Variable-Number Tandem-Repeat Analysis Methods for Molecular Strain Typing of Human *Brucella melitensis* Isolates from the Middle East", JOURNAL OF CLINICAL MICROBIOLOGY, July 2009; **47**: 2226–2231.

69. Glanders, Iowa State University Center for Food Security and Public Health fact sheet.

70. Derbyshire BJ, "The eradication of glanders in Canada", Can Vet J Volume **43**, September 2002

71. Heitor F, "Social relationships in a herd of Sorraia horses", Behavioural Processes **73** (2006) 231–239

72. Houpt KA, "Investigating equine ingestive, maternal, and sexual behavior in the field and in the laboratory", J. Anim. Sci. 1991; **69**:4161-4166.

73. Baguelin M, et. al., "Control of equine influenza: scenario testing using a realistic metapopulation model of spread", R Soc Interface. 2009 Apr 1. [Epub ahead of print].

74. Romero CM, et. al., "Genome sequence alterations detected upon passage of *Burkholderia mallei* ATCC 23344 in culture and in mammalian hosts", BMC Genomics 2006, **7**:228.

75. Song H, et. al., "Simple Sequence Repeat (SSR)-Based Gene Diversity in *Burkholderia pseudomallei* and *Burkholderia mallei*", Mol. Cells OT, 237-241, February 28, 2009.

76. U'Ren JM, et. al., "Tandem repeat regions within the *Burkholderia pseudomallei* genome and their application for high resolution genotyping", BMC Microbiology 2007, **7**:23

77. Staples JE, et. al., "Epidemiologic and Molecular Analysis of Human Tularemia, United States, 1964–2004", Emerging Infectious Diseases **Vol. 12**, No. 7, July 2006.

78. Niebylski ML, et al., "Characterization of an Endosymbiont Infecting Wood Ticks, Dermacentor andersoni, as a Member of the Genus *Francisella*", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 1997; **63**: 3933–3940; See also: Baldridge GD, "Transovarial transmission of Francisella-like endosymbionts and Anaplasma phagocytophilum variants in Dermacentor albipictus (Acari: Ixodidae). J Med Entomol. 2009 May;46(3):625-32

79. Goethert HK, et. al., (a)"Genotypic Diversity of *Francisella tularensis* Infecting Dermacentor variabilis Ticks on Martha's Vineyard, Massachusetts", JOURNAL OF CLINICAL MICROBIOLOGY, Nov. 2004; **42**: 4968–4973; (b) "A metapopulation structure for perpetuation of *Francisella tularensis* BMC Microbiology 2009, **9**:147.

80. Peterson JM, et. al., "Multiple *Francisella Tularensis* Subspecies and Clades, Tularemia Outbreak, Utah", Emerging Infectious Diseases **Vol. 14**, No. 12, December 2008.

81. Vogler AJ, et. al., "An optimized, multiplexed multi-locus variable-number tandem repeat analysis system for genotyping Francisella tularensis", Letters in Applied Microbiology Sept. 2008.

82. Pandya GA, et. al., "Whole genome single nucleotide polymorphism based phylogeny of *Francisella tularensis* and its application to the development of a strain typing assay", BMC Microbiology 2009, 9:213.

83. Vogler AJ, et. as., "Phylogeography of Francisella tularensis: Global Expansion of a Highly Fit Clone", JOURNAL OF BACTERIOLOGY, Apr. 2009; **191**: 2474–2484.

84. Johansson A, et. al., "Extensive Allelic Variation among *Francisella tularensis* Strains in a Short-Sequence Tandem Repeat Region", Journal of Clinical Microbiology **39**: Sept. 2001, p. 3140–3146.

85. Gurcan S, et. al., "Characteristics of the Turkish Isolates of *Francisella tularensis*" Jpn. J. Inf. Dis. **61**: Sept. 2001, p. 3140–3146.

86. Currie BJ, "Advances and remaining uncertainties in the epidemiology of *Burkholderia pseudomallei* and melioidosis" Transactions of the Royal Society of Tropical Medicine and Hygiene (2008) **102**: 225—227.

87. Dance DAB, "Ecology of *Burkholderia pseudomallei* and the interactions between environmental *Burkholderia* spp. And human–animal hosts", Acta Tropica **74** (2000) 159–168.

88. Currie BJ, et. al., "A Cluster of Melioidosis Cases from an Endemic Region is Clonal and is Linked to the Water Supply Using Molecular Typing of Burkholderia pseudomallei Isolates", Am. J. Trop. Med. Hyg. 2001; 65:177-179.

89. Chantratita N, et. al.,"Genetic Diversity and Microevolution of *Burkholderia pseudomallei* in the Environment", PLoS Neglected Tropical Diseases 2:e182.

90. U'Ren JN, et. al., "Fine-Scale Genetic Diversity among *Burkholderia pseudomallei* Soil Isolates in Northeast Thailand", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Oct. 2007; 73: 6678–6681.

91. Pearson T, et. al., "VNTR analysis of selected outbreaks of *Burkholderia pseudomallei* in Australia", Infection, Genetics and Evolution **7** (2007) 416–423.

92. Holdeman LV, "The Ecology and Natural History of *Clostridium botulinum*", J. Wildlife Dis. 1970; 6:205-210.

93. Caya JG, et. al., "*Clostridium botulinum* and the Clinical Laboratorian", Arch. Pathol. Lab. Med. 2004; 128:653-662.

94. Luquez C, et. al., Distribution of Botulinum Toxin-Producing *Clostridia* in Soils of Argentina", APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 2005; 71: 4137–4139.

95. Hill KA, et. al., "Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and Clostridium butyricum type E strains", BMC Biology 2009; **7**:66.

96. Ayyalasomayajula S, "A Network Model of H5N1 Avian Influenza Transmission Dynamics in Domestic Cats", Zoonoses Public Health. **55** (2008) 497–506.

97. Craft, ME et. al. "Distinguishing epidemic waves from disease spillover in a wildlife population", Proc Roy Soc B Online, 25 February 2009. (Contact model for wild animals).

98. Xiao Y, et. al., "Pair approximations and the inclusion of indirect transmission: Theory and application to between-farm transmission of Salmonella", Journal of Theoretical Biology **244** (2007) 532–540.

99. Bohm M, et. al., "Dynamic interactions among badgers: implications for sociality and disease transmission" Journal of Animal Ecology 2008; 77: 735–745.

100. Kalinowski ST, "Evolutionary and statistical properties of three genetic distances", Molecular Ecology (2002)11:1263–1273.

101. Messenger S, Dzenitis J, and Velsko S., "Rapid determination of spontaneous mutation rates in bacteria and DNA viruses", Lawrence Livermore National Laboratory Report, October 15, 2005.

Appendix 1.
"Rare strains", typing system resolution, and a genetic "match probability" for microbial forensics

In this appendix we derive a probabilistic measure of strain "rarity" and show that the perception of how rare a strain is depends on the resolution of the strain typing system. This simple theory also provides a straightforward definition of "match probability" in the context of microbial forensics, because it answers the question: if we compare the genetic sequences of two isolates obtained from two arbitrary infected hosts, what is the probability that they are identical?

As a zoonotic pathogen spreads geographically, it creates a network of infected animal hosts whose size expands as[1]:

$$N_{hosts} \approx (\langle k \rangle^{G+1} - 1)/(\langle k \rangle - 1),$$

where G is the number of generations of host-host transmission, and $\langle k \rangle$ is the average number of secondary infections caused by an infected host. For the discussion in this appendix, genetic sequences that define the diversity of the pathogen over the transmission network are taken to be the consensus sequences of the pathogen isolates obtained from hosts in this network. The *rarity* of a genotype can be unambiguously defined as the probability of finding that (consensus) genotype in an isolate sampled from an arbitrary host from the network.

