LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-TR-215305

Resolution in forensic microbial genotyping

Stephan P. Velsko

September 12, 2005

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Resolution in forensic microbial genotyping

Stephan P. Velsko
Lawrence Livermore National Laboratory

UCRL-TR-215305

August 26, 2005

Abstract

Resolution is a key parameter for differentiating among the large number of strain typing methods that could be applied to pathogens involved in bioterror events or biocrimes. In this report we develop a first-principles analysis of strain typing resolution using a simple mathematical model to provide a basis for the rational design of microbial typing systems for forensic applications. We derive two figures of merit that describe the resolving power and phylogenetic depth of a strain typing system. Rough estimates of these figures-of-merit for MLVA, MLST, IS element, AFLP, hybridization microarrays, and other bacterial typing methods are derived from mutation rate data reported in the literature. We also discuss the general problem of how to construct a "universal" practical typing system that has the highest possible resolution short of whole-genome sequencing, and that is applicable with minimal modification to a wide range of pathogens.

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Executive summary

The resolution of a microbial strain typing system is related to its exclusionary power in an investigation of bio-terror or bio-crime cases. Resolution is thus a key parameter for differentiating among the large number of strain typing methods that could be applied to pathogens involved in such events. In this report we develop a first-principles analysis of strain typing resolution using a simple mathematical model to provide a basis for the rational design of microbial typing systems for forensic applications.

The salient result derived from this analysis is a simple figure-of-merit (F.O.M.) for strain typing systems that is directly related to the probability that the typing method will recognize that two microbial lineages separated by a small number of mutations are, in fact, distinct. A second figure-of-merit describes the "phylogenetic depth" of the typing system, which is a property complementary to resolution. These figures-of-merit are determined by the mutational spectrum of the microbe and by the number of loci effectively probed by the typing system. Rough estimates of these figures-of-merit for MLVA, MLST, IS element, AFLP, and other bacterial typing methods are derived from mutation rate data reported in the literature.

As a result of this analysis, it is possible to provide *quantitative* support to a number of general conclusions about typing systems:

- In general, typing systems that examine the largest number of mutational loci with the highest mutation rates have the highest resolution. However, equivalent resolution can come from either a relatively few markers with high mutation rates (e.g. VNTRs) or from a large number of more slowly mutating markers (e.g. SNPs.) Because substitution rates are generally around 4 orders of magnitude slower than VNTR mutation rates, a typing system would need to examine thousands of SNPs in order to match the resolution provided by a few VNTRs.

- All typing systems based on a single kind of mutation (e.g. VNTRs) will have limited resolution and limited "universality." Resolution will be limited because once the first few markers with the highest mutation rates are scored, subsequent markers add marginally to the resolution. Universality will be limited because the number of such markers, as well as the rate of mutation of a given marker type, can change both among species and *within* a species.

- Resolving power is *additive* for independent mutations. This fact, along with the previous observation, argues for "mixed" or polyvarietal typing systems that exploit all the fast mutational loci, regardless of type (VNTR, IS, DR, etc.) To exploit this observation in practice, there is a critical need for a "universal" approach to marker identification and assay development.

- For bacterial typing, to approach the maximum resolution short of complete genomic sequencing, it is necessary to probe both fast mutational loci and mutations that have low rates but are distributed over very many loci in the genome (e.g. SNPs.) Hybridization arrays provide a potential means for genome-wide analysis of SNPs. However, type I error rates (false detections of mutational differences) rise rapidly in this regime, potentially compromising the exclusionary value of the typing system. Thus, in the presence of finite error rates, adding additional loci to those already probed by extended multi-locus typing systems alone will provide only marginal improvements in resolution for many pathogens of interest.

- Rapidly mutating organisms such as RNA viruses are much easier to type at high resolution than bacteria. Only a small fraction of an RNA viral genome must be examined to differentiate lineages separated by fewer than 100 generations.

- Careful consideration must be given to the construction of the "diversity panels" of micro-organisms that are used to evaluate the resolving power of typing assays. It is advisable that closely related islolates, such as those generated by serial passage, be included in these panels systematically, and that the widest possible selection of geographically diverse isolates be used.

For forensic applications, the selection of a single standard typing method would advantageous for simplifying QA/QC, proficiency testing, and other requirements related to admissibility. The analysis in this report is intended to motivate discussion of the general problem of how to construct a "universal" practical typing system that has the highest possible resolution short of whole-genome sequencing, and that is applicable with minimal modification to a wide range of pathogens. Although multi-locus typing assays currently command the widest interest among assay developers, it is possible that the highest resolution "universal" typing systems could be based on other technologies, such as restriction fragment analysis, hybridization based assays, or mass spectrometric approaches.

Table of Contents

1. Introduction

Microbial genetic typing, or *strain typing,* is a key element of microbial forensics that may often be crucial for elucidating the chain of events that connect the death or illness of victims, or the contamination of property, with an act of bio-terrorism [1-8]. In forensic applications, genetic typing is used to establish the degree of relatedness among samples of the agent organism found at the crime scene and in other venues during the course of the investigation. One of the most important properties of a strain typing method used for microbial forensics is its *resolution*, which is a measure of its ability to differentiate between strains that are nearly identical in DNA sequence. The resolution of a strain typing method is proportional to its *exclusionary power*, i.e. its ability to eliminate the need to investigate certain pathways for the acquisition of an agent by a terrorist or to make implausible certain routes of infection of a victim. Resolution requirements for forensic strain typing are generally more exacting than those for clinical typing, since forensic typing may be called upon to identity DNA differences for which there is no discernable phenotypic consequence. However, there are strong similarities between the resolution requirements for epidemiological strain typing methods and those used in forensic investigations [9-13, 14].

In this regard, it is generally understood that the complete DNA sequence of a microbe is the ultimate high resolution "fingerprint" of that organism, but complete genome sequencing is still too time-consuming and expensive a task to apply systematically as a strain-typing tool for large sets of pathogen isolates. Thus, a number of simpler techniques have been developed that are quicker and more convenient to apply to large sample sets, but have less resolution than whole genome sequence comparisons. In almost all cases, these techniques have been developed *ad hoc* for clinical, epidemiological, or ecological applications, not explicitly for microbial forensics. Recent reviews identify and discuss more than a dozen distinct genetic strain-typing methods that have been proposed and demonstrated in the literature [15-19]. Many more variants can be found in the literature on specific organisms.

The large number of available techniques naturally raises the question of whether there is a smaller subset of techniques among them that can be selected as "universal" forensic strain typing technologies. Such a down-selection is both desirable and necessary because it would greatly streamline the formulation of forensic QA/QC guidelines for laboratory analysis, simplify courtroom evidentiary standards, and, in the case of an investigation, would ultimately reduce the time and cost of typing large numbers of suspect samples by reducing to manageable size the set of isolates that must be resolved by whole genome sequencing. Although resolution is not the only consideration in this regard, it is clearly an important one.

It should be recognized that this question is not merely of academic interest. Recent legislation embodied in the Terrorism Preparedness and Response Act of 2002[20] and the accompanying House Conference Report [21] sets forth, in effect, a requirement for high resolution strain typing in conjunction with the registration of select agent holdings within the U.S. It is clearly the intention of the framers of this act that select agents be "fingerprinted" with sufficient resolution that agents used in a bio-terror incident be traceable to laboratories of origin. No consideration has been given to the technical feasibility or mode of implementation of this capability.

Formulating a strategy for choosing a standard strain typing system requires us to answer several related questions about resolution:

- Does a strain typing technique need to be specially tailored to each pathogen to achieve the highest possible resolution? If so, how does one identify the optimum technique for a given pathogen?

- Can typing resolution be significantly improved by combining two or more techniques either in parallel, or in tandem?

- Is there a systematic way to approach the development of new higher resolution techniques that push typing significantly closer to sequencing?

The scientific literature contains many individual studies where two or more strain typing methods have been compared on a common set of bacterial isolates of a given species [22-25]. These studies often compare older, more established techniques such as RFLP, Ribotyping, or PFGE with newer proposed methods, often based on PCR technology. However, only a few of these studies [26-31]specifically address organisms from the select agent list, and none were undertaken with forensic considerations in mind.

In 2001, a major strain typing inter-comparison exercise was initiated under joint sponsorship by the DTRA Advanced Systems and Concepts Office and the NNSA Chemical Biological National Security Program [32]. This multi-agency "round-robin" activity focused on *Bacillus anthracis*, *Yersinia pestis*, *Brucella* spp. and E. *coli* O157:H7, and included evaluation of MLVA, PFGE, RFLP, AFLP, MLST, IS-100, and RiboPrinter typing systems. In addition to resolution, a number of other considerations, including reproducibility, data transferability, and requirements for sample size and purity, were considered in this study. However, the design of this inter-comparison study did not emphasize resolution, since each technique was applied to a set of 25 widely diverse strains of each of the 4 bacteria, rather than to a set of strains known to be clonally related [33] with very high genetic similarity.

Direct experimental comparisons of this sort are an important element of any down-selection process (especially for understanding practical issues that can influence the ultimate utility of a technique.) However, there is a clear need for more theoretical guidance that can help narrow the number of techniques that need to be evaluated experimentally, provide an improved basis for the design of inter-comparison studies, and help guide future development of improved methods by providing a framework for understanding the optimization of techniques for different agents. Toward this end, this paper attempts to provide a simple "back of the envelope" quantitative analysis of the resolution properties of microbial forensic strain typing methods that provides answers to many of these questions.

As a step towards understanding resolution requirements, we consider in section 2 of this report a generic bio-terrorism scenario, and describe the ways that strain typing would be used in the investigation of such an event. One of the most important characteristics of a typing system is its ability to quickly eliminate from the investigation as many alternate sources (i.e. laboratories or natural reservoirs) as possible by showing that the genetic "fingerprint" of the attack strain does not match them as closely as it does others. The importance of this capability to any particular investigation depends, of course, on two factors: (1) how widely a group of closely related strains are held – i.e. the number of distinct laboratories, natural reservoirs, or infected individuals (e.g. for HIV) that are potential sources for the attack strain, and (2) the degree to which other evidence can be used to exclude "suspect lineages" that are not resolved by strain typing. While statistical data on US holdings of select agents is not yet available, we argue that the situation faced by the recent anthrax investigation is not unique -in many cases we can expect to find multiple laboratories or other potential sources for a bio-agent possessing strains whose lineages are separated from a clonal parent [33] by no more than a few thousand generations. Ultimately, the number and size of such clonal sets, and the time and cost of whole genome sequencing establishes the basic resolution requirements for a forensic typing system.

Section 3 summarizes our best current understanding of the types and rates of mutational changes that have been observed in bacteria and viruses. Since much of this information has been obtained for just a few particular micro-organisms, its extrapolation to the full range of select agent pathogens represents a necessary, but possibly erroneous generalization. Nonetheless, while it is possible that surprising deviations from these generalizations may yet be discovered, we believe that the picture developed here has sufficient qualitative validity to provide a basis for estimating the resolution figures-of-merit of a variety of existing strain typing methods.

In section 4 we introduce a simple model for strain typing resolution that can be derived from considerations of the number and rate of the mutations probed by each technique. We consider explicitly the probability that two strains that are

clonally related by a lineage of N generations will be distinguished by the typing method, if they differ by one or more mutations. In this model, a simple figure of merit (FOM) emerges that characterizes the resolution of the typing system. When two typing systems are combined, the figure-of-merit of the combined system is generally smaller than the sum of the figures-of-merit of the two typing systems, unless they probe completely separate sets of mutational loci with independent mutation rates. It is also easy to show that the high mutation rates of RNA viruses permit high resolution typing even when the typing system used has a relatively low figure-of-merit.

The theory outlined in Section 4 assumes that typing systems are "error free," in the sense that any mutation that is detectable is reported, and there is no potential for reporting a mutation that, in fact, does not exist. Thus, for example, whole genome sequencing is assumed to result in a completely accurate sequence, with a base-calling error rate of zero. In practice, of course, this is not true, and the inclusion of error rates is important for understanding how to optimize typing systems. However, we have deferred the discussion of the influence of error rates on inferences about matches between genetic fingerprints to section 6.

Section 5 provides some estimates of the resolution figures-of-merit of several major strain typing methods. In order to analyze the intrinsic resolution capabilities of these methods, we focus on their fundamental principles and do not dwell on differences in implementation. Available data on mutation rates and the distribution of various kinds of mutations on bacterial chromosomes is used to produce rough estimates of the resolution FOM. The estimates support the general observation that the highest resolution systems are those that characterize the largest number of the highest mutation rate loci. However, it is also true that once these "fast" markers are captured, adding additional loci provides rapidly diminishing returns.

Section 6 introduces the effects of error rates on the overall resolution of typing systems. Both type I errors, i.e. those that falsely report genomic differences between two isolates when, in fact, there are none, and type II errors, i.e. those that miss actual genomic differences, are considered. The complete theory is formulated in terms of hypothesis testing, and sets the stage for discussions of optimization. We show that, unless typing error rates are quite small, an optimum typing system may be one that examines only a limited number of the fastest mutating loci. Technologies such as hybridization arrays, which examine very large numbers of SNPs and other slow mutations, must have very low error rates before they can be exploited to provide high genotyping resolution.