For simplicity, we will consider the clonal expansion of a pathogen, and ignore the possibility of genetic exchange. Assume that an isolate taken from an infected host is characterized by a consensus genetic sequence with G mutable loci, and that each mutable locus j has a mutation rate γ_j per generation. Assume that we have a typing system that examines L out of the G possible loci. Following standard practice, we will define a strain to be one with a certain set of allele states at the L loci. The probability that we will see no change in the L loci (i.e. the bacterium is the "same strain") when two isolates are separated by N_{gen} generations is[2]

$$P_0(L,N_{gen}) = e^{-\Gamma_L N_{gen}}$$

where the rate of genomic change associated with the system of L typing loci is:

$$\Gamma_L = \sum_{j=1}^{L} \gamma_j$$

On the other hand, if we look at the entire sequence of G loci, the probability of observing the same sequence after N generations is

$$P_0(G,N_{gen}) = e^{-\Gamma_G N_{gen}}$$

where the rate of change associated with the whole genome is:

$$\Gamma_G = \sum_{j=1}^{G} \gamma_j$$

As discussed in reference A1.3, every typing system captures only some fraction of the total genomic mutation rate $\alpha_L = \Gamma_L/\Gamma_G << 1$ and Γ_G is approximately 0.005 per generation, when the entire mutational spectrum is considered. A typical typing system like MLVA captures about 20% of the genomic rate, i.e. $\Gamma_L \approx 0.001$ gen^{-1}.

Without loss of generality, we may suppose that a reference genotype denoted "genotype 0" is associated with the (historical) index host in the network. What is the probability of observing genotype 0 in an isolate obtained from an arbitrary host in the network? If the pathogen undergoes G_{host} generations of expansion in an infected host prior to transmission to the next host, and we consider an isolate from a host that is g transmission generations away from the index host, then the total number of generations separating the index host isolate from the isolate in question is

$$N_{gen} = G_{host} \bullet g.$$

Let the joint probability of observing genotype 0 in a host that is also separated from the index host by g generations in a network of $\mathcal{G}$ total generations of transmission be denoted $P(0,g|\mathcal{G})$. By the rules of conditional probability we can write:

$$P(0,g|\mathcal{G}) = P(0|g,\mathcal{G})P(g|\mathcal{G}),$$

where $P(0|g,\mathcal{G})$ is the probability of observing genotype 0 *given* that the host is separated from the index host by g transmission generations, and $P(g|\mathcal{G})$ is the probability that an arbitrary host chosen from the network will be separated from the index by g generations. Referring to the discussion above, we can write:

$$P(0|g,\mathcal{G}) = \exp(-\Gamma_J \bullet \mathcal{G}_{host} \bullet g),$$

where J is either G or L, depending on whether we are using the whole genome sequence or the typing system. $P(g|\mathcal{G})$ is simply the fraction of infected hosts at generation g out of the total number of hosts in the network:

$$P(g|\mathcal{G}) = \langle k \rangle^{\mathcal{G}} / N_{hosts}$$

The probability of observing genotype 0 is then the sum of the joint probability $P(0,g|\mathcal{G})$ over all generations g:

$$\mathbf{P(0|\mathcal{G}) = \sum_{g=0}^{\mathcal{G}} P(0|g,\mathcal{G}) \bullet P(g|\mathcal{G})}$$

which leads to

$$P(0|\mathcal{G}) = \frac{1 - \langle k \rangle}{1 - \langle k \rangle^{\mathcal{G}+1}} \frac{1 - (\langle k \rangle e^{-\Gamma_J \mathcal{G}_{host}})^{\mathcal{G}+1}}{1 - (\langle k \rangle e^{-\Gamma_J \mathcal{G}_{host}})}$$

Figure A1.1 displays $P(0|\mathcal{G})$ as a function of the infected host population size for a transmission network with $\langle k \rangle = 1.1$, assuming $\mathcal{G}_{host} = 20$ generations, $\Gamma_G = 0.005$ gen^{-1} and $\Gamma_L = 0.001$ gen^{-1} (representing a typical MLVA typing system.) Note that when "genotype 0" is defined by the typing system, the probability of finding it in an arbitrary isolate is greater than 10%, even for a very large network of infected hosts. In contrast, when the whole (consensus) genome is used to define the genotype, the probability is less than 1% even for a small-sized outbreak network. Thus, when whole genome sequencing is used as the genotyping method, every "strain" will appear to be rare.

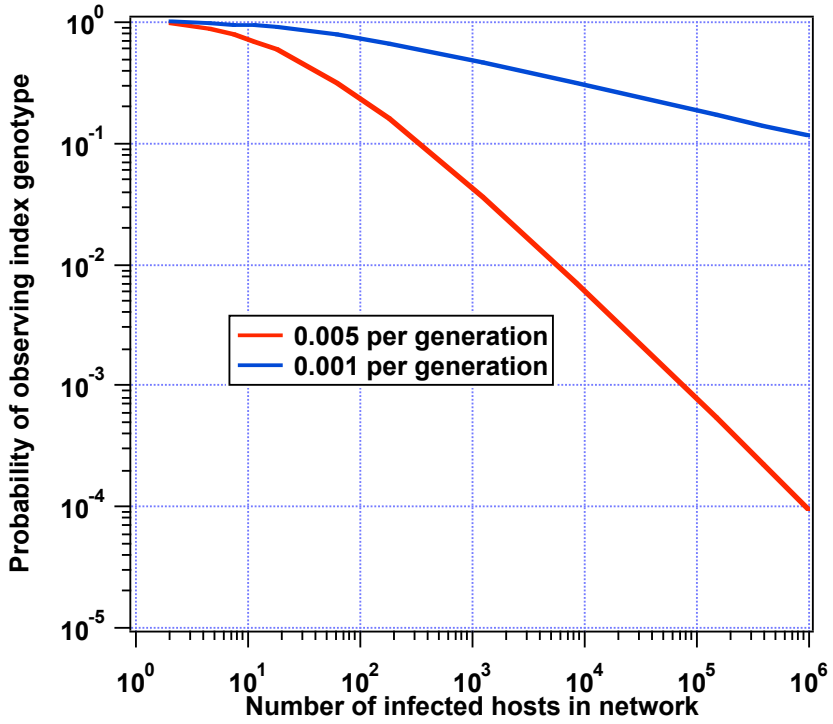

Figure A1.1 Effect of typing system resolution on estimates of strain "rarity".

Another important observation is that the "rarity" of a strain will depend on the average connectivity of the transmission network. Figure A1.2 shows how $P(0|\mathcal{G})$ varies with

<k> for fixed values of Γ and G_{host}. This dependence is simply explained by the fact that when <k> is near 1, it takes many generations of host-host transmission to create a large number of infected hosts. Thus, an arbitrary pair of isolates is more likely to be separated by long chains of what are, in effect, bottlenecking serial transfers. Conversely, when <k> >> 1, only a few generations of host-host transmission are required to generate the same number of infected hosts. Thus, pairs of isolates tend to be separated by only short serial transfer chains, and the probability of observing two isolates with the same genotype is concomitantly higher.

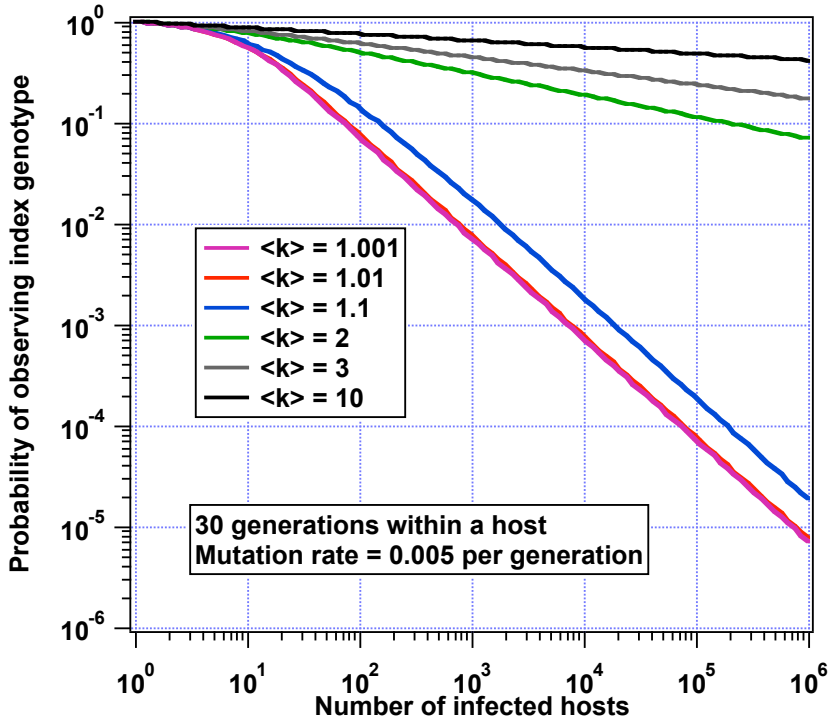

Figure A1.2 Effect of the average number of secondary infections caused by an infected host on the "rarity" of a genotype sampled from a network of infections.