Finally, in section 7 we discuss several challenges in the path toward the development of an optimal and "universal" forensic strain typing system. An important issue related to current practice is the use of sets of widely diverse isolates to screen typing methods for resolving power. We point out that while

this method finds the set of "permissive" mutations (i.e. those that are less constrained by selective pressures,) it does not necessarily identify the "fast" ones. That is, it is entirely possible that one can find high diversity among mutations that have slow rates and low diversity among mutations with high rates, especially when the isolates are separated from each other by more than ten thousand generations. Therefore, we recommend the use of standard sets of clonally related isolates that are generated by serial passage to develop and validate assays for forensic strain typing.

Although we emphasize issues which impact forensic applications in this report, microbial strain typing is of interest to microbiologists for many reasons. An excellent review of the interrelationships between microbial genetic typing, taxonomy, phylogenetics, mutation, and epidemiology, can be found in reference 12, which is recommended to the reader seeking background in this area.

2. Microbial forensic scenarios and inferences

The resolution requirements for a strain typing method used in an investigation will, in general, depend on the context in which it is used. Figure 1 shows the generic stages in a bio-terror incident starting with acquisition of a pathogen by the terrorist or criminal, subsequent development of that pathogen into the bio-terror "weapon" through cultivation and possible genetic modification, dissemination by some route to infect the primary victims and finally possible contagious spread of the disease to secondary victims. (Of course, in many cases the genetic manipulation step will be absent, in some cases even the propagation step may not apply, and not all agents are contagious.)

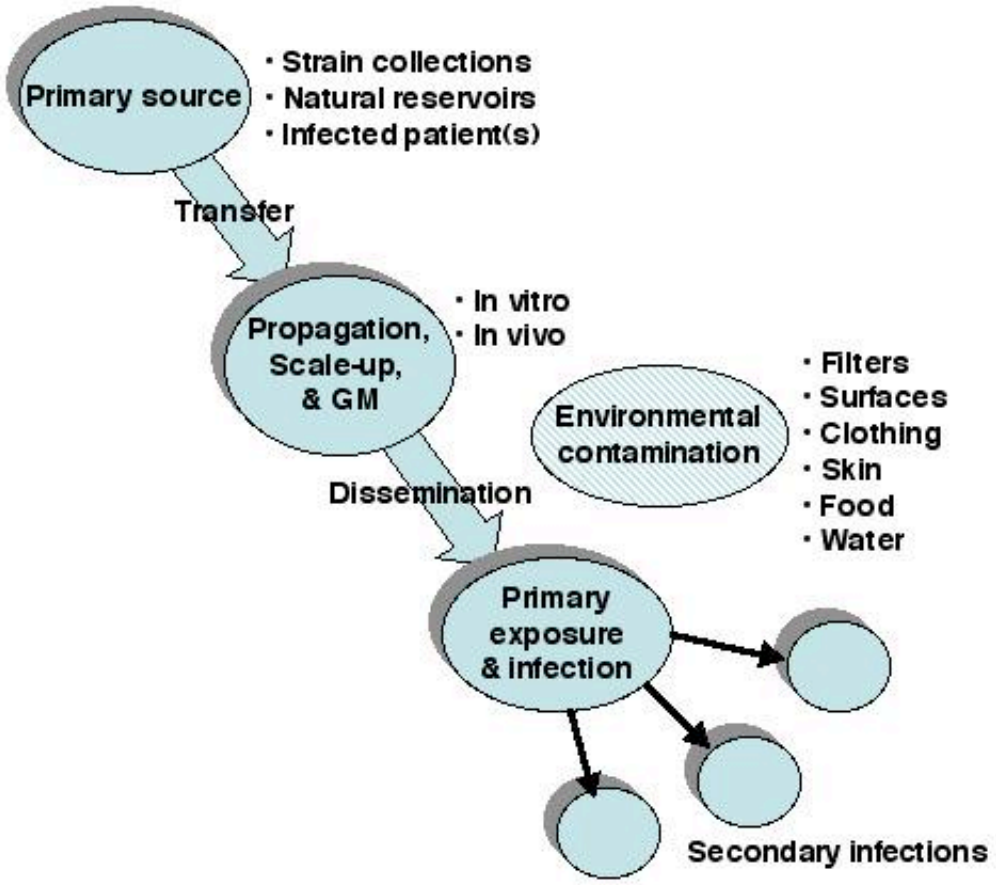

Figure 1. General steps in a bioterror incident, and corresponding sources of samples collected in the subsequent investigation.

Given this generic scenario, there are two entry points for law enforcement involvement. If the act of dissemination is accompanied by an overt threat or warning, or is witnessed and reported as a suspicious event, law enforcement investigation may begin immediately, and any related sickness or deaths will be investigated as presumptive consequences of the event. If the act of dissemination is covert, the fact of bio-terrorism may only be decided after medical and epidemiological investigation. As the anthrax letters incident demonstrated, these otherwise distinct response components may only coalesce through a rather complex chain of occurrences as the magnitude and geographic scope of the event becomes clearer to the investigators [34]. Nonetheless, once law enforcement investigation begins, the collection of samples and subsequent strain typing acquires a purpose and character distinct from clinical and epidemiological investigation.

Referring to Figure 1, we can distinguish between two basic contexts within a law enforcement investigation that engender different strain resolution requirements in practice. The first context is where strain typing analysis is used

to establish matches among evidence samples obtained during the investigation, including:

- Clinical samples from victims, or infected perpetrators
- Samples collected from surfaces, food, water, etc. at a putative crime scene, for example, a contaminated mail room or a restaurant salad bar
- Samples seized during the investigation of a site where it is suspected that the agent was generated or stored.
- Samples from Biowatch or other environmental samplers after routine screening indicates the occurrence of a pathogen release

Collection of such samples by law enforcement agencies will be motivated by non-microbial evidence (e.g. victim histories, witness testimony, tips, and other investigative leads and deductions) and effectively pre-supposes that the microbes found in those samples have a reasonably high *a priori* likelihood of being the ones involved in the crime or terrorism incident. Thus, the need for extremely high-resolution comparisons of sample DNA to establish relatedness between samples is generally (but not always) diminished. In fact, the resolution requirements here are similar to those for epidemiological strain typing, in that other evidence can outweigh differences in genetic fingerprinting when declaring that two samples are related to the same outbreak [10, 12].

An exception to this situation is where a pathogen that has been used in a crime or act of terror must be distinguished from a closely related one that exists as a natural background or as an endemic strain at relatively high levels in the geographical location of the incident. By "relatively high" we mean that there is a non-negligible probability that an environmental sample may contain the background pathogen, or that a clinically diagnosed infection may be due to the endemic pathogen. It is important to note that the background pathogen may be present either in a natural reservoir (including human reservoirs – e.g. HIV) or because of a previous release incident (e.g. anthrax in the US mail system.) In such cases, the highest resolution may be necessary to distinguish between the background and released strains when trace environmental or clinical DNA is compared.

The 2002 incident of plague cases reported in New York City illustrates some of these points[35]. Investigation of the New Mexican home of the infected patients, as well as personal testimony provided strong *a priori* evidence that the Y. *pestis* infections were acquired naturally. Pulsed Field Gel Electrophoresis (PFGE), a standard clinical typing technique, showed that the clinical Y. *pestis* isolates had the same band pattern as isolates from New Mexico, consistent with the likely *a priori* hypothesis. A higher resolution technique, MLVA, in fact revealed some detectable genetic differences among the strains, but these were

not considered significant in light of the other evidence supporting the *a priori* hypothesis of natural infection.

Strain typing is also used in a second, distinct investigative context in which DNA "fingerprints" of forensic samples are compared to those of a previously collected library of samples (or a data base of genetic typing information) to provide evidence that can narrow the range of *potential* primary sources of a bacterial or viral strain that has been used in a crime or biological terrorism incident. The term "primary source" may refer to an infected individual, isolates stored in a research laboratory, or an environmental reservoir. To facilitate this, both the law enforcement [6, 36] and the scientific community [7] have recommended the establishment of a national strain repository that contains:

- Samples of all select agent pathogens in the possession of U.S. microbiology laboratories, referenced to the laboratory from which they were obtained,

- Relevant isolates of select agents from well-defined natural reservoirs, along with geographic data on the extent of each reservoir,

- A database of genetic "fingerprints" of each sample in the collection that can be compared with the "fingerprints" of isolates obtained in the course of the investigation.

Ultimately, this system is envisioned to extend to pathogens obtained from sources from around the world, although the prospects of a comprehensive domestic database are much more likely within the foreseeable future. Much of the legal basis for such pre-event collection within the U.S. is contained in the legislation establishing the CDC select agent program[37], the Bioterrorism Preparedness and Response Act of 2002[20], and the related House Conference Report [21]. (Note, however, that clinical samples from human reservoirs of potential agents used in some bio-crimes, e.g. Human Immunodeficiency Virus (HIV) and Hepatitis C Virus (HCV), would not be included in this library.) Since pathogens stored in scientific laboratories and contained in natural reservoirs are considered to be likely targets for acquisition by criminals or terrorists, such a library is clearly an important tool to help investigators identify potential sources quickly. However, since the library may have been collected prior to any particular crime or incident, there is, in general, no *a priori* probability that any particular source is connected to a crime when it occurs. Hence, resolution of small genetic differences is essential to provide exclusionary evidence that reduces the number of plausible "primary sources" of the microbe. It is this use of genetic typing that forms the core of our discussion of resolution requirements.

It is seldom the case that only one or two possible sources of a given pathogen strain exist. This is believed by some to be true for smallpox, for example, since

the only known surviving viable isolates reside at the Centers for Disease Control in the United States and at two laboratories in the Russian Republic. More generally, however, a given pathogen may be distributed over many laboratory collections and separate reservoirs. Figure 2 shows a typical pattern of migration of a pathogen strain. Residing originally in a geographically distinct natural reservoir (Reservoir 0 in figure 2) it may migrate to other reservoirs via one of the myriad ways provided by the modern global system of transport of people, animals, and commodities. Historically, a number of pathogens, including B. *anthracis*, and more recently West Nile virus, have migrated this way. In addition, scientific collection efforts may bring isolates of a strain to a particular laboratory for study (Laboratory 0 in Figure 2) but, especially if it is a strain with unusual properties (including, e.g. high virulence,) isolates derived from the original culture may be sent to multiple other laboratories for study. These laboratories may, in turn, provide the strain to yet other laboratories. The most notable example to date may be the Ames anthrax strain, which was held by at least 7 geographically separate laboratories within 20 years of its original collection from a natural source[38]. Larger numbers of laboratories holding related strains of Y. *pestis*, *Brucella*, *Vibrio cholerae*, and other select agent pathogens have been registered with the CDC select agent program, although statistical data concerning the number of laboratories holding particular pathogens is not publicly available.

When pathogens are transferred from the wild to the laboratory, or from one laboratory to another, the isolates undergo one or more stages of plating and serial transfer in order to ensure that sufficient quantities of pure strain are held in stock. A population of pathogens generated by a single serial transfer can, depending on the length of time the colonies are allowed to grow, differ from the parent population from which it was derived by as many as 30 generations, although 20 is a commonly accepted "average" or typical number. Propagation on differing laboratory media, at different growth temperature, or other conditions can lead to eventual (unintentional) selection of mutations that shift the majority genotype away from that of the original "wild type" strain. Changes in DNA sequence can also be observed upon spread of an infection through a host population. Thus, genetic modifications become fixed in pathogens that migrate from one reservoir to another in the wild. It is not unusual to observe significant phenotypic differences after a few thousand generations in the laboratory, or a single passage through a host. Within a given laboratory, and sometimes within an infected host, it is even possible to have

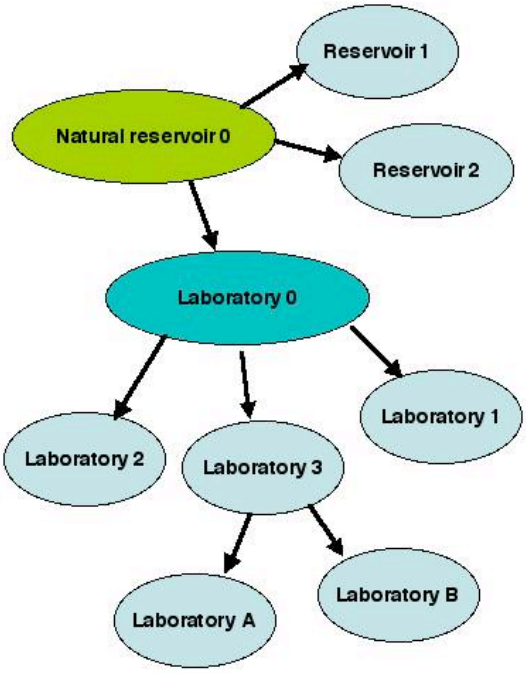

Figure 2. Propagation of a pathogen from a natural reservoir (0) to other geographically distinct reservoirs (1,2) and from the primary collecting laboratory (0) to other collaborating laboratories (1,2 … A,B). More complex exchanges are possible, but are not shown here. The set of isolates from all of these sources may form a set of "suspect lineages" in a bio-terror investigation.

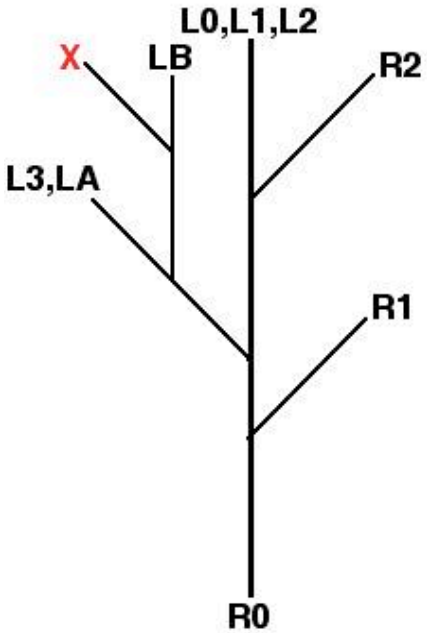

Figure 3. Hypothetical "suspect lineages" of strains from the potential primary sources referred to in figure 2. R0, R1 etc. refer to reservoirs, L0, L1 etc. to laboratories. The red X represents the attack strain involved in the chain of events shown in Figure 1. This strain is separated from the original R0 strain by 3 mutational events, and is derived by a single mutational event from strain LB.

several genetically distinct, but not phenotypically distinguishable lines originating from the same original isolate. (This possibility must be considered in any thorough investigation of a suspect laboratory or when strains from clinical isolates are compared.) In addition to genetic changes that occur during cell division, mutations can also apparently occur during stationary phases of microbial growth, as will be discussed below.

As an investigation (epidemiological or forensic) proceeds, a hierarchy of techniques will be used to identify the microbial agent first to the species level, and then to strain level. At each stage, more previously plausible suspect sources (or "suspect lineages") are effectively excluded. It is generally accepted that this process of elimination should occur at least to the level of species identification, and preferably to the strain level, on as short a timescale as possible in order to better guide treatment, quarantine and other public health decisions. However, the level of strain identification required for determining medical and public health response may have considerably less resolution than is necessary to leave a manageable number of plausible sources open for further law enforcement investigation.

The situation that exists after standard clinical-level identification of the attack strain may look like that illustrated in Figure 3. Here a number of sub-strains that are indistinguishable from the attack strain (but actually differ from it by a small, but as yet unknown number of mutations,) are associated with the array of potential sources described by Figure 2. Large clusters of unresolved isolates that match the attack strain are not necessarily a problem if they are all from the same location, i.e. lab or reservoir. But it seems likely that in practice the isolates in large unresolved strain clusters will come from *different* locations, due to natural infiltration of pathogens (e.g. anthrax) into geographically separated areas, or because of exchange of cultures from lab to lab. (Laboratory exchanges may not always be a matter of record, especially if they occurred prior to the passage of select agent legislation.) The largest unresolved clusters are likely to be "interesting" strains that are widely shared among members of the microbiological community.

In the absence of other evidence it becomes necessary to apply strain typing techniques that can distinguish smaller numbers of mutational differences among bacterial genomes to reduce the number of "suspect lineages" that connect plausible sources to the actual strain used in the attack. (Note that while multiple sources may possess strains with exactly identical genomes, it cannot be assumed that the attack strain exactly matches any of them, especially if it is derived from a clinical sample.) Clearly, if the number of unresolved lineages is small enough, direct sequencing of the entire genomes may be considered the simplest and most definitive approach to higher resolution. However, given the current cost and time needed for whole genome sequencing of bacteria, this number is likely to be quite small. It is probably safe to argue that in the anthrax

letters investigation, the availability of a strain typing assay that could quickly resolve even a few of the Ames isolates would have been considered highly valuable. However, in the absence of data on the diversity of the complete collection of other select agent strains currently held by U.S. laboratories, it is not clear if the resolving power of current typing systems for them would be considered adequate, or not.

As typing resolution increases, the probability of obtaining an "exact match" between the attack strain and one or more of the suspected source strains decreases. This is true even if one of the suspect source strains is, in fact, the direct progenitor of the attack strain, because of mutations accumulated during the generations of growth that separate them. On top of this, every typing system, including whole genome sequencing, has a finite error rate in declaring a match between two loci on different genomes. Thus, it is important to be able to estimate the probability that typing or sequencing error *alone* can explain the observed differences between fingerprints. In practice, the fingerprints derived from strain typing or sequencing are used to make two kinds of argument. The simplest is that the fingerprint of the attack strain is "closer" to certain of the suspect source strains than others, thus excluding the less well-matched sources from further consideration. A more sophisticated argument is that the attack strain is more likely to have been derived by mutation from certain (perhaps one) of the source strains than others. A variant of this second argument has been used in HIV cases where it is related to proofs of the direction of infection [39]. A discussion of the statistical issues involved in these arguments is beyond the scope of this paper. However, it is important to note that the admissibility of measures of the degree of relatedness of pathogens has been established in the courtroom[39].

3. Mutational spectra and rates

In this section, we will briefly review the current understanding of the possible types of genetic change that might occur between the organism obtained from the original source, and the organism(s) recovered from various locations in the generic scenario described in Figures 1, 2 and 3. The frequency at which different mutations are observed within a group of related bacteria is determined both by the rate at which they appear in the genome of the replicating cell and the permissiveness with which their environment allows them to be retained and propagated. Our knowledge regarding the kinds of mutations most likely to be observed is gained through comparative genomics of many bacteria. Rates of spontaneous genetic change are obtained through experiments in which bacteria are propagated over many generations. Both of these endeavors have made steady progress in recent years as DNA sequencing technologies and bacterial sequence datasets have become more widely available. The major types of spontaneous mutations that have been observed and studied are, in rough order of ubiquity:

Single nucleotide substitutions (SNSs). Substitutions do not change the length of the sequence in which they occur. Such mutations are assumed to be found with the highest probability on synonymous positions within open reading frames or within non-coding DNA because such substitutions are "neutral" and have little effect on phenotypic properties that can be acted on by selective pressures. At non-synonymous loci within open reading frames, where SNSs either change an amino acid or interrupt translation, selective pressures reduce the probability of observing them.

Single nucleotide insertions and deletions (Indels) These mutations are mediated by strand misalignment during replication, and are often associated with nearby direct repeats, monomeric runs, and palindromic repeats [40, 41]. When these occur in a gene, they are usually deleterious to gene function since they cause frame-shifts. In intergenic regions they may be neutral, or have subtle effects on gene regulation. They are likely to be observed only in genes that are not essential for bacterial survival in the environment in which the bacteria are propagating. Thus, the number of loci where single nucleotide indels will be observed is presumably much smaller than for SNSs. In this paper we will group single nucleotide insertions and deletions that occur within long monomeric tracts along with the VNTRs (see below). Thus the term "indels" will refer exclusively to single nucleotides at non-repeat or short repeat loci.

In the literature, SNSs and Indels are sometimes collectively referred to as Single Nucleotide Polymorphisms, or SNPs, although other authors appear to equate SNPs with SNSs alone.

Insertions/deletions at variable number tandem repeat (VNTR) and single nucleotide repeat (SNR) loci. These mutations occur within long segments of tandemly repeated short DNA sequences, including monomeric tracts [42, 43]. Such structures are often built into genes for surface proteins, for example, to provide mutational diversity that allows the population to respond quickly to environmental changes, including for example, host immune response [44-47]. The number of VNTR and SNR loci is much smaller than the number of genes in bacterial genomes.

Insertions/deletions of insertion sequence (IS) elements. Insertion sequences are small (< 2.5 kb) segments of DNA that encode a transposase gene, and are mobile, i.e. they are spontaneously inserted and deleted from among multiple loci on a bacterial genome [48-50]. They have been observed in a variety of pathogenic bacteria where they also play a role in providing diversity for quick environmental response. Several different insertion sequence types can exist in the same organism, and variation can include deletion, insertion, and movement of elements within the genome. More complicated transposable elements are also possible, collectively referred to as transposons[50]. Like VNTR loci, the number of IS elements and other transposons in a genome is much smaller than the number of genes.

Spontaneous deletions and sequence duplications. Segments of DNA that are not IS elements or tandem repeat units can also be spontaneously duplicated within, or deleted from, bacterial genomes. This process creates a class of "spaced repeats" or "interspersed repeats," i.e. multiple regions of identical sequence separated by intervening regions of unrelated sequence [51]. Deletion and duplication events involving these elements and/or their intervening sequences occur by various mechanisms [52]. The earliest use of such elements for typing were the REP and ERIC assays [53] The direct repeats (DR) region in the *M. tuberculosis* genome is characterized by the presence of multiple copies of a 36 bp repeated sequence separated by non-repetitive spacer sequences. Different strains of *M. tuberculosis* have different numbers of repeated elements and variable spacers, and this has been exploited to produce a strain typing method [54, 55]. Close repeats (CRs) are pairs of homologous segments 8 to 10 nucleotides long, separated by spacers several nucleotides long. Between 100 and 1000 CRs are present in typical bacterial genomes[56], although very little is known about their mutational activity.

Loss of plasmids is one of the most common large deletion events observed during bacterial growth, but large deletions from within chromosomal DNA are also possible. Very little data is available on the rates of these mutations, however a well-known 102 kbp deletion of the pigmentation locus in *Yersinia pestis* occurs with a spontaneous frequency of order 10^{-3}[57].

Sequence Inversions. These mutations involve the inversion of short DNA segments at specific sites within the genome. The inversion is catalyzed by site specific recombinases which can promote inversions at rates as high as 0.1 per generation[58, 59]. In the ideal form, sequence length is not altered. Along with VNTRs, large deletions and duplications, DNA inversions are a common "molecular switch" mechanism for controlling expression of bacterial phase variation[58].

Homologous recombination. This long-studied process results in the exchange of large DNA segments between distantly related parts of a microbial chromosome. It represents a mechanism for generating large scale genomic changes, including the movement or duplication of whole genes.

Finally, a rare, but possible mutational event is lateral transfer of DNA between bacterial pathogens through transformation or transduction mechanisms. For simplicity, we will not discuss these further, but note that there are only a few circumstances that would generally present opportunities for this kind of genetic exchange in the context of the events described in Figure 1. These circumstances apply to samples that have been obtained after passage through human or animal gut, or where the bacteria may have undergone vegetative growth in soil, sewage, or other environments where they commingled intimately with other bacteria. However, the possibility of such events could be important in

some investigations when it is necessary to distinguish natural from artificial changes in antibiotic resistance, or pathogenicity.

It is known that the distribution of mutations and mutation rates over the genome is not uniform. Some regions of the genome are highly conserved because they represent either genes for which mutations are generally deleterious to bacterial survival, or perhaps, error correction is especially high due to epigenetic mechanisms, or both. Such regions often contain "housekeeping genes" whose function is critical to bacterial metabolism. In these regions, synonymous SNSs are nearly the only observed mutation. Alternatively, "hot spots" or hyper-variable regions also exist. These regions are often associated with "contingency genes," which have evolved rapid mutational mechanisms to promote population survival in the face of rapidly changing environmental stresses such as the immune systems of hosts[44]. It is also important to note that the "deleterious" nature of many mutations depends critically on the environment in which the organism is growing. In a rich nutritional environment characteristic of laboratory cultures, mutations that disrupt the function of many genes may be effectively "neutral" because those genes are not essential to survival.

Experimental studies of mutational spectra imply that the number of times a specific type of genomic mutation is observed within a large clonal population, is a function of both the average rate of that type of mutation and the number of loci in which that type of mutation can be observed. An example drawn from the literature is shown in Figure 4. In this experiment, clones with spontaneous mutations in a particular gene (pyrE in *Sulfolobus acidocaldarius*) were isolated, and the sequence of the mutated gene determined[60]. These mutations are decidedly *not* neutral, since they strictly represent gene-disrupting events. Nonetheless, the spectrum observed here illustrates a general principle that is borne out by many other experiments on the mutation spectra of genes, and appears to apply equally well to the entire genome. Note that ≈70% of the observed mutations were single nucleotide insertions or deletions within a single "hot spot" locus (a monomeric poly-A repeat.) The remaining 30% of the mutations are SNSs, small indels, and other types distributed among many other loci within the gene. Thus, an assay that only detected mutations at the hot spot locus with 100% probability would, in fact, only detect mutations in the *gene* 70% of the time. Moreover, if we assume that the mutational rates associated with these mutations are typical of what can be found at similar loci across the entire genome, then the non-"hot spot" mutations must be a far more significant fraction of the total genomic mutation rate. That is, since the number of mutational "hot spots" is generally smaller than the number of genes, the occurrence of "slower" mutations such as SNSs and indels be a significant, if not dominant fraction of the total, because they can be found on a much larger number of loci within the genome.

Similarly, recent whole genome sequence comparisons of several Ames strain B. *anthracis* isolates by Read, et. al. show that a set of isolates whose lineages are derived from a common ancestor strain collected from the wild fewer than 25 years earlier had diverged after transfer to other laboratories, or diffusion to other natural reservoirs[61]. The mutations observed were SNPs, SNRs, VNTRs, small indels, and large deletions. The variety of observed mutational types is significant, since not all of the strains were completely sequenced at the time of this publication. It is likely that the number and variety of observed polymorphism will only increase when the complete genomes of all the isolates are compared. Thus, the Ames data is consistent with the argument that *many* kinds of mutation have significant probability of occurrence even in lineages separated by a relatively small number of generations.