One consequence of this analysis is that it is easy to see why outbreak structure is a critical parameter for making estimates of the "match probability". Figure A1.3 illustrates two extreme types of outbreak topology. Example A is typical of food outbreaks, where bacteria from a single infected farm animal (or food worker) can contaminate a batch of a commercial food product and infect a large number of people. With a low incidence of human-human secondary infections, nearly every human isolate

will be related to the contaminant by a single transmission step. Example B is representative of an outbreak of a disease with low infectivity, where no "superspreader" events take place[4]. Most isolates are related to the index case by relatively long chains of transmission. Some disease outbreaks exhibit significant numbers of "superpreader" events, where one host can infect a large number of other hosts, increasing the value of <k>, and therefore lie between the limits represented by A and B.

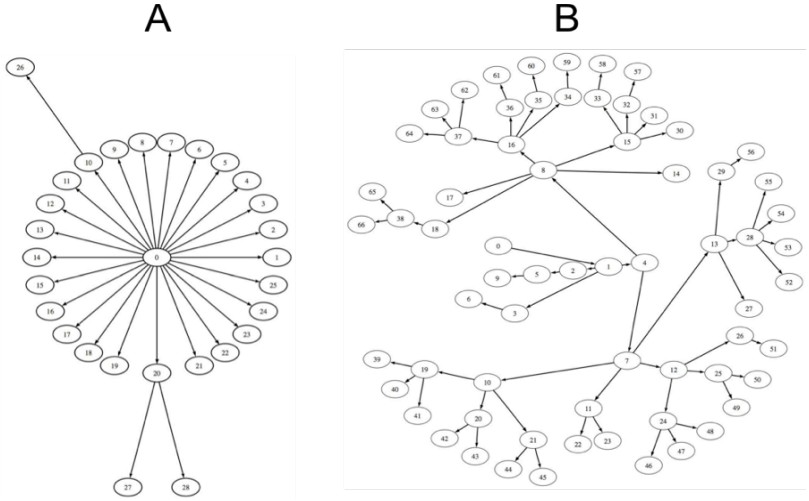

Figure A1.3. Two transmission networks with different <k> values. A: a network representative of a food contamination case, <k> ≈ N, the number of infectees; B: a network for a disease with low infectivity <k> ≈ 2.

Finally, it must be pointed out that the theory outlined above is approximate, since it assumes a simplified description of the disease transmission process. First, the initial equation in this appendix relating the size of the outbreak N_{hosts} and the number of generations of infection $\mathcal{G}$ is only true on average. Actual disease networks are stochastic samples from a process described by a probability distribution for the number of new hosts infected by an infected host. The random process that generates each outbreak transmission tree leads to a distribution of trees with different $\mathcal{G}$ and N_{hosts} values. If our only information is the size of the outbreak, then $\mathcal{G}$ is uncertain, and the outbreak is characterized by a probability distribution $P(\mathcal{G}|N_{hosts})$. A more exact calculation of $P(0|N_{hosts})$ must account for this. Secondly, we assume that outbreaks are stationary –

that is, their statistical properties are independent of time, or the number of generations of transmission. All real outbreaks evolve as the fraction of remaining susceptible hosts decreases, and the ability to generate new infections grows less probable. This will affect the network topology, and modify the relationship between $\mathcal{G}$ and N_{hosts}. Thus, a more realistic SIR model for outbreaks is necessary to accurately gauge $P(0|N_{hosts})$.

References for Appendix 1.

A1.1. This formula is based on the observation by Newman that in a network characterized by a mean coordination number z, the average number of neighbors a distance d from any node is z^d. To obtain the total number of neighbors up to a distance D, simply sum the geometric series to obtain $(1-z^{D+1})/(1 - z)$. M.E.J. Newman, "Models of the Small World – A Review", J. Stat. Phys. 101, 819 – 841, (2000).

A1.2. Gillespi, JH, Population Genetics 2nd Ed. (Johns Hopkins University Press, Baltimore, 2004.) Ch. 2.

A1.3.Velsko, SP, "Resolution in forensic microbial genotyping", Lawrence Livermore National Laboratory Report, UCRL-TR-215303, August 6, 2005.

A1.4. Lipsitch M, et. al. Transmission Dynamics and Control of Severe Acute Respiratory Syndrome", Science **300**, 1966, (2003).

Appendix 2.

The analogy between bacterial population genetics and mtDNA and Y chromosome population genetics

If we consider the human population as all humans who ever existed, then it can be looked at as two separable, but interlocking networks, one in which all the nodes are male and the other all female. The male network is determined by father-son lines of descent with inheritance of the Y chromosome, while the female network is determined by mother-daughter lines of descent with inheritance of the mitochondrial DNA (mtDNA). Figure A2.1 illustrates these networks by showing a portion of the genealogy of British royal descendents, beginning with Albert and Victoria. Each male carries a (somatic) population of Y chromosomes derived from a single Y chromosome inherited from his father. Each female carries a population of mtDNA molecules derived from a small set of mtDNA molecules inherited from her mother. The sequences of Y and mitochondrial DNA associated with each node are typically consensus sequences obtained from somatic cell samples.

For each generation of father-son or mother-daughter transfer, there is a probability that one or more changes in the consensus sequence will occur. These mutation rates have estimated by several authors. For example, Xue, et. al. measured the substitutional mutation rate of Y DNA to be 3×10^{-8} per nucleotide per generation for a 10 Mb region that excluded gaps, repetitive sequences and palindromes[1]. Decker et. al. performed a study of Y-STR mutations at 17 loci and reported data for 389 father-son pairs[2]. A total of 23 single locus mutation events and 1 two-locus mutational event were observed. This data fits well to a Poisson distribution with $\gamma = 0.06$ per generation. Several groups have measured the rate of base substitutions in the HVRI and HVRII regions of mtDNA using persons linked by known genealogies. Combining the results of studies quoted in a recent paper by Sigurðardo´ttir et. al.[3], the substitution rate is roughly $\approx 1 \times 10^{-2}$ gen^{-1}. Sigurðardo´ttir, et. al. also estimate that the mutation rate associated with insertions and

deletions from a poly C tract within HVRII is also nearly 1 x 10^{-2} per generation. Table A2.1 summarizes some order-of-magnitude estimates of mutation rates for mitochondrial and Y chromosome DNA and compares them to data for bacteria and viruses in host-host infection events.

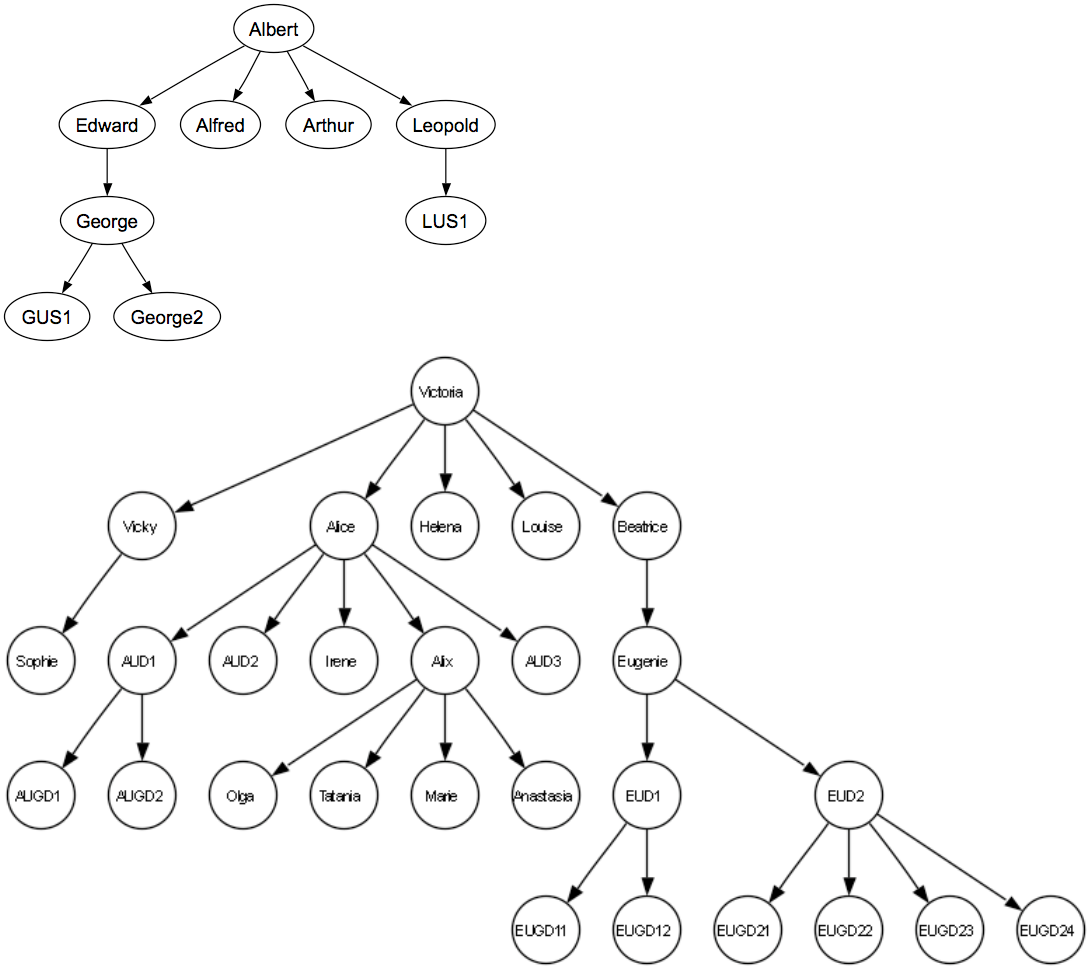

Figure A2.1 Transmission networks for Y DNA and mtDNA based on the descendents of Prince Albert (3 generations) and Queen Victoria (4 generations). UD = un-named daughter, US = un-named son, UGD = un-named grand-daughter.

Similarly, the network topology determines the prior probability that two randomly chosen males or females will be connected by a pedigree with a given number of generations. For human networks, the Galton-Watson process, which posits a probability distribution for the number of offspring of a given node, is a classic model that can also be used to describe transmission networks for microbial outbreaks. There is thus a natural analogy between bacterial populations in infected hosts and the populations of Y

or mitochondrial DNA in humans. The network framework described in reference xx and in section 3 of this report can be used *mutatis mutandis* in both human and microbial cases. It should be noted, however, that there are some subtle differences between these two applications. For example, haplotype frequencies in traditional mtDNA and Y-STR databases are based on a sampling population of living humans, whereas network probabilities consider *all* nodes past and present. It is also true that much more accurate information about the relationships between known network nodes usually exists when using mtDNA or Y-DNA for determining familial relationships than is the case when microbial DNA is used to identify relationships between isolates.

Table A2.1 Estimated mutation rates for viruses, bacteria, Y-DNA and mtDNA.

Organism	Genomic region and haplotype system	Transmission event	Average # of changes in haplotype per transmission event
Virus	Whole genome substitutions	Host - host	≈ 1
Bacteria	Whole genome substitutions	Host - host	≈ 0.01
Bacteria,	Whole genome MLVA system	Host - host	≈ 0.03
Y-DNA	10 Mbp Euchromatic region substitutions	Father - son	≈ 0.3
Y-DNA	17 locus STR system	Father - son	0.03
mtDNA	HVRI and HVRII substitutions	Mother - daughter	0.01
mtDNA	HVRII poly C tract	Mother - daughter	0.01

References for Appendix 2.

1. Xue, Y. et. al., Curr Biol. (2009); 19:1453-1457.

2. Decker AE, et. al., Forensic Science International: Genetics 2 (2008) e31–e35

3. Sigurðardo´ttir et. al., Am. J. Hum. Genet. 66:1599–1609, 2000

Appendix 3.

The microbial "paternity equation"

As a pathogen population propagates along the branches of an outbreak transmission tree, the process of genetic change is a stochastic process that can be characterized by a distribution function describing the probability of observing changes in the consensus sequence after M steps along a chain of infected nodes. The most general form for this distribution for a single step, M = 1, between any two nodes (here denoted 1 and 2 respectively), is:

$$P(S_1,S_2|M=1,\tau, t_1,t_2)$$

where S_1 and S_2 represent the consensus sequences of the microbial populations in each of the two nodes, τ represents the time between infection of node 1 and the transmission event between 1 and 2. The parameters t_1 and t_2 represent the time intervals between infection of each node and the time when isolates are obtained from each of them. (We assume that an isolate represents a sample of a node's population that is "frozen in time" with respect to the course of the infection. If isolates are subjected to additional *in vitro* culture or animal passages, we may consider the passaged samples new isolates associated with nodes that represent the populations of the pathogen in the culture vessel or laboratory animal.)

Clearly, inferences about the relationship between two nodes implied by sequences S_1 and S_2 are based on some quantitative comparison between S_1 and S_2. This quantitative comparison metric is some numerical function of the two sequences. In this derivation, we will assume that the comparison metric is a single scalar quantity, although there is no fundamental reason why it could not be multidimensional. We will denote the comparison metric by $\delta = \delta(S_1,S_2)$ and refer to δ as the "genetic distance" although it need not be a traditional genetic distance measure[4].

For simplicity, our inferential framework assumes, as do most other models of molecular evolution, that the random process is Markovian, and that the Markov process describing evolutionary change is time-reversible, which is also almost universally assumed in phylogenetic theory[14]. The assumption of reversibility has the effect of making the probability function depend only on the absolute number of steps that separate two nodes in the transmission tree. Thus, the transmission tree is regarded as an undirected graph with M computed as the number of edges connecting two nodes regardless of whether they are connected by a chain through intermediate nodes, or are descended from a common ancestor node.

In addition, two other random processes play a role in determining the probability of observing a particular δ value. These arise from the uncertainty in the times t_1 and t_2 that isolates are obtained from a node relative to the time the node is infected, and uncertainty in τ, the time that pathogen transmission occurs relative to the time the transmitting node

was infected. These factors can affect the probability of observing a certain genetic difference between S_1 and S_2 because, the genetic diversity of the subpopulation of pathogens changes as the population size expands, and because of selective pressures and genetic drift during later stages of infection. To take these factors into account we can define probability distributions $P(t_1)$, $P(t_2)$, and $P(\tau)$, and average over them to obtain:

$$P(\delta|M) = \iiint P(\delta|M;t_1,t_2,\tau)\, P(t_1)\, P(t_2)\, P(\tau)\, dt_1 dt_2 d\tau \qquad (0)$$

This averaging accounts for the fact that the values of t_1, t_2 and τ are never known precisely.

The transmission tree associated with an outbreak is also generated by a random process. Disease transmission depends on particular mechanisms (e.g. airborne transfer by droplets, transmission by insect vectors, or the oral-fecal route) that are mediated by various kinds of social contacts. Each transmission tree generated in an actual outbreak can be thought of as a random sample from an ensemble of all possible outbreak trees that are consistent with the underlying mechanisms of transmission for that pathogen, and the underlying contact network for disease transmission. The probability $\mathcal{P}(M)$ that a pair of nodes drawn randomly from the tree will be related by M steps is defined on this ensemble of possible trees.

Consider an arbitrary sub-tree T drawn from the ensemble of outbreak trees {T} associated with outbreaks of the pathogen in question. Imagine that two nodes are chosen at random from this tree, the pathogen isolates from each node are sequenced and consensus sequences S_1 and S_2 are obtained, from which we calculate the value of $\delta(S_1, S_2)$. The joint probability of observing a particular δ value for a pair of nodes that are separated by M steps is given by:

$$P(\delta,M) = P(\delta\,|M)\bullet\mathcal{P}(M), \qquad (1)$$

It must be noted that equation (1) implicitly assumes that the relationship between δ and M is independent of the particular tree, but is only a function of host-pathogen interactions and the host-host transmission mechanisms for the disease in question, and that every node and every transmission event in the tree is governed by the same probability distribution. Normalization clearly requires that $\sum_M \mathcal{P}(M) = 1$. The probability that two nodes are separated by more than M_0 steps is

$$\mathcal{P}(M > M_0) = \sum_J \mathcal{P}(J), \text{ where J runs from } M_0+1 \text{ to } \infty, \qquad (2)$$

and the joint probability that two nodes exhibit a genetic difference δ and are separated by $M>M_0$ steps is

$$P(\delta, M > M_0) = \sum_J P(\delta|J)\mathcal{P}(J) \quad (3)$$

where J runs from $M = M_0+1$ to ∞.