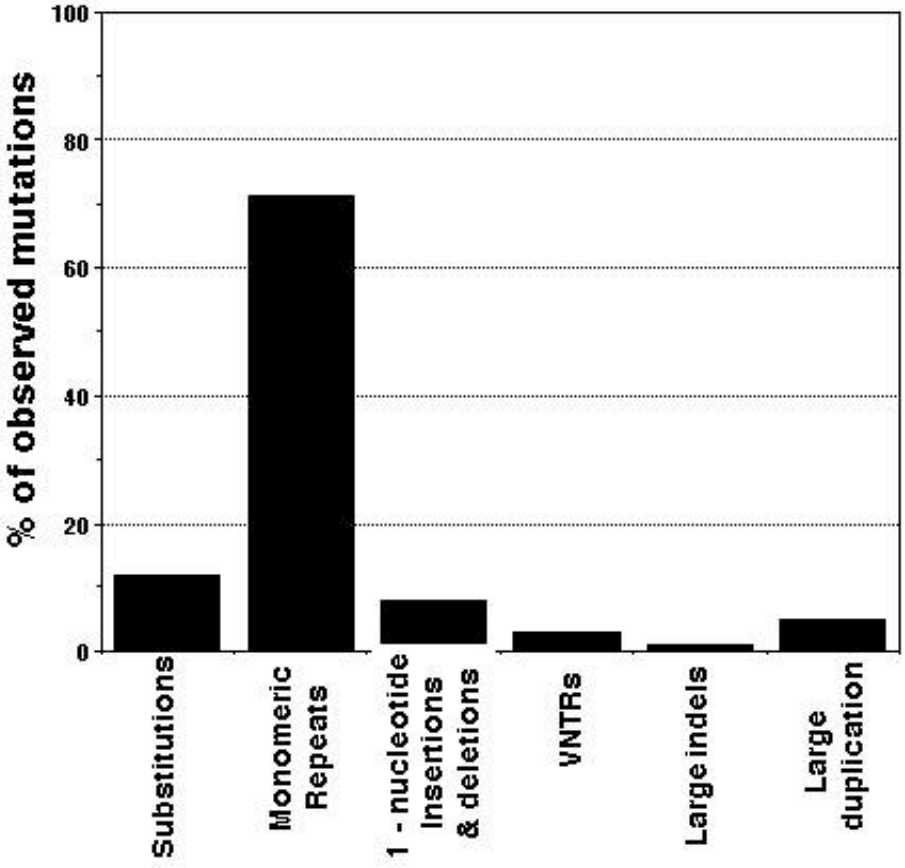

Figure 4. Observed mutation rate spectrum for the pyrE gene in S. acidocaldaris. Data from Grogan, et. al., [60].

We have attempted to collate and summarize the available data on the rates of mutations and the number of loci at which they can be found in table 1. It must be recognized at the outset that the values in table 1 represent only the broadest generalizations, and ignore considerable variation both within the genome of a

given organism and among those of different species. Although there is considerable information on the rates of single nucleotide substitutions in bacteria and viruses, there are serious gaps in experimental data for more complex mutations such as duplications, large deletions, and inversions. Very little data exists for select agent pathogens in particular, and our values are drawn from a wide range of other types of bacteria and viruses that have been studied in the literature. Nonetheless, the values in table 1 can be argued to be our current best guess, and perhaps the attendant uncertainties will stimulate further, more focused work in this area.

Typical SNS rates lie between 10^{-10} and 10^{-9} per generation per base pair in bacteria and DNA viruses[62]. It has been estimated that in RNA viruses in natural reservoirs substitutions at approximately 30% of genomic sites are neutral or under very weak selective pressure. However in bacteria, both the larger fraction of non-coding DNA and the permissive environment of nutrient rich laboratory cultures may permit a slightly higher fraction of effectively neutral sites. Thus, we have assumed that substitutions can be observed in a nominal 40% of the genome.

Table 1. Contribution of various types of mutation to the total genomic mutation rate.

	Type of mutation	Average rate per generation per locus γ_m	Nominal number of loci per genome[2] N_1	Contribution to the overall genomic mutation rate. Γ_m	References
Bacteria	SNS[1]	5×10^{-10}	2×10^{6}	1×10^{-3}	62
	Single nucleotide indels	5×10^{-10}	1×10^{5}	5×10^{-5}	-
	VNTR	3×10^{-5}	30	1×10^{-3}	63,64
	SNR	3×10^{-4}	10	3×10^{-3}	63
	IS elements	1×10^{-5}	30	3×10^{-4}	65 - 68
	Large deletions & duplications	$> 1 \times 10^{-6}$	30	$>3 \times 10^{-5}$	69
	Inversions	$>1 \times 10^{-5}$	10	$>1 \times 10^{-4}$	70
RNA viruses	SNS	3×10^{-5}	3×10^{3}	1×10^{-1}	71 - 73

[1]Single nucleotide substitution (primarily transitions)

[2]Assuming a genome size of 5×10^{6} bp for bacteria, 1×10^{4} bp for RNA viruses.

We could find no specific information on the observed rates of occurrence of neutral single nucleotide insertions and deletions in bacteria or viruses. Therefore in table 1 we *assume* that single nucleotide insertions and deletions (excepting those at monomeric repeat loci, which are lumped with VNTRs in this discussion) generally occur with rates similar to SNSs. In addition, we have assumed that the nominal number of loci at which small indels can be observed as fixed mutations is between one and two orders of magnitude smaller than for substitutions, due to the intrinsic frame shifting character of indels within open reading frames. More precise values await a systematic investigation of rates and available sequence data.

Mutation rates at VNTRs and SNRs in several select agent bacteria have been the focus of recent studies by Keim, et. al. VNTR and SNR loci exhibit varying mutation rates, with the fastest approaching 1×10^{-3} per generation[63,64]. Typically, greater than 10 and fewer than 100 VNTR loci show rates greater than 10^{-5} per generation and have significant diversity among bacterial strains of the same species. Between 3 and 30 high diversity SNR loci may be present in a bacterial genome. Thus, in table 1 the number of VNTR and SNR loci could be expanded, but at the expense of reducing the average rate.

The rates of insertion, deletion and rearrangement of IS elements have been studied in *E. coli* and *Mycobacterium tuberculosis*. A study of 10,000 generations of *E. coli* showed that a system of approximately 40 IS elements of various types had an overall mutation rate of $\approx 1 \times 10^{-3}$ per generation[65]. A study of 2000 generations of E. coli found IS mediated mutations in the Rbs operon occurred at a rate of 5×10^{-5} per generation[66]. IS element fingerprinting of *M. tuberculosis* has become a standard clinical typing method, and changes in IS fingerprint patterns are sometimes observed in samples drawn from the same patient. A rough estimate of 0.008 changes per IS element per year has been given in this bacterium[67]. If we assume 300 generations per year, this would also be consistent with transposition rates of order 10^{-5} per generation per element. Using data from Martusewitsch, et. al. rates between 10^{-4} and 10^{-6} per generation per element can be calculated for the bacterium *Sulfobulus sofataricus*[68]. Although it is known that *Y. pestis* and *B. anthracis* both possess IS elements, no rate data is currently available. However, it is clear from the above data that, at least in some bacteria, insertion sequence transposition rates can approach those of many VNTRs.

The rates of more complex mutations such as large deletions, duplications, and inversions are less well characterized. However, one study observed rates around 3×10^{-6} per generation for deletion between 101 bp non-palindromic repeats, and rates greater than 10^{-3} per generation for deletion between palindromic repeats on plasmids in E. coli [69]. Phase variation due to inversions

have been generically characterized as occurring at rates greater than 10^{-5} per generation[58,70].

The rates of spontaneous substitutions in RNA viruses are typically much faster than in bacteria, or DNA viruses. Overall mutation rates were estimated as nearly 1 per generation per genome for lytic viruses like poliomyelitis or influenza A[71,72], and ≈ 0.1 per generation per genome for retroviruses. Hepatitus C virus, which along with HIV has been the subject of forensic interest in criminal cases, exhibits mutation rates of order 10^{-5} per nucleotide per replication[73]. Studies of foot-and-mouth disease virus indicate that per mutation rates between 10^{-3} and 10^{-6} per site per replication are exhibited by this pathogen[74]. Very little information on mutations other than substitutions is available, and it may be that the compact nature of viral genomes does not permit a significant number of other mutational types. As we will show in the next section, the high overall mutation rate of RNA viruses makes it easier to achieve high resolution typing with assays that observe relatively small fractions of the entire genome.

The approximate contribution of each type of mutation listed in table 1 to the overall genomic mutation rate Γ_g is given by the product of the average mutation rate γ_m and the nominal number of loci N_l for that type of mutation. Thus, the overall genome mutation rate of the "typical" bacterium would be the sum of the Γ_g values in table 1, which is approximately 5×10^{-3} per generation. This value is somewhat larger than the "universal" value of 3.4×10^{-3} estimated by Drake based on substitution rates in DNA based micro-organisms[75]. However, given that we have attempted to include other mutations besides substitutions, the crudeness of the approximations in both calculations, and the uncertainty in many of the values, the order-of-magnitude agreement seems reasonable.

Finally, it should be mentioned that the above discussion explicitly assumes that mutation is associated with cell division. In fact, there is a body of evidence that mutations can also occur in stationary phase cells[76]. There is also evidence that stress conditions, such as the nutrient starvation that accompanies stationary phase, increase the rate of mutation[77]. The spectrum of stationary phase mutations can be very different from that associated with cell division, and the rates are calculated per time rather than per generation. In the case of E. coli, for example, rates of IS element rearrangements during stationary phase are of order 10^{-5} per hour of storage time[76]. Thus, for example, strains stored on agar stabs for a year could nominally accumulate as many mutations as a lineage ≈ 10,000 generations long.

4. A simple model for strain typing resolution

Typing systems provide two kinds of information about genetic relationships between isolates: The most basic information is whether the genomes of two isolates differ at all. Many typing systems go beyond this, and also identify the types and number of mutational differences. While this second capability is usually of greatest interest to microbiologists, typing system resolution is only dependent on the first. Thus, in this section and subsequent ones we will consider only the ability of a typing system to decide if two strains are genetically identical or not. This permits the formulation of resolution in terms of the statistics of simple binary hypothesis testing.

Moxon, et. al. developed a simple model to investigate certain statistical properties of mutations related to the existence of contingency and housekeeping loci within a bacterial genome[44]. We have utilized this basic picture to describe the ability of a strain typing system to detect mutational differences between clonally related isolates. To simplify this discussion, we assume that the typing system is error free in the sense that alleles are always correctly identified. Thus, any genetic difference detected by the typing system is always "real," and no false mis-matches are possible. In section 6 we will relax this assumption and examine the consequences of finite error rates.

In the following analysis, a typing system is said to "examine" a locus on a genome for particular types of mutation if it can detect those mutational changes at that locus. A typing system "reports" changes in the examined loci, but in some typing systems the actual locus where the change has taken place is not reported. For example, typing systems that rely on restriction fragment length determinations may examine all loci for which a mutational change will lead to a detectable change in one or more band positions, but which change has occurred at which locus is not always possible to deduce. By their nature, PCR based typing systems usually report both the locus and mutational type.

Consider a strain that is clonally related to a strain obtained from a primary source, as in figures 2 and 3. Along any lineage connecting the strain in question to the original source strain the probability of finding some number of mutations is a function of the mutation rates at various loci within the genome, and the number of generations that separate the two strains. A strain typing system examines a set of m_s loci, each of which is characterized by a mutation rate γ_{sj}, i.e. the number of mutations per generation. In addition to the set of m_s loci that are examined by the typing system, there is a set of m_u *un-examined* loci characterized by mutation rates γ_{uk}. The sum of m_s and m_u is the complete set of loci in the pathogen's genome, G.

$$m_s + m_u = G \tag{1}$$

In general, G is approximately twice the number of base pairs in the genome, representing each nucleotide position that can undergo a substitution, and every site between two nucleotides where an insertion or deletion can occur. In this model, absolutely conserved loci are defined by $\gamma = 0$.

The probability that particular loci (examined or unexamined) are unchanged after N generations is given by:

$$P_{s0j} = e^{-\gamma sjN} \tag{2a}$$

$$P_{u0k} = e^{-\gamma ukN} \tag{2b}$$

The probability that, after N generations, no mutation occurs among the m_s loci examined by the typing system is given by:

$$P_{s0} = \prod P_{s0j} = e^{-\Gamma sN} \tag{3}$$

Where

$$\Gamma_s = \Sigma \gamma_{sj} \tag{4}$$

Similarly, for the unexamined loci:

$$P_{u0} = \prod P_{u0k} = e^{-\Gamma uN} \tag{5}$$

Where

$$\Gamma_u = \Sigma \gamma_{uk} \tag{6}$$

The probability that no mutation has occurred anywhere on the genome is given by:

$$P_0 = P_{s0} \bullet P_{u0} = e^{-\Gamma g} \tag{7}$$

Where

$$\Gamma_g = \Gamma_s + \Gamma_u \tag{8}$$

is the whole genome mutation rate, as discussed by Drake[75].

The probability that there is at least one mutational difference between the genomes of two isolates separated by N generations is

$$P_1 = 1 - P_0. \quad (9)$$

Figure 4 shows the variation of P1 with the number of generations for a value of Γ_g approximately equal to Drake's "universal" value for bacteria and DNA viruses. It is clear that for two islolates separated by 100 generations or more, there is a high probability of finding at least one mutational difference. For RNA viruses, which have approximately 30 times higher mutation rates, the probability is much higher.

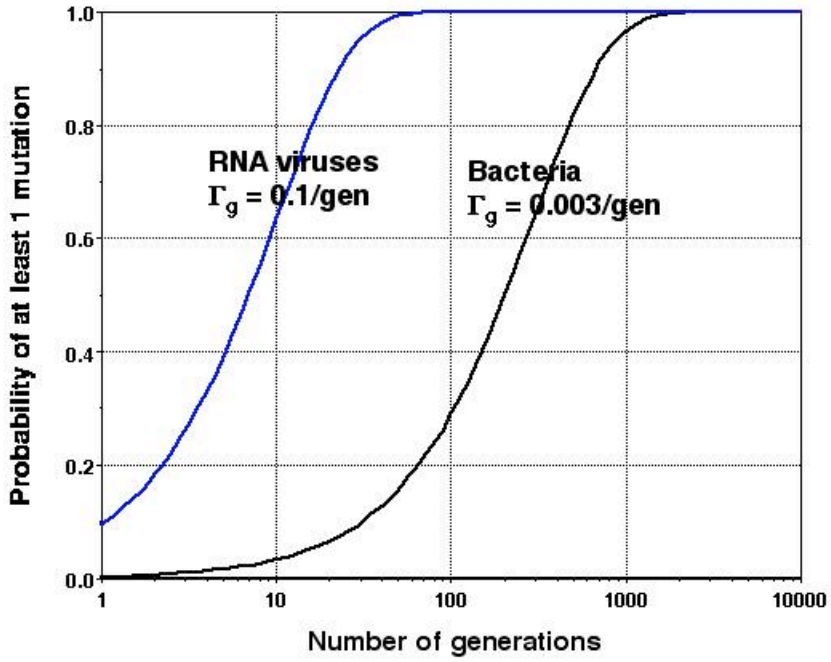

Figure 4. Probability of finding at least 1 mutation in a lineage of N generations.

The conditional probability that the typing system will detect a mutation that has occurred within the genome after N generations, *given* that there is at least one mutation somewhere on the genome, is simply the probability that at least one mutation has occurred among the m_s examined loci divided by the probability that there is at least one mutation on the entire genome:

$$P_d = (1 - P_{0s})/(1 - P_0) \quad (10)$$

Conversely, the probability that the typing system will falsely declare as identical two isolates separated by N generations, even though there is at least one mutational difference somewhere on the genome is:

$$P_f = P_{0s}(1 - P_{0u})/(1 - P_0). \quad (11)$$

Normalization by $(1 - P_0)$ in equations 10 and 11 is necessary to adjust P_d and P_f for cases where there is no mutation in *either* the examined or unexamined parts of the genome. Its inclusion makes P_d and P_f conditional on the existence of at least one mutation being present. Note that

$$P_d + P_f = 1. \quad (12)$$

Using the definition

$$\alpha = \Gamma_s/\Gamma_g \quad (13)$$

Equation 10 can be written in terms of α, the entire genome mutation rate Γ_g, and the number of generations N:

$$P_d = (1 - e^{-\alpha\Gamma gN})/(1 - e^{-\Gamma gN}) \quad (14)$$

The quantity α is essentially the fraction of the entire genome mutation rate that is captured by the typing system that examines a set of m_s loci characterized by a set of mutational rates $\{\gamma_j\}$. This can be regarded as a figure of merit that describes the resolving ability of the typing system.

The influence of α on the resolution of the typing system (now defined in terms of the probability that the system will correctly recognize when two isolates have at least one mutational difference,) is shown in Figure 5, which illustrates two points. First, for resolving isolates that are separated by fewer than a few thousand generations, it is necessary for the typing system to have an α value of 0.1 or higher. Second, note that for $N < 100$ generations, the value of P_d is asymptotic to α. Thus α is essentially the probability of detecting a single mutational difference if it occurs at all within the genome.

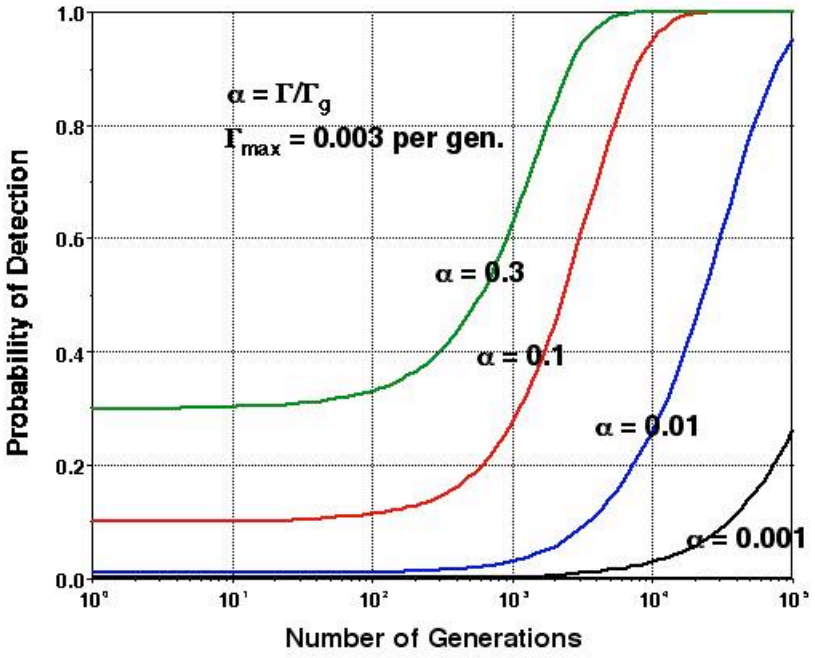

Figure 5. P_d as a function of N for various values of α.

It might be noted that the expression of P_d in terms of α and Γ_g is somewhat arbitrary, and that use of Γ_s and Γ_g would be equally valid. However, the former choice of variables is conceptually convenient because α has a natural interpretation in terms of probability, as explained above. A typing system with a figure-of-merit α can resolve two isolates separated by N_α generations with high probability if

$$N_\alpha \geq (\alpha\Gamma_g)^{-1} \tag{15}$$

Since the resolution of a typing system depends on both α and the genomic mutation rate Γ_g, it could be argued that α is not a good figure of merit because it does not cleanly separate the technology from the organisms it is applied to. However, although Γ_g will, strictly speaking, vary from bacterium to bacterium, the work of Drake implies that this variation is small[75]. Thus, for purposes of comparing typing systems, it seems quite valid to assume a common, constant bacterial or viral Γ_g value. It is important to recognize, however, that α is not necessarily transferable from one class of organisms to another, e.g. bacteria to viruses, even if the same types of mutations are probed by the typing system.

Typing systems for RNA viruses typically rely on sequencing of just one or two genes or gene fragments, usually amounting to no more than 10% of a (nominal 10kbp) viral genome, and a nominal α value of ≈ 0.1. Nonetheless, isolates separated by fewer than 100 generations are easily resolved. The reason for this

is illustrated in Figure 6, which shows the effect of the overall genomic mutation rate Γ_g on the resolution power of typing systems with equivalent α values.

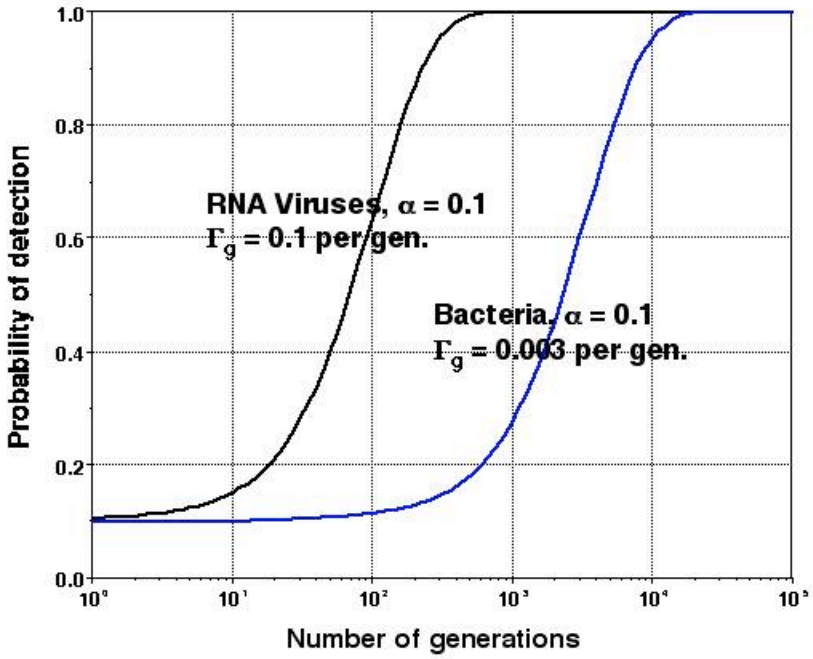

Figure 6. Comparison of typing systems with equivalent α values applied to bacteria and viruses.

The definition of Γ_s in equation 4 provides a basis for discussing the potential for increasing the resolution of strain typing by combining several techniques. Clearly, if two techniques examine completely independent sets of loci and mutational types, then the Γ_s values of each can be added together to obtain the Γ_s (and α) of the composite technique. Thus, for example, combining a typing system that looks exclusively at VNTRs with one that looks exclusively at SNSs mutations will additively increase the value of α. However, more complex typing systems that examine intersecting sets of mutations will not, in general be additive. To calculate the α value of the composite system it is first necessary to identify all of the examined loci to eliminate double counting of the mutations in common.

Finally, the relative value of incrementing α either by combining typing systems, or by substituting one system for another can be understood from equation 14. In figure 7 we show the how the resolution of a bacterial typing system (expressed in terms of the number of generations separating two isolates that can be resolved with 95% probability) varies with α. For α values smaller than 0.1, the increase of resolution with α is very steep. But once α is larger than 0.1 The resolvable lineage length decreases roughly linearly, until α approaches its

limiting value of 0.95. Thus, doubling a from 0.1 to 0.2 reduces the number of generations from ≈10,000 to ≈5000.

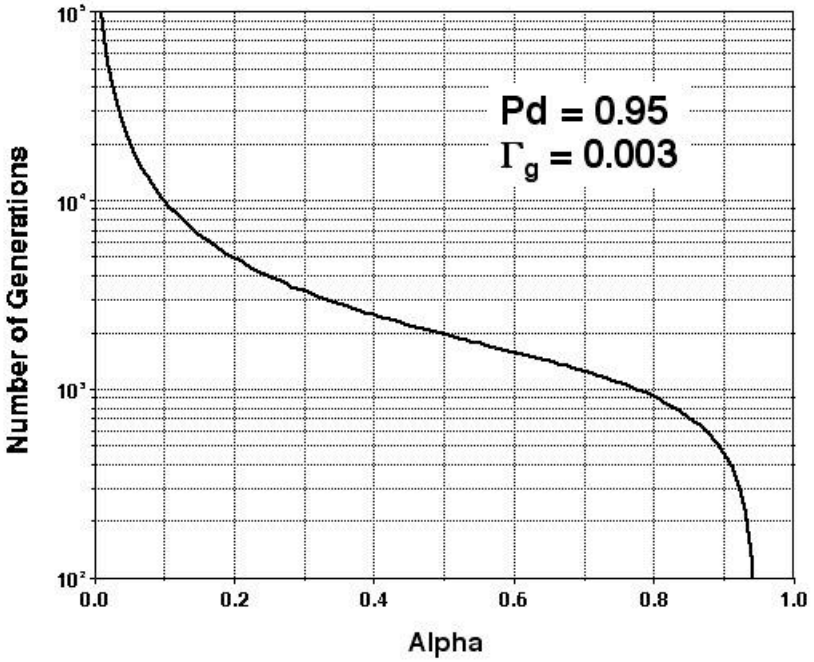

Figure 7. Resolution versus α. Resolution is expressed as the number of generations separating two resolvable isolates. The detection probability has been fixed at 0.95.

As we will discuss in the next section, the typing systems that have been developed in practice typically select certain classes of mutations and examine a certain number of mutational loci in that class. Before estimating the α values of those systems, it is of interest to formulate an idealized typing system that is optimized in the sense that it possesses the largest possible α value for a given number of examined loci. The experimental picture of mutation rates summarized in section 3 suggests that there is probably considerable overlap among the mutation rates belonging to different classes of mutations. Thus, for example, some IS element transpositions may be faster than some VNTRs, and the rates of certain duplications may exceed the rates of certain deletions, and so forth. The optimized typing system is simply one that that examines the m fastest mutational loci, regardless of class. We can thus imagine that any given microbial genome is characterized by an ordered set of mutational loci starting with the most rapidly mutating locus and continuing the series according to decreasing mutation rate. In Appendix 1 we formulate a model for such an ordered set of mutation rates that is consistent with the known empirical constraints on the fastest individual locus, and the overall genomic mutation rate. For an optimized typing system that examines m loci, α is calculated by summing the rates of the first m loci, according to equation (4). The overall

mutation rate Γ_g is simply the sum of all the rates in the series, up to the total number of loci.

This idealized optimal typing system plays an important role in our discussion of error rates in section 6. It allows us to establish a plausible systematic relationship between α and the number of examined loci, which affects the accumulation of errors defined by per locus error rates. Typing systems that are not optimized in this sense accumulate more error for the same α value. Thus, the calculations in section 6 represent the best that could be achieved for the given error rates.

While α has great utility for defining resolution, it should be noted that two typing systems with equal α values are not necessarily equally useful. For example, consider a typing system based on a single mutation that causes phase variation in a bacterium [58,59]. The extremely high mutation rate associated with this marker may represent a very high fraction of the overall mutation rate, and yet its utility for differentiating lineages is clearly questionable compared to a typing system that derives the same α value by interrogating a large number of more slowly changing loci. Part of this difference in utility stems from the problem of homoplasy in quickly mutating loci, but more fundamentally it is due to the lack of significance of observing that two isolates differ by a change at a single, very mutationally labile site.