Note that

$$\mathcal{P}(M \leq M_0) = 1 - \mathcal{P}(M > M_0) \quad (4)$$

and

$$P(\delta, M \leq M_0) = \sum_J P(\delta|J)\mathcal{P}(J) \quad (5)$$

where J runs from $M = 1$ to M_0

From equations (2) - (5) we can calculate the conditional probabilities

$$P(\delta|M > M_0) = P(\delta, M > M_0)/\mathcal{P}(M > M_0) \quad (6)$$

and

$$P(\delta\ |M \leq M_0) = P(\delta, M \leq M_0)/\mathcal{P}(M \leq M_0) \quad (7)$$

We can now use (6) and (7) and Bayes's theorem to calculate the probabilities that $M > M_0$ or $M \leq M_0$ *given* an observed δ value for isolates derived from the two nodes:

$$P(M>M_0|\delta) = P(\delta|M>M_0)\mathcal{P}(M>M_0)/[P(\delta|M>M_0)\mathcal{P}(M>M_0) + P(\delta|M\leq M_0)\mathcal{P}(M\leq M_0)] \quad (8)$$

and

$$P(M\leq M_0|\delta) = P(\delta|M\leq M_0)\mathcal{P}(M\leq M_0)/[P(\delta|M\leq M_0)\mathcal{P}(M\leq M_0) + P(\delta|M>M_0)\mathcal{P}(M>M_0)]. \quad (9)$$

Equations (8) and (9) provide weight-of-evidence expressions relating the measured δ value for a pair of isolates to the probability that they were drawn from nodes related by more than or fewer than M_0 transmission events respectively. When $M_0 = 1$, then these equations provide the probability that the two isolates are related by direct transmission. (Strictly, the probability functions $P(\delta|M)$ and $\mathcal{P}(M)$ are not defined for the case $M = 0$ since they refer to two distinct nodes from the network, so the condition $M \leq 1$ is equivalent to $M = 1$.)

Equation (9) with the condition $M = 1$ can be re-written in the form:

$$P(M=1|s_1,s_2) = \left[1 + \frac{P(s_1,s_2|M>1)}{P(s_1,s_2|M=1)} X \frac{\mathcal{P}(M>1)}{\mathcal{P}(M=1)} \right]^{-1} \quad (10)$$

where we have made explicit the dependence on S_1 and S_2, the sequences determined for the "victim" and "suspect" nodes. This form is analogous to the equation used to determine the probability of paternity or other familial relations in human DNA forensics.

It is easy to see that other hypothesis tests can also be defined within this framework. For example, the distribution $P(M=M_0|S_1,S_2)$, and its complement $P(M\neq M_0|S_1,S_2)$ where M_0 is an arbitrary number have utility for certain kinds of forensic cases where entire transmission chains must be reconstructed. Regardless of the precise form of the hypothesis test, calculations of the posterior probability depend, through equations (3) – (7) on the sampling distributions $P(\delta|M)$ and $\mathcal{P}(M)$, which are fundamental quantities that must be estimated through field and laboratory studies.

Appendix 4.
Accelerated determination of spontaneous mutation rates in bacteria

Mutation rates at specific genetic loci are key parameters for microbial forensics, but current methods for determining them require passage experiments that may extend over many months, if not years. As a result, comparatively little is known about such rates. In this appendix we demonstrate that there is a way to significantly accelerate the determination of bacterial mutation rates by optimizing the design of passage experiments. Our core concept is based on the observation that the precision of a mutation rate that is determined by a serial passage experiment depends only on the product of the number of generations and the number of lineages generated in an experiment. Modern techniques for analyzing DNA amplicons can process samples much more quickly than the time it takes to propagate bacteria over a large number of generations. As a consequence, *minimum time* experiments usually require significantly more replicate lineages and significantly fewer generations of growth than are traditionally used. Rather than propagating 10 replicate lineages for 10,000 generations, the minimum time experiment might consist of 1000 lineages propagated for 100 generations each. A consequence of the need to handle such large numbers of lineages in parallel is that automated replicate serial transfer is required to execute these time-optimized experimental designs. It is plausible that such massively parallel (in vitro) serial passage and sample processing could be carried out in an integrated microfluidic module. In addition to greatly reducing the time to complete these experiments, a number of quality control issues such as contamination prevention and sample mix-up are best addressed by such a completely automated system.

Classical mutation rate experiments focused on mutations that deactivate particular metabolic genes so that mutants could be identified by growth in selective media. These experiments are straightforward, but are restricted to the particular genetic loci for which metabolic selectivity can be identified. Experiments that are relevant to microbial forensics examine the genetic sequences of a more general selection of loci to determine changes in allele state. Examples are studies of substitution rates in *E.coli* [6] and VNTRs in B. anthracis[2] and *Y.pestis*[9] that were carried out over the last 10 years. Both PCR based genetic typing assays and genomic sequencing can be used to identify allele state changes.

There are a variety of published experimental protocols that have been used to perform targeted mutation rate experiments. Serial passage experiments can be differentiated by whether each new culture is initiated by a single clone transferred from the previous batch ("bottlenecking") or whether the inoculum contains a large number of micro-organisms from the previous culture (serial dilution.) In the context of the inference-on-nets concept, these two protocols mimic whether the infection transmission process involves a small sample of the infecting bacteria or a large sample, and therefore the probability of a change in the consensus genotype upon transmission.

Reference 9 describes a serial passage experiment which starts by agar plating a dilute suspension of bacteria assumed to be generated from a single clone. A number N of the resulting colonies are picked and streaked onto N new agar plates to form N_{lin} replicate lineages. After subsequent growth on the new plates, a colony is selected from each plate and streaked onto a new plate, and the procedure repeated M times. Under the assumption that each colony originates from a single bacterium, and represents n_{gen} generations of growth, the total number of generations accrued by each lineage is

$$N_{gen} = M \bullet n_{gen}. \quad \text{(A4.1)}$$

At the end of the passage period, a colony picked from each of the N final plates is processed for DNA analysis. PCR is used to amplify the loci that are of interest for mutational analysis, and the amplicons are analyzed. For many types of loci this simply consists of measuring the presence or size of the amplicons, but may also involve base composition or sequence determination.

Serial passage experiments can also be carried out in liquid cultures by repetitive serial dilution. In this case each replicate culture is initiated by a number N_0 of bacteria and grows to N_1 bacteria after n_{gen} generations:

$$N_1 = N_0 \bullet 2^{n_{gen}} \quad \text{(A4.2)}$$

In both agar plate and serial dilution type passage experiments n_{gen} is experimentally estimated from the logarithm of the ratio N_1/N_0, which can be determined in various ways. In agar plating, it is usually assumed that the chosen colonies were initiated by a single bacterium ($N_0 = 1$) and that colony size is related to N_1. In liquid culture, calibrated optical density measurements are used to estimate N_1 and N_0.

In some serial transfer protocols, cells are not transferred during or immediately after the exponential growth phase. This may be deliberate in some cases[7], but in others it is simply a consequence of opting for the convenience of having laboratory personnel perform the transfer at the same time each day. Thus, each passage of n_{gen} generations may consist of a lag, exponential, and resting phase. Such experiments might be good representations of multiple passages experienced by bacteria that are transferred to different laboratories. Since there is some evidence that mutation can occur during resting phases, it is not clear that these experiments result in rates that are identical to what would be observed if the culture were always in exponential phase growth throughout the passage experiment.

There are two phenomena that can affect the accuracy of mutation rate estimates and must be monitored for in each lineage. The first of these is adaptive mutations that cause significant increases in growth rate, hence re-defining the value of the generation time (τ_{gen}) for that lineage. In continuous culture these can result in "sweeps" in which a newly mutated genotype with increased fitness rapidly takes over the culture. The

second important phenomenon is the generation of "mutators" which have significantly increased mutation rates. There is some evidence that mutators may arise as a response to stressful growth conditions such as starvation[12].