This latter notion can be captured quantitatively by the probabilistic concept of "surprise" which is related to Shannon entropy[78]. In Appendix 2 we derive a second figure of merit by taking this approach. This F.O.M., designated β is a measure of the phylogenetic depth of the typing system and is complementary to α. In Appendix 2 we present a crude mapping of some existing typing systems onto the space defined by α and β.

5. Estimates of α for some current strain typing methods

In this section we will estimate the resolution figure-of-merit values for several strain typing methods. The number of strain typing methods reported in the literature is too large to permit explicit treatment of all of them. Thus, we have restricted our analysis to four commonly used techniques - MLST, MLVA, RFLP and AFLP - and one technique that is currently the subject of intense R&D: hybridization arrays. Even with this restriction, there is not enough information on exact mutation rates and numbers of mutational loci available to permit precise calculations of α. Therefore we have relied on the rough estimates of rates and numbers of mutational loci given in table 2. As a result, only order-of-magnitude differences should be considered significant, and it is not possible to discern differences that might make one technique better than another for certain species, for example. Although the resulting comparison is quite crude, it

illustrates the method and paves the way for more refined calculations as experimental data of more breadth and precision become available. In addition, the analysis is useful for underscoring the kinds of information needed from experiments and bio-informatics analysis to support the design of optimized high resolution typing assays.

Our choice of methods covers the most common strategies for obtaining genetic typing information, and includes both PCR and restriction fragment based approaches. Our representation of these methods is idealized. We do not consider the practical aspects of their implementation, even if these can sometimes influence resolution (e.g. gel uniformity for gel-based RFLP methods.) In each case, we calculate Γ_s by asking which mutations will lead to a detectable difference in the allele or band pattern examined by the typing method. We then add up the assumed rate constants for those mutations that will be so detected.

Multi-locus sequence typing. (MLST) This method chooses a small number (usually 6 – 10) of specific genes[79,80]. Highly conserved regions flanking the genes, or segments of the genes, are identified and PCR primers are designed to amplify the intervening genes or gene segments. The resulting amplicons are sequenced, and the sequence itself is the allele. To produce a reliable MLST assay considerable sequence information must be available for at least one strain (the reference strain) that is reasonably close to the cluster of strains one is trying to resolve, so that appropriate conserved regions for primers can be identified. Typically, the set of genes are chosen from among the highly conserved "housekeeping genes" of the reference organism so that there is a very high probability that the same primer sequences will be present in the test strains. This, of course, also typically constrains the amount of variation that can be observed in the intervening sequences. Thus, the mutational spectrum is, with very high likelihood, constrained to SNSs. A typical MLST assay may generate amplicons around 500 bp long for 7 housekeeping genes, constituting a total of 3500 loci. Of these loci we can expect ≈ 40% to show variation (as discussed in section 3,) thus $\Gamma_s \approx 3500 \times 0.4 \times 5 \times 10^{-10} = 7 \times 10^{-7}$ per generation. If we use Drake's value for $\Gamma_g = 0.0034$, then $\alpha = \Gamma_s/\Gamma_g = 2 \times 10^{-4}$.

Multi-locus VNTR analysis (MLVA). In this technique, reference sequence information is used to identify VNTR loci, and high diversity VNTR loci are chosen for the assay since they are thought to be likely to have high mutation rates[63,64]. Conserved regions flanking each chosen VNTR are used to design primers for PCR amplification. Like MLST, the reliability of the assay is based on the relatively low probability of mutations occurring in the primer regions. The length of the amplicons generated by the PCR amplification of the region surrounding the VNTR are the alleles. A typical MLVA assay may interrogate ≈ 30 sites on the genome, so that $\Gamma_s \approx 30 \times 3 \times 10^{-5} \approx 1 \times 10^{-3}$ per generation. Thus, $\alpha \approx 0.3$. MLVA type assays that interrogate SNRs in addition to VNTRs will have

higher α values. For example, adding 5 SNR loci with average mutation rates of 1×10^{-4} per generation will increase α to ≈ 0.5 [63].

Restriction fragment length polymorphism (RFLP). Unlike MLST and MLVA, the RFLP method needs no reference sequence information. Instead it makes use of the fact that a significant number of sequences recognized by restriction enzymes are present in all bacterial genomes, and are relatively uniformly distributed over the whole sequence. A restriction enzyme is used to cut the DNA into fragments, and one of several methods are used to measure the fragment lengths. With the exception of substitutions that, by chance create or remove a restriction site, this form of RFLP will not detect any SNSs, since they do not change the length of the sequences in between the restriction sites. However, *any* mutation involving insertion or deletion could, *in principle*, be detected. This includes VNTRs, SNRs, IS elements, larger indels, and duplications. Thus, referring to the values in table 2, this idealized version of RFLP would have an α value close to 1. (To calculate this accurately one would need better values for both the individual mutation rates and the overall genomic mutation rate.) In practice, however, a number of factors greatly reduce the resolving power of RFLP.

One factor that reduces RFLP typing resolution is the fragment length resolution that is characteristic of practical gel or CE based methods. Since it is difficult to resolve fragment lengths to better than 1 nucleotide out of 1000, it is not possible to detect changes in VNTR or SNR loci for a significant number of bands in a typical RFLP digest. The distribution of fragment sizes can be modified to produce more fragments with smaller average length by using more frequently cutting restriction enzyme (say one that recognizes a 4 bp restriction sequence instead of a 6 or 8 bp sequence.) However, fragments greater than 1000 bp long will still comprise a significant fraction of the distribution.

A second factor is that not all fragments are examined by the typing system. This can be due to high and low fragment length cutoff values (high molecular weight bands that are poorly resolved are often left out of RFLP analysis,) or because of other filtering mechanisms that select only certain fragments for examination. A typical example is the use of IS element hybridization probes in the RFLP analysis of *Mycobacterium tuberculosis*[22]. The *M. tuberculosis* genome may possess ≈5000 restriction sites, but at most 30 or so IS6110 elements. Clearly, the lengths of only 30 of the possible 5000 fragments will be examined in this assay (in fact, we can generally expect fewer than 30 bands because by chance some will contain more than one IS6110 element and be doubly labeled.) Thus, the value of α is reduced greatly since mutations that reside in the fragments that do not contain IS6110 elements cannot be detected. A lower bound on α can be estimated from this technique by assuming that the only mutations detected are those that involve insertion, deletion or transposition of

the IS6110 element itself. For 30 IS elements with an average γ of 1×10^{-5}, $\alpha \approx 0.09$.

Amplified restriction fragment polymorphism (AFLP). This complex variant of RFLP involves both restriction fragmentation of the genome and selective PCR amplification of a selected set of fragments. The genome is cut with two restriction enzymes, one that recognizes a 6 bp restriction site and one that recognizes a 4 bp site. PCR primer recognition sequences are ligated to each end of the fragments[81]. In practice, only fragments that are terminated by the 6 - site on one end, and by the 4-site at the other end are amplified by the subsequent PCR protocol. Since the 6-sites are far less numerous than the 4-sites, the number of fragments that are selected by this method reflects the number of 6-sites. The average length of these fragments however, is a reflection of the number of 4-sites. In the absence of any other selection, a bacterial genome might yield 1000 - 2000 such fragments with an average length around 300 bp. However, it is common to design primers that hybridize well only to fragments that have particular single or double nucleotide subsequences at each end[81]. In the case that single nucleotides are selected, the number of amplified fragments is reduced by a factor of 1/16, while the number is reduced by 1/256 for double nucleotide selective primers. Typically between 50 and 100 fragments are amplified and then sized by gel electrophoresis or capillary electrophoresis.

For a *single* AFLP assay then, a $\approx 3 \times 10^4$ bp (non-contiguous) section of the genome is examined. Like RFLP, SNSs will not be detected, but all possible insertions and deletions can be detected in principle. Thus, for a single AFLP assay the expectation value of α must be very close to $3 \times 10^4 / 5 \times 10^6 = 6 \times 10^{-3}$, for our nominal 5 million bp genome. This seems at first glance to be only slightly better than the resolving power of MLST. However, many different AFLP assays can be applied to the same genome by changing the selective primers and by choosing different pairs of restriction enzymes. In this way, much larger sections of the genome can be examined and the value of α increased. Using assays that cover all possible singly selective primer pairs, for example, should roughly raise α by a factor of 16, since independent fragments are probed by each combination. On the other hand, since the location of the various restriction sites is random, there will be overlap between the parts of the genome covered by different restriction enzyme pairs. Thus, additivity would not be valid, α would increase sub-linearly as additional assays are included, and the computation of and exact value of α for a set of assays using different enzyme pairs would be complicated.

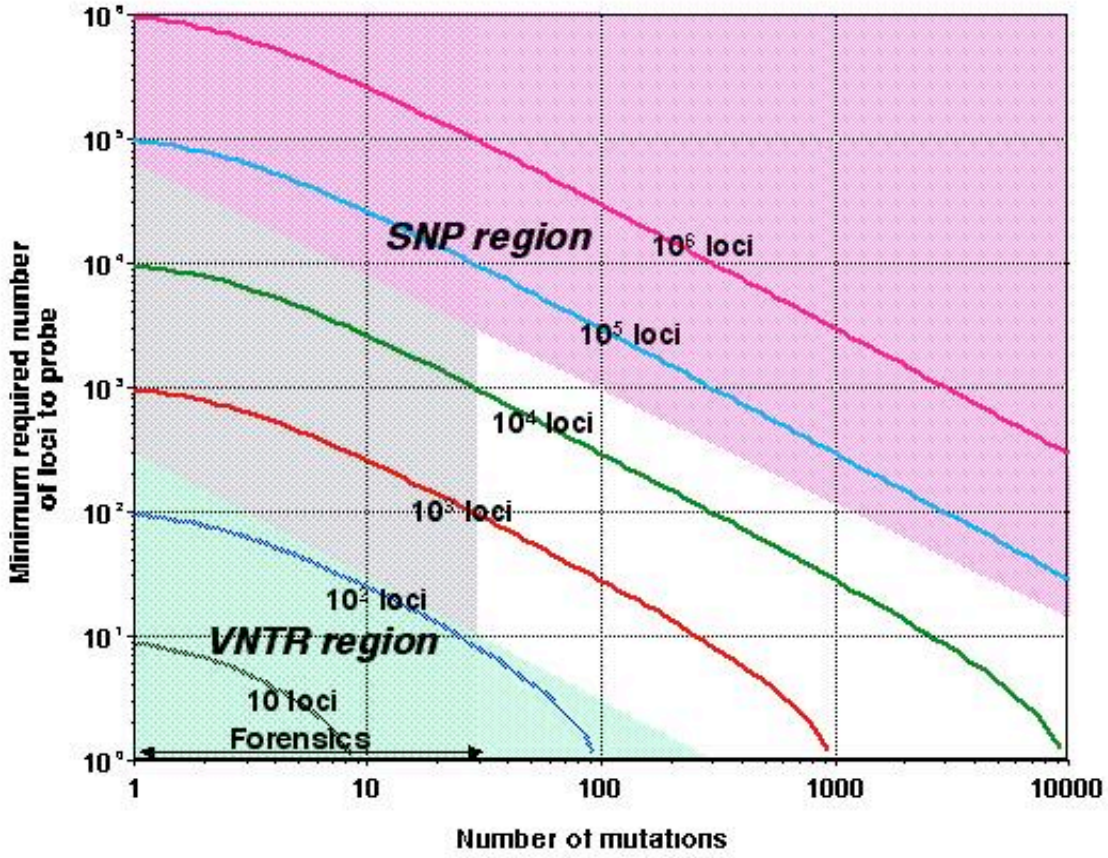

Figure 8. Minimum number of loci that would need to be probed in order to detect at least one mutation if M mutations are distributed over L loci. The detection probablility was fixed at 90%. The region in gray represents the small number of mutations that might separate suspect isolates in a forensic situation.

High Density Hybridization Arrays. Oligonucleotide micro-arrays are an emerging technology for mutational screening of Human DNA as well as bacteria and viruses[83-87]. They are one of the few *potentially* practical ways of detecting SNSs over significant fractions of the bacterial genome without direct sequencing. Because SNSs are distributed with fairly uniform probability over such a large fraction of the genome, and because the probability of a substitution occurring at a given site is so low, only a technique that monitors a very large number of loci can hope to detect a few SNSs occurring within a short lineage. Figure 8 shows the number of loci it is necessary to examine in order to detect at least one of M mutations uniformly distributed over L loci at random. Clearly, it is necessary to examine of order L loci in order to be confident of detecting at least one of a few mutations. Hybridization arrays achieve this by printing millions of microscopic patches of short (25-75 bp) segments of an entire reference genome onto a glass slide. Segments of the genome of a test strain that match perfectly to some reference segment will hybridize to it, while segments that don't match will hybridize only weakly or not at all. A chip reader is used to determine the number and location of weakly or non-hybridized

patches, which closely indicate the number and types of mutation present in the test strain. To use such a chip in the forensic scenario above, it would be necessary to sequence 1 genome (say the attack strain) and use it as the reference. The unresolved cluster of strains that remain after conventional typing would then be tested against the resulting chip.

In principle, variants of this technique could be used to provide the entire genetic sequence of a pathogen, and thus would have $\alpha \approx 1$. However, this has not yet been achieved in practice. However arrays that determine genome-wide SNSs and short indels, as well as other types of mutation have been demonstrated. Referring to the model in Appendix 1, a tiled microarray configured to detect all possible single nucleotide substitutions and short indels (several million loci corresponding to a very large, high density array) might achieve an α of ≈ 0.3 and would be completely complimentary to typing systems that target the relatively small number of other, faster mutations. However, this technology suffers from significant occurrences of cases where matching sequences fail to hybridize with the reference oligomers for various reasons. Such failures are indistinguishable from the presence of mutations, and thus constitute false positive detections of genetic differences. As will be discussed in sections 6 and 7, the effectiveness of hybridization array technology for high resolution forensic typing will depend critically on managing these error rates.

6. The effect of errors on typing resolution

In the previous sections we have implicitly assumed that the typing system was error-free in the sense that any mutation that occurred among the examined loci would be detected, and only mutations that actually occurred would be reported. Thus, a "false match" can occur only because all mutations are present among the unexamined loci, and "false mis-match" cannot occur at all. In reality, of course, there is a finite (if small) probability that a real change among the examined loci will be "missed" by an assay, or that a change of allele at some locus will be reported, even no such change has occurred. Therefore, in this section we examine the consequences of such errors.

To begin we will re-cast the results of section 4 in terms of conditional probabilities associated with binary hypothesis testing. The outcome of a typing assay performed on a pair of isolates that form a "suspect lineage" is either D_0, the decision that they have identical genomes, or D_1, the decision that they differ by one or more mutation. Let H_0 be the hypothesis that two genomes are identical, that is, that no mutations have occurred after N_g generations. Similarly, H_1 is the hypothesis that they differ by at least one mutation. P_0 as defined in section 4 is the a priori probability associated with H_0, i.e. $P(H_0) = P_0$, and $P(H_1) = 1 - P_0$. In addition, let h_0 be the hypothesis that no mutation has occurred among the loci examined by a typing system. From section 4, p_{0s} is the

probability associated with h_0. The hypothesis that at least one mutation has occurred among the examined loci is h_1. P_d can be recognized as the conditional probability that at least one mutation has occurred among the examined loci given that H_1 is true. Thus:

$$P(h_1 \mid H_1) = P_d \tag{16}$$

Similarly, P_f is the conditional probability that no mutations have occurred among the examined loci given that H_1 is true:

$$P(h_0 \mid H_1) = P_f \tag{17}$$

Since it is always true that h_0 is implied by H_0 we have:

$$P(h_0 \mid H_0) = 1 \tag{18}$$

Similarly, h_1 can never be true if H_0 is, thus:

$$P(h_1 \mid H_0) = 0 \tag{19}$$

We will consider two types of error associated with each locus that is examined by a typing system. Let ε_{1j} be the probability that the typing system reports a change in locus j at which no mutation has actually occurred (false detection.) Similarly, let ε_{2j} be the probability that the typing system fails to report a change in locus j at which a mutation has, in fact, occurred (missed detection.) The probability that the typing system correctly reports that no change has occurred among the m_s examined loci, given the hypothesis h_0 (that no change has actually occurred) is just the probability that each locus is reported correctly:

$$P(D_0 \mid h_0) = \Pi(1 - \varepsilon_{1j}) = P_c \tag{20}$$

The probability that the typing system mistakenly reports that at least one change has occurred among the examined loci, given h_0 is:

$$P(D_1 \mid h_0) = 1 - P_c \tag{21}$$

For simplicity, we will assume throughout this analysis that all ε_{1j} are equal, and denote this error as ε_1.

Calculating the conditional probability of missed detections is more complex. In Appendix 3 we show that the probability that the typing system falsely reports

that no mutations have occurred among the m_s examined loci when at least one mutation has occurred is approximated by:

$$P(D_0 \mid h_1) \approx \varepsilon_2, \tag{22}$$

Where ε_2 is the average value of ε_{2j} among the examined loci. Given this approximation, we also have:

$$P(D_1 \mid h_1) = 1 - \varepsilon_2. \tag{23}$$

Now consider the case where there are no mutational differences between the two isolates (H_0). The conditional probability that the outcome of the typing assay will result in the decision that the two isolates are identical is given by:

$$P(D_0 \mid H_0) = P(D_0 \mid h_0)P(h_0 \mid H_0) + P(D_0 \mid h_1)P(h_1 \mid H_0) = P_c \tag{24}$$

Similarly, the conditional probabilities for the other possible outcomes given the hypotheses H_0 or H_1 are given by:

$$P(D_1 \mid H_0) = (1 - P_c) \tag{25}$$

$$P(D_0 \mid H_1) = P_c P_f + \varepsilon_2 P_d \tag{26}$$

$$P(D_1 \mid H_1) = (1 - P_c)P_f + (1-\varepsilon_2)Pd. \tag{27}$$

Of primary interest are the probabilities of type I errors (false indication of mis-match) and Type II errors (false indication of match) that are given by equations (25) and (26) respectively. Note that the probability of false mis-match depends on the number of loci that are examined, through equation (22). This gives it a parametric dependence on the resolution parameter α, such as that implied by the model for γ_j developed in Appendix 1. In contrast, the probability of false match primarily depends on P_f, the probability that mutations have occurred in the un-examined part of the genome, assuming that ε_2 and ε_1 are small quantities. The behavior of the type I and II errors as a function of the number of loci that are examined by the typing system are shown in Figure 9. These calculations utilize the model in Appendix 1, and set $\varepsilon_1 = \varepsilon_2 = 1 \times 10^{-6}$. Note that the probability of obtaining a false match rapidly decreases as α increases, reaches a plateau in the region where α is approximately constant, then falls to nearly zero (actually to ε_2) as α approaches 1. The probability of obtaining a false match does not achieve significance until a very large number of loci are examined. Thus, for typing systems that examine considerably fewer loci than

whole genome sequencing type II errors are most significant, while typing systems that examine very large numbers of loci, e.g. hybridization arrays, primarily suffer from type I errors.

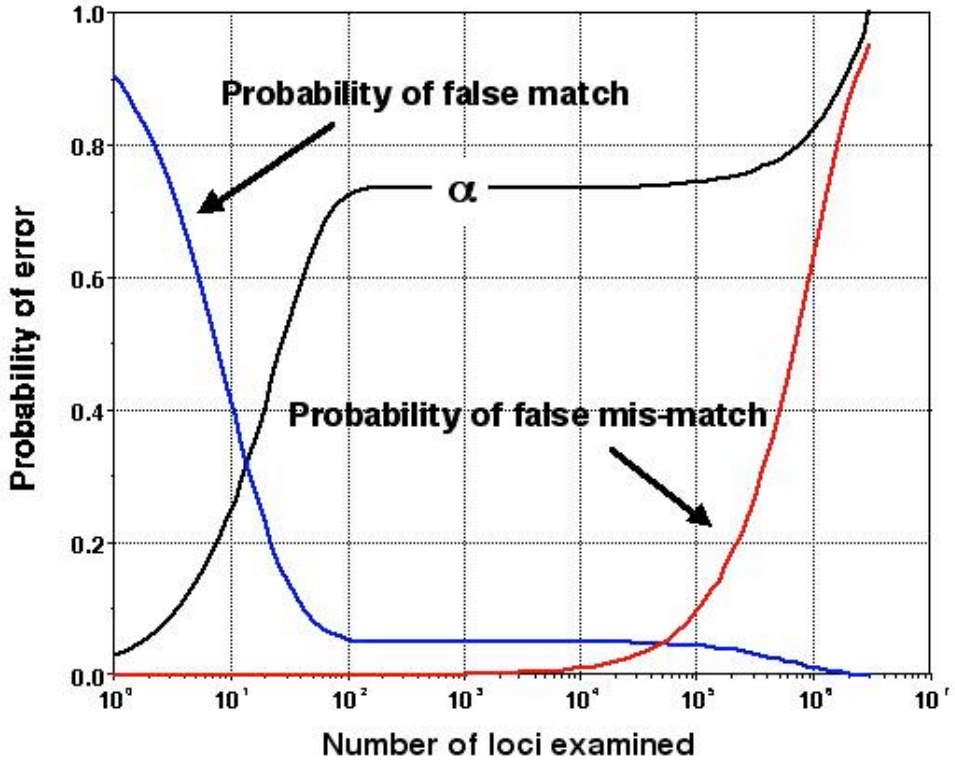

Figure 9. Probability of false match (Type II) and false mis-match (Type I) errors as a function of the number of loci examined by a typing system. The calculation uses the model for relating a to the number of loci developed in Appendix 2. A total of 3×10^6 loci are assumed, and the two isolates are separated by 1000 generations.

One way to illustrate the effect of type I errors on resolution is to determine the minimum number of generations that must separate two isolates be before we are confident that a positive detection indicates a real difference and not simply typing error. To do this, we must consider $P(H_1 \mid D_1)$, the probability that a detection of a mutational difference reported by the typing system implies a real difference between two isolates. According to Bayes's theorem,

$$P(H_1 \mid D_1) = P(D_1 \mid H_1)P(H_1)/P(D_1) \tag{28}$$

and

$$P(D_1) = P(D_1 \mid H_0)P(H_0) + P(D_1 \mid H_1)P(H_1) \tag{29}$$

In the error free case, $\varepsilon_1 = 0$, $P(H_1 | D_1) = 1$, regardless of N_g. Figure 10 shows how increasing values of ε_1 force the minimum number of generations to higher values in order to guarantee that a reported difference implies a real mutational difference with 95% probability.

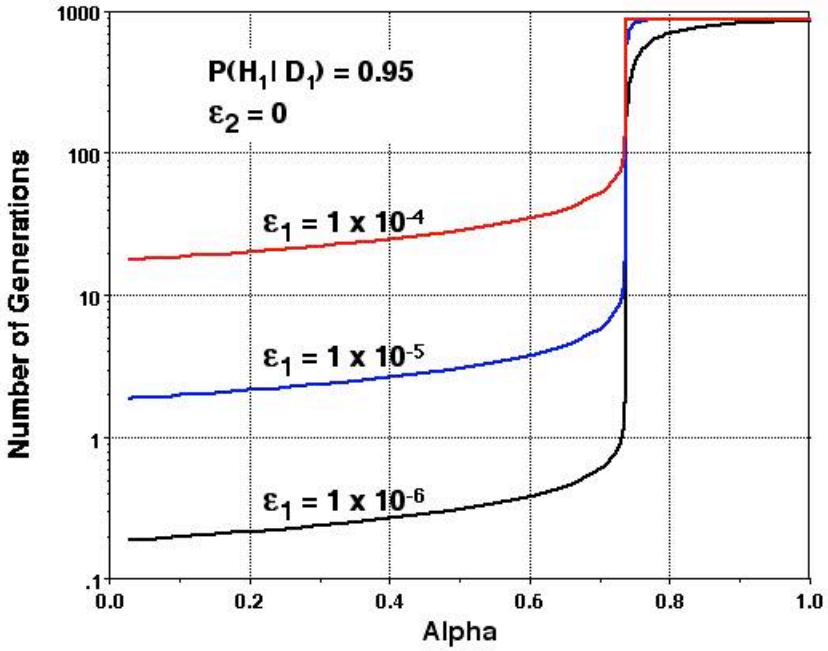

Figure 10. Minimum number of generations required to guarantee $P(H_1 | D_1) = 0.95$, for various values of ε_1. The model of Appendix 1 was used to relate α to the number of loci probed by the typing system.

The very steep rise around $\alpha \approx 0.7$ represents the region where the number of loci examined by the typing system continues to increase but the typing system figure of merit is not changing significantly. In the $\alpha > 0.7$ region, the number of generations is given by:

$$N_g \approx \ln[1 + (P_{H1D1}/(1 - P_{H1D1}))(1 - P_c)]/\Gamma_g \qquad (30)$$

Here $P_{H1D1} = P(H_1 | D_1) = 0.95$, and $\Gamma_g = 0.0034$ per generation. This equation is exact when $\alpha = 1$, in which case P_c is evaluated for ε_1 and the total number of examined loci. For the values of ε_1 used to generate Figure 10, $P_c \approx 0$ for $\alpha > 0.7$ and $N_g \approx 880$. More stringent criterion for $P(H_1 | D_1)$ will result in higher values of N_g, i.e. isolates will have to be separated by more generations before it is more than 95% likely that the typing system will definitively indicate they are genetically different.

Thus, type I errors have an important effect on resolution when the typing system has a very high α value, such as might be achieved by whole genome sequencing or by very large hybridization arrays. This is the cost of including large numbers of slowly mutating loci in the typing system.

Next we must examine the effect of missed detections on the confidence with which a typing system can declare two isolates have identical genomes.

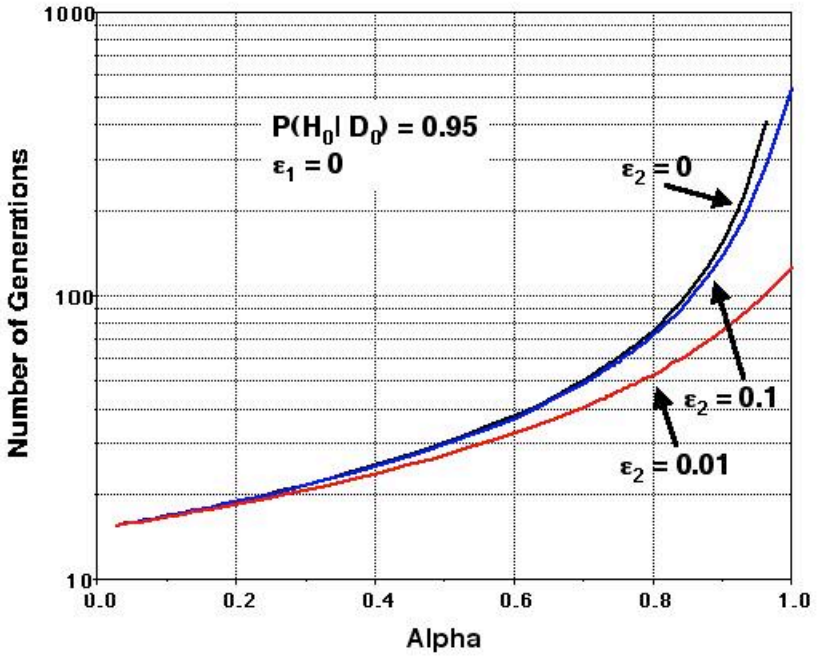

Figure 11. Maximum number of generations separating two isolates for which a reported match implies with high confidence that there are no mutational differences between them.