The underlying model of mutation rates assumes that the appearance of mutations at each locus is an independent Poisson random process characterized by a rate constant γ. While this is a plausible framework for estimation, there are a number of ways that actual mutation rates might not conform to this simple model[15]. For example, the rate of mutation at a given locus could depend on the allele state. (Mutation rate data for VNTRs in *Y. pestis* suggest that this may be the case[16].) In addition, there are reasons to believe that mutations at one locus may have effects on the rate of mutations at other loci. The existence of mutator variants is one very obvious example of this. Nonetheless, the Poisson approximation has been adapted by others in this field and provides an uncomplicated basis for discussing the gross features of passage experiments and their optimization.

Consider an experiment with N_{lin} replicate lineages, each consisting of N_{gen} generations. In any one lineage the probability of no mutations in a certain locus after N_{gen} generations is given by:

$$P_0(N_{gen}, 1) = \exp(-\gamma N_{gen}) \qquad \text{(A4.3)}$$

where γ is the rate constant for that mutational locus. In N_{lin} independent identical lineages, the probability of observing no mutations is:

$$P_0(N_{gen}, N_{lin}) = P_0(N_{gen},1)^{N_{lin}} = \exp(-\gamma N_{gen} N_{lin}) \qquad \text{(A4.4)}$$

Thus, the probability of observing at least one mutation in that locus in a passage experiment consisting of N_{lin} lineages of N_{gen} generations each is given by:

$$P_1(N_{gen},N_{lin}) = P_1 = 1 - \exp(-\gamma N_{gen} N_{lin}) \qquad \text{(A4.5)}$$

If $P_1 = 95\%$, then out of 100 identical replicate experiments consisting of N_{lin} lineages with N_{gen} generations each, 95 experiments would exhibit least one mutation among the N_{lin} lineages examined. Conversely, if a locus we are interested in has a mutation rate of γ, in order to do an experiment that has a 95% chance of observing at least one mutational event in that locus we would need to generate N_{lin} identical lineages of N_{gen} generations each. (We have also derived an expression for the dependence of P_1 on the uncertainty in the experimentally derived mutational rate constant, not shown here.)

If, during a single experiment we observe m_μ mutational events at a locus of interest among the N_{lin} lineages, an estimate of the mutation rate at the observed locus is given by (see reference 1):

$$\gamma \approx m_\mu / N_{lin} N_{gen} \quad \text{(A4.6)}$$

On the other hand, the variance of γ is given by:

$$Var(\gamma) \approx \gamma / N_{lin} N_{gen} \quad \text{(A4.7)}$$

Equations (6) and (7) are valid when mutational events are rare, i.e. the fraction of lineages that exhibit a mutational event is much smaller than 1. The relative uncertainty in an estimate of γ is then given by

$$\sigma_{rel} = 1/(\gamma N_{lin} N_{gen})^{1/2} \quad \text{(A4.8)}$$

Note that in all these equations only the product of N_{gen} and N_{lin} is important, so that an experiment with a single lineage with a certain number of generations is equivalent to one with two lineages with half the number of generations each. In the next section we will demonstrate how this fact can be exploited to minimize the time required to complete a passage experiment without compromising the accuracy of the determined mutational rate constant.

The time needed to complete a passage experiment is the sum of three terms:

$$T_{exp} = T_{gr} + T_{pr} + T_{an} \quad \text{(A4.9)}$$

where T_{gr} is the total time spent growing the culture, T_{pr} is the time consumed in preparing the initial culture for the experiment and the final (passaged) cultures for DNA analysis, and T_{an} is the time required to analyze the passaged cultures for the presence of mutations. In most experiments, T_{pr} is negligible compared to the time spent on growth and analysis. This is because the most time consuming element, preparation of the N_{lin} passaged samples for DNA analysis, can be done in parallel on all the samples simultaneously. Therefore, in subsequent derivations we will assume $T_{pr} = 0$. If we define τ_{gen} to be the generation (doubling) time of the culture, and T_{1an} to be the time it takes to analyze the DNA amplicons from one passaged sample (i.e. one lineage) then:

$$T_{exp} \approx \tau_{gen} \ln(2) N_{gen} + T_{1an} N_{lin} \quad \text{(A4.10)}$$

Recall that equation (A4.5) defines the required product of N_{lin} and N_{gen} that is needed in order to ensure that mutations with rates up to γ will be detected among the N_{lin} lineages with probability P_1. For a robust passage experiment designed to accurately assess mutation rates greater than a certain value, say γ_0, P_1must be chosen to have a high value (0.95 for example; alternatively we can choose to constrain σ_{rel}, see Appendix.) This sets

a constraint on the value of the product $N_{lin} \bullet N_{gen}$. Under this constraint, equation (A4.10) can be minimized with respect to N_{gen} to give:

$$(N_{gen})_{Min} = [-T_{1an} \bullet \ln(1 - P_1) / \tau_{gen} \bullet \ln(2) \bullet \gamma_0]^{1/2} \qquad \text{(A4.11)}$$

The associated number of lineages for this minimum time experiment is given by:

$$(N_{lin})_{Min} = -\ln(1 - P_1)/\gamma_0 \bullet (N_{gen})_{Min} \qquad \text{(A4.12)}$$

and the minimum value of T_{exp} is given by:

$$(T_{exp})_{Min} = 2[-T_{1an} \bullet \ln(1 - P_1) \bullet \ln(2) \bullet \tau_{gen}/\gamma_0]^{1/2} \qquad \text{(A4.13)}$$

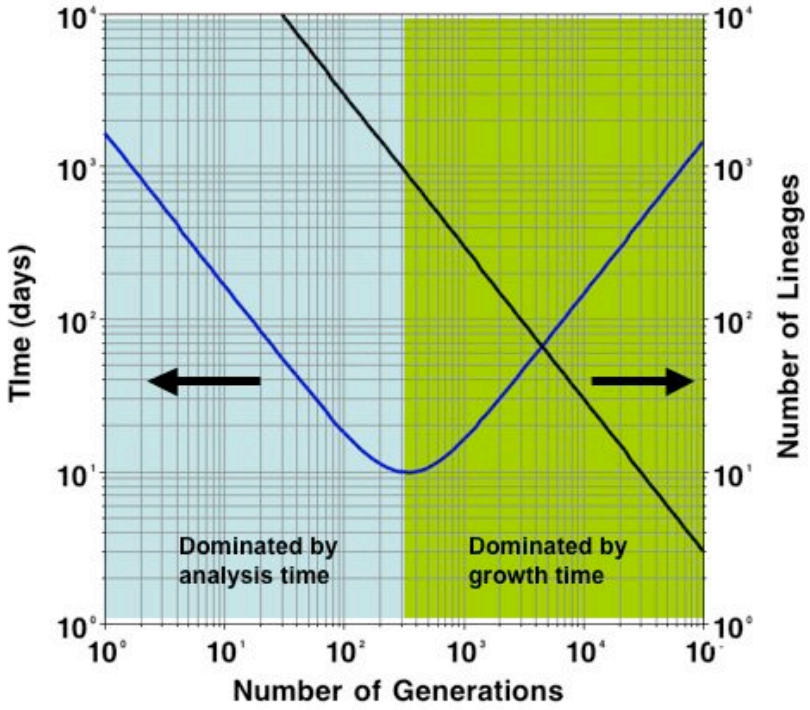

Figure 2. Time and required number of lineages as a function of the number of generations. Parameters: $\gamma_0 = 1 \times 10^{-5}$/gen; $P_0 = 0.95$; $\tau_{gen} = 0.5$ hr; $T_{1an} = 8$ min.

To illustrate the degree of time reduction that is possible with an optimum choice of experimental design, consider a serial passage experiment in which we wish to observe mutations with rates as low as 10^{-5} per generation with a confidence level of $P_1 = 0.95$. We will assume that τ_{gen} is 30 minutes and that T_{1an} is 8 minutes (a typical value for a commercial capillary electrophoretic sequencer, see below.) The values of T_{exp} and N_{lin} that are obtained as N_{gen} is varied is shown in Figure 2.

Note that a "typical" experiment that propagates bacteria for 10^4 generations requires only around 30 lineages, but will take more than 100 days to complete. In contrast,

Figure 2 indicates that only 10 days are required to gain the same data from approximately 1000 lineages that have been propagated only 300 generations each.

From equations (A4.11) and (A4.12) the ratio of the optimum N_{gen} and N_{lin} values is:

$$(N_{gen})_{Min}/(N_{lin})_{Min} = T_{1an}/[\ln(2)\bullet\tau_{gen}] \qquad (A4.14)$$

Thus, as long as the analysis time is short compared to the generation time, the optimum will lie with smaller N_{gen} and larger N_{lin}. This is likely to be the case for most bacteria. E. coli has a laboratory generation time of about 20 minutes in optimal media. This represents one of the shortest bacterial generation times. Most pathogens have slower growth rates, for example M. tuberculosis has a generation time of about 12 hours. In contrast, standard CE analysis of PCR amplicons can be done at rates faster than 10 minutes per sample, and new technologies such as the TIGER electrospray mass spectrometer based system can analyze close to one sample per minute[17].

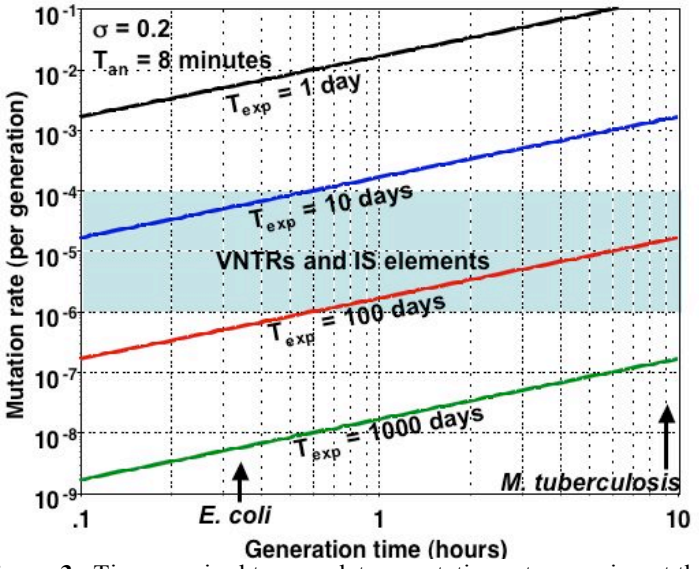

Figure 3. Time required to complete a mutation rate experiment that can determine a mutation rate of a given value (y axis) with a precision of 20%, for a bacterium with a given generation time (x axis). This calculation assumes that allele state at a given locus can be determined in 8 minutes, e.g. by a CE instrument. The blue band indicates the range of mutation rates that have been reported previously for VNTR and IS element mutations.

Figure 3 shows the time that would be needed to complete a time-optimized experiment that can determine a mutation rate with a reasonable precision (20%) assuming that the per-sample analysis time is 8 minutes. Note that for bacteria with generation times up to several hours, less than 100 days (≈ 3 months) are required to determine precise rate constants of 10^{-5} per generation or greater, and considerably less than 1000 days (2.7 years) are required to determine such rates in slowly growing bacterium like *M. tuberculosis*. Note that a strong implication of this figure is that many experiments to date are not reporting very precise measurements.

The number of lineages required for time-optimized determination of precise mutation rates is shown in Figure 4. Under assumptions identical to those used in Figure 3, experimental designs with between 10^3 and 10^4 lineages are needed to obtain precise values of VNTR and IS element mutation rates. Note also the consistency between Figures 3 and 4, since 10^4 samples would require a total analysis time of approximately 55 days, or a total experiment time of 110 days.

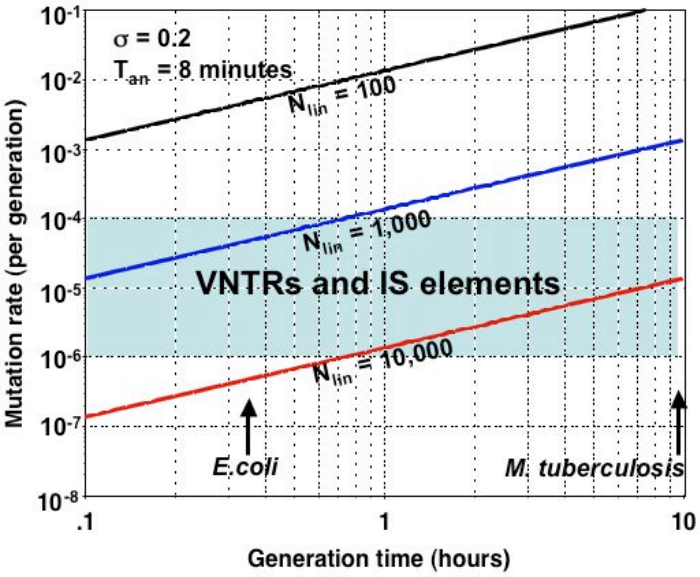

Figure 4. Number of lineages required for a time-optimized experiment to determine a mutation rate (y axis) with a precision of 20%, as a function of bacterial generation time.

This analysis shows that a time-optimized experimental design requires us to handle the growth, processing, and analysis of the very large number of parallel lineages. Each of the three basic steps in a mutation rate experiment must be considered: (1) Serial transfer and growth over a large number of generations and/or lineages; (2) processing of samples for mutation identification assays; and (3) measurement of allele states at selected loci. Finally, one must consider the integration of the three steps.

A superior approach to carrying out such highly multiplexed parallel processes in practice is to utilize recent advances in the microfabrication of bioreactors, combined with cutting-edge microfluidic platforms. Microfluidic platforms are very well suited for replication and manifolding to permit highly parallel processing. This means that many samples can be handled at once so the total elapsed processing time is no longer than it takes to handle a single sample. The movement of fluids through the system is computer-controlled and such platforms are envisioned as an ultra-compact and integrated replacement for laboratory robotics. Thus, the requirements for automation, integration, and parallel processing can be met using this approach, and such a system is inherently safer than standard culturing because the fluid handling is nearly completely closed. Since it is automated, it minimizes the operator exposure per sample, and the entire complex fluidic system is small and easier to decontaminate than a laboratory robot. A microfluidic apparatus with thousands of channels for sample handling and growth could easily fit inside a standard biosafety cabinet.

In principle, passaging experiments require only a simple modifications of existing microbioractor and microfluidic system designs. Figure 7 shows how a simple paired cell system can be used to perform repetitive serial transfers. Each cell must be outfitted to allow purging and cleaning prior to the next growth cycle, and a system for transferring one or more bacteria from one chamber to the next must be implemented. Transfer of a known quantity of bacteria between growth chambers is probably the most challenging aspect of this scheme, especially if seeding with a single clone is desired. Nonetheless, automated cell sorting in microfluidic chips has been demonstrated, and it is likely that this problem can be solved by similar principles[30].

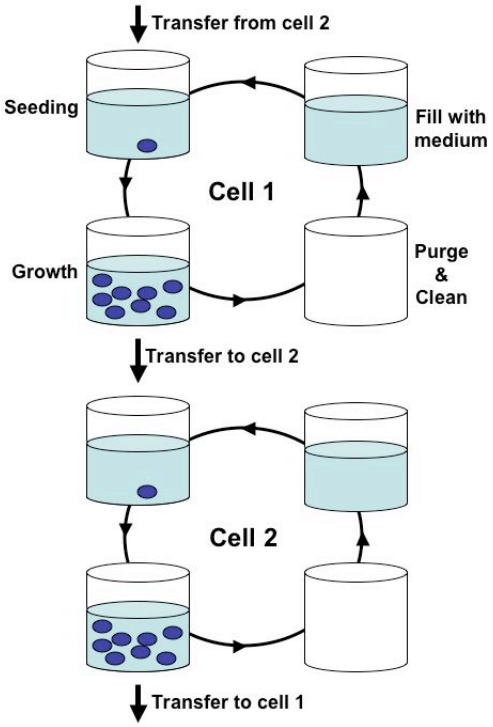

Figure 7. Two cell system for automated serial passage experiments.

An important design consideration stems from a practical constraint on the minimum volume of the reactor cell that can support a given number of generations of cell division and a desired final cell density. If ρ is the desired final cell density after the growth cycle, N_{gen0} is the number of generations in one growth cycle, then V is given by:

$$V = 2^{Ngen0}/\rho \qquad (15)$$

Figure 8 shows that there is a narrow range of minimum working volumes that will support a typical growth cycle in which 20-30 generations of growth lead to final cell densities in the 10^8/ml to 10^{10}/ml range. In fact, volumes between 10 and 100 μliters are ideal for micro-fabrication of highly multiplexed reactor units.

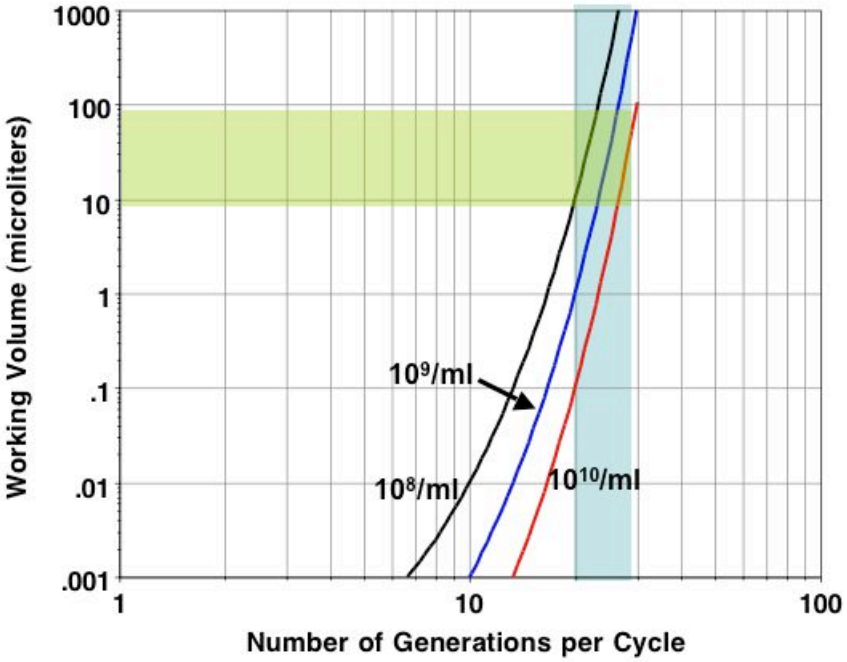

Figure 8. The working volume required to support a given number of generations of growth at a specified final cell density.

References for Appendix 4.

1. Lenski, R.E. and Keim, P., "Population genetics of bacteria in a forensic context", in Micobial Forensics (R.G. Breeze, B. Budowle, and S.E. Shutzer, eds.) (Elsevier Acasdemic Press, San Diego CA, 2005), pp.355-369.

2. Keim, P., et. al. "Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales", Infection Genetics and Evolution 2004, **4**, 205-213.

3. Lowell, J. L., et. al. "Identifying Sources of Human Exposure to Plague", J. Clinical Microbiology, 2005, 43, 650-656.

4. Wendel, A., "Advanced Systems and Concepts Office Overview", presentation for the UNM Neutrino Workshop, February 5, 2004.

5. Keim, P. "Microbial Forensics: A Scientific Assessment", ASM position paper.

6. Lenski, R.E., et. al., "Rates of DNA Sequence Evolution in Experimental Populations of Escherichia coli During 20,000 Generations" J. Mol. Evol. (2003) **56**, 498-508.

7. D. Papadopoulos D., et. al. "Genomic evolution during a 10,000-generation experiment with bacteria", Proc. Natl. Acad. Sci. USA (1999) **96**, 3807-3812.

8. Warren, R.L. et. al. "Advanced Microbial Forensic Initiative", in SRS Technologies report TR04-507, p. 13-29, (2004).

9. Girard, J.M., et. al. "Differential plague-transmission dynamics determine Yersinia pestis population genetic structure on local, regional, and global scales", Proc. Natl. Acad. Sci. USA (2004),**101**, pp.8408-8413.

10. Dykhuizen, D.E. and Hartl, D.L., "Selection in Chemostats", Microbiology Reviews 1983; **47**, 150-168.

11. Helling, R.B., et. al. "Evolution of Esherichia Coli During Growth in a Constant Environment", Genetics (1987) **116**, 349-358.

12. Hersch, M.N., et. al., "Adaptive mutation and amplification in Escherichia coli: two pathways of genome adaptation under stress", Research in Microbiology 2004, **155,** 352-359.

13. Husimi, Y. et. al., "Cellstat – A continuous culture system of a bacteriophage for the study of the mutation rate and the selection process at the DNA level", Rev. Sci. Instrum. (1982), **53**, 517-522.

14. Drake, J.W. and Hwang, C.B., "On the Mutation Rate of Herpes Simplex Virus Type I", Genetics 2005, **170**, 969-970.

15. Rosche, W.A. and Foster, P.L. "Determining Mutation Rates in Bacterial Populations",ID Methods, 2000, **20**, pp.4-17. Available online at www.idealibrary.com.

16. Keim, P., personal communication.

17. Sampath, R. et. al. "Rapid Identification of Emerging Pathogens: Coronavirus", Emerging Infectious Diseases 2005, **11**, 373-379.

18. For example, the Genomic Solutions HiGro high capacity microwell plate growth system (www.GenomicSolutions.com).

19. Kostov, Y. et. al., "Low-Cost Microbioreactor for High-Throughput Bioprocessing", Biotechnology and Bioengineering 2001, **72**, 346-352.

20. Szita, N., et. al., "Development of a multiplexed microreactor system for high-throughput bioprocessing", Lab Chip 2005, **5**, 819-826.

21. Maharbiz, M.M. et. al. "Microbioreactor Arrays With Parametric Control for High-Throughput Experimentation", Biotechnology and Bioengineering, 2004, **85**, 376-381.

22. Zanzotto, A. et. al., "Membrane-Aerated Microbioreactor for High-Throughput Bioprocessing", Biotechnology and Bioengineering 2004, **87**, 243-254.

23. Kim, J.W. and Lee, Y.H., "Development of a microfermenter chip", J. Kor. Phys. Soc. 1998, **33**, S462-S466.

24. Belagadde, F.K. et. al. "Long-Term Monitoring of Bacteria Undergoing Programmed Population Control in a Microchemostat", Science 2005, **309**, 137-140.

25. Song, J.W., et. al., "Computer -Controlled Microcirculatory Support System for Endothelial Cell Culture and Shearing", Anal. Chem. 2005, **77**, 3993-3999.

26. Leclerc, E., et.al., "Microfluidic PDMS (Polydimethylsiloxane) Bioreactor for Large-Scale Culture of Hepatocytes", Biotechnol. Prog. 2004, **20**, 750-755.

27. Sin, A., et. al., "The Design and fabrication of Three-Chamber Microscale Cell Culture Analog Devices with Integrated Dissolved Oxygen Sensors", Biotechnol. Prog. 2004, **20**, 338-345.

28. www.bioprocessors.com

29. www.Fluidigm.com

30. Takahashi K, et. al., "Non-destructive on-chip cell sorting system with real-time microscopic image processing", J. Nanobiotechnology 2004, **2**:5; published online.

31. Quake, S.R., and A. Scherer. 2000. From Micro- to Nanofabrication with Soft Materials, *Science* **290**, 1536-1540.

32. Ungar, M.A., H.-P. Chou, T. Thorson, A. Scherer, and S.R. Quake. 2000. Monolithic Microfabricated Valves and Pumps by Multilayer Soft Lithography, *Science* **288**, 113-116.

33. Xin, Z., et. al., "High-Throughput DNA Extraction Method Suitable for PCR", BioTechniques 2003, **34**, 820-826.

34. Greenspoon, S.A., "Application of the BioMek 2000 laboratory automation workstation and the DNA IQ system to the extraction of forensic casework samples", J. Forensic Sciences 2004, **49**, 29-39.

35. Smit, M.L., et. al., "Automated extraction and amplification of DNA from whole blood using a robotic workstation and and integrated thermocycler", Biotechnol. Appl. Biochem. 2000, **32**, 121-125.

36. Auroux, P.A. et. al., "Miniaturized nucleic acid analysis", Lab Chip 2004, **4**, 534-546.

37. Erickson, D, and Li, D., "Integrated microfluidic devices", Analytica Chimica 2004, **507**, 11-26.

38. Chovan T., and Guttman A., "Microfabricated devices in biotechnology and biochemical processing", TRENDS in Biotechnology, 2002, **20**, 118-121.

39. Saunders, N.J. et. al. "Mutation rates: estimating phase variation rates when fitness differences are present and their impact on population structure", Microbiology 2003, **149**, 485-495.