The probability that there are no mutational differences between two isolates, given that the typing system reports none (D_0) is given by:

$$P(H_0 \mid D_0) = P(D_0 \mid H_0)P(H_0)/P(D_0) \tag{31}$$

Where

$$P(D_0) = P(D_0 \mid H_0)P(H_0) + P(D_0 \mid H_1)P(H_1) \tag{32}$$

Figure 11 shows the maximum number of generations that can separate two isolates beyond which the probability $P(H_0 \mid D_0)$ is less than 0.95. In other words, for isolates separated by more than the indicated number of generations, the probability $P(H_1 \mid D_0)$ that there are mutational differences between them even though the typing system reports none is greater than 5%.

In the low α region, the minimum number of generations is independent of ε_2, and is given by:

$$N_g = -\ln(P_{H0D0})/G_g \quad (33)$$

For $P_{H0D0} = 0.95$ and $G_g = 0.0034$ per generation, $N_g \approx 15$, as indicated in Figure 11. More stringent values of $P(H_0|D_0)$ will drive N_g to lower values. For α values close to 1, N_g is given by:

$$N_g \approx \ln[P_{H0D0}\varepsilon_2/(P_{H1D0} + P_{H0D0}\varepsilon_2)]/\Gamma_g \quad (34)$$

The effect of ε_2 is to constrain the value that N_g can have in the $\alpha \approx 1$ limit.

This analysis of errors implies that the development of practical high-resolution strain typing systems will always involve a balance between type I and type II errors. Since type I errors are those where the typing system falsely reports genetic differences between isolates which are, in fact, identical, the consequence is the false exclusion of a possible "suspect lineage." Since type II errors are those in which the typing system fails to report genetic differences between isolates that are, in fact, different, the consequence is an unnecessarily large set of isolates that require whole genome sequencing to resolve. It is of some interest to consider a simple formal analysis of choosing the optimal balance, based on the model we have developed in the previous sections. The analysis is based on the standard Bayesian cost function approach[88].

Given the conditional error probabilities $P(D_1|H_0)$ and $P(D_0|H_1)$ we can construct a simple cost function based on the individual costs of the two types of error, and find the α value of the optimal typing system that minimizes the total cost. That is, we define:

$$C = C_{fe}P(D_1|H_0) + C_{md}P(D_0|H_1) \quad (35)$$

where C_{fe} is the cost of falsely declaring two isolates to be different (false exclusion) and C_{md} is the cost of a missed detection of genetic differences. (We have implicitly assumed that the cost of correct typing is 0 in this expression.) It is difficult to calculate, of course, the exact costs of these errors in monetary terms. However, it is plausible that C_{md} is approximately the marginal cost (perhaps both time and money) of an additional whole genome finished sequence, plus the cost of pursuing other aspects of the investigation on a source that could have been excluded earlier. C_{fe} is the cost that would accrue because of time lost in an investigation due to neglect of a suspect source that was falsely excluded on the basis of the faulty strain typing assay, as well as the cost of

having to re-investigate that source later. In any case, it will become apparent that the conclusions of this analysis do not depend strongly on the exact values of these parameters.

To cover a plausible range of cost scenarios we have calculated C as a function of a for three cases: $C_{fe} = 10 \times C_{md}$ (cheap sequencing); $C_{fe} = C_{md} = 1$ (similar costs) and $C_{md} = 10 \times C_{fe}$ (expensive sequencing.) These cases are illustrated in Figures 12 – 14. Here, the cost curves were calculated for various values of N_g and fixed values of ε_1 and ε_2.

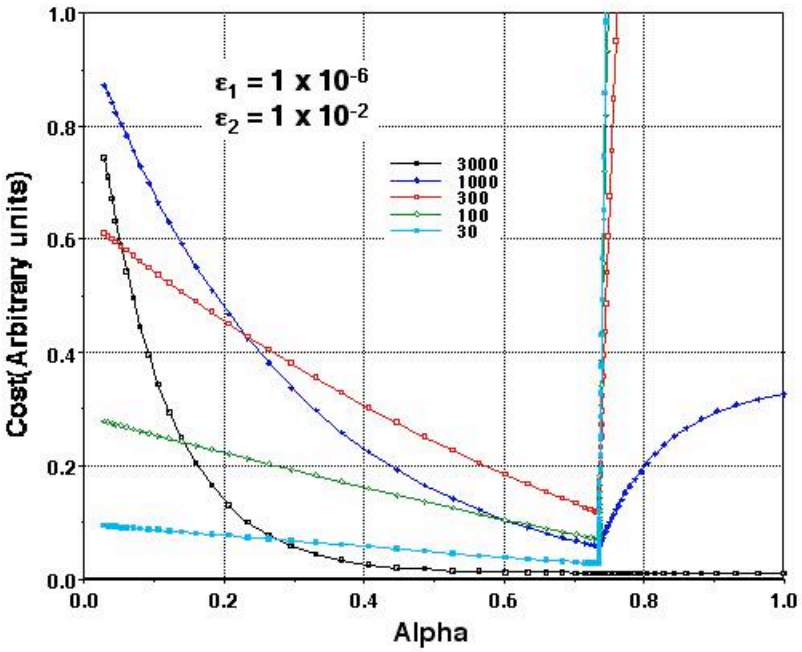

Figure 12. Cost curves for the "cheap sequencing" case.

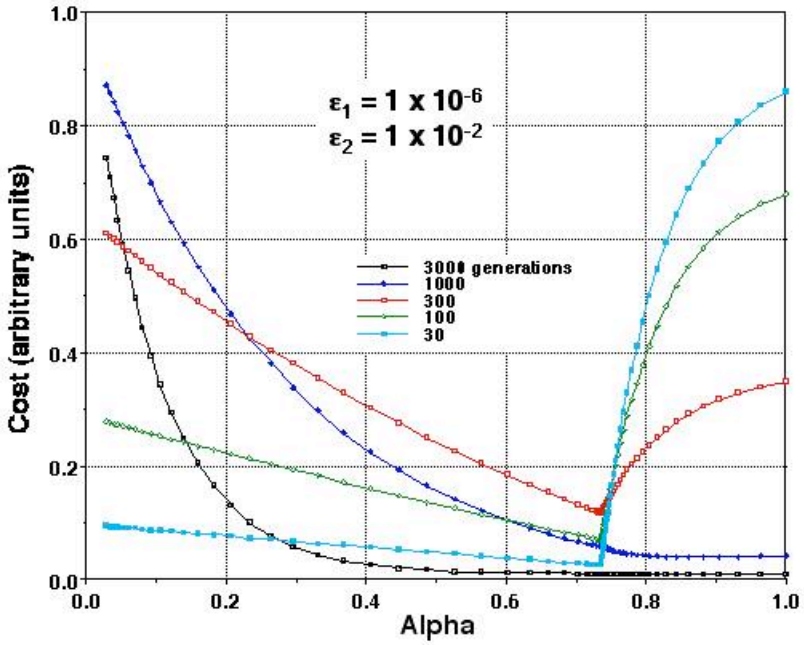

Figure 13. Cost curves for the case of equal costs for false detection and missed detection.

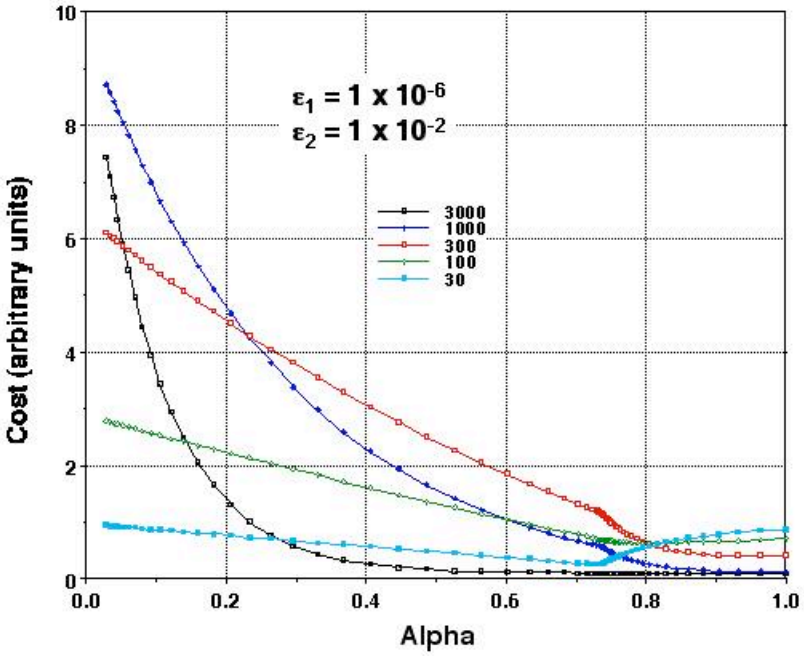

Figure 14. Cost curves for the "expensive sequencing" case.

It is clear from these figures that, for a large fraction of cases where N_g is less than 3000 generations, there is a minimum in cost at the break point around α = 0.7, where the fastest mutations have been incorporated into the typing system,

but the accumulation of type I error probability that takes place as the large number of SNS-like mutations are incorporated has not yet begun. For $N_g \geq$ 3000 generations, the likelihood of mutations becomes so significant that the correct detection of at least one real mutational difference becomes more likely than making a type I error, so the curves transition to very flat minima with low cost even in the a> 0.7 region.

In order to achieve minima closer to $\alpha \approx 1$ it is necessary to have ε_1 values of 10^{-7} or smaller. This places a rough bound on the maximum probability of false detection type errors that can be tolerated in high density gene-chip typing systems before their advantages for high resolution strain typing can be achieved. In this regard, it is of interest that the SNP error rate for whole genome sequencing itself is typically of order 10^{-5}, so that in the case of the Ames anthrax isolates it was necessary to use separate PCR based SNP assays to verify the identity of the dozens of apparent SNPs that resulted from comparative sequencing[61]. While the error rate associated with high density gene chip arrays is not widely reported, observations in references 89 and 90 suggest that failure to hybridize (which amounts to a false detection of mutational difference) often occurs at rates of order 10^{-2}. At these rates, even the notion of using a combination of re-sequencing array data and confirmatory SNP PCR assays may not be feasible, since a typical array re-sequencing of a bacterial genome is likely to produce thousands of apparent SNPs that would need to be assayed individually for accuracy. However, more recent work with tiling arrays that are designed to provide more uniform Tm values across array elements, and specialized analysis software suggest that rates around 10^{-6} are possible[87].

7. Towards the systematic development of optimized strain typing systems for microbial forensics

The above analysis provides a conceptual basis upon which to formulate a strategy for future development of higher resolution forensic strain typing techniques. However, the technological question of how to create an optimal high-resolution typing system is of practical relevance only if it is clear that there is a need for such a system. It may be the case that one or more existing strain typing systems are sufficient to reduce clusters of unresolved strains among U.S. holdings to sizes that are easily manageable in a bio-terror investigation, thus satisfying the requirement implied by current select agent legislation. This point can only be rigorously decided by empirical means. Thus, an important study towards establishing requirements might consist of choosing the highest resolution strain typing system available for each select agent pathogen from the current commonly used systems, and using it to type isolates from as many of the CDC registered laboratories as possible. The number and size of unresolved clusters of strains, combined with estimates of the time and cost of resolving

them by whole genome sequencing will determine the need for advanced typing R&D.

Even in the absence of this data, however, one can use a less rigorous argument to illustrate the potential need for higher resolution systems. The performance of typing systems is often characterized by a resolution index, i.e. a parameter that summarizes the ability of the typing assay to differentiate among a set of test strains. One example of such an index was defined by Hunter [91]as:

$$S = 1 - \{(\Sigma n_j(n_j-1))/N(N-1)\} \quad (36)$$

where N is the total number of isolates in the collection of test strains, and n_j is the number of isolates having the jth fingerprint, and the sum extends over the K distinct fingerprints generated by the typing system when applied to the N isolates. A typing system that would be currently regarded as high-resolution typically has $0.95 \leq S < 0.99$. Equation (36) can be used to calculate how the expected average size of an unresolved cluster of strains depends on the value of S and the total number of strains being typed. The result of such a calculation is shown in Figure 15.

From Figure (15) it is clear that when a strain collection with 1000 or more isolates is typed by a system with $S < 0.99$, clusters of average size greater than 10 will be statistically expected. In actual practice, of course, there will be some distribution of cluster sizes depending on the origin and clonal relationships among the isolates. Large clusters might be expected for "interesting strains" that have been passed to many laboratories, or collected from major outbreaks. In any case, the prospect of quickly resolving clusters of 10 or more isolates by whole genome sequencing is likely to represent a less-than-desirable challenge during a bio-terror investigation, thus motivating a need for higher resolution typing.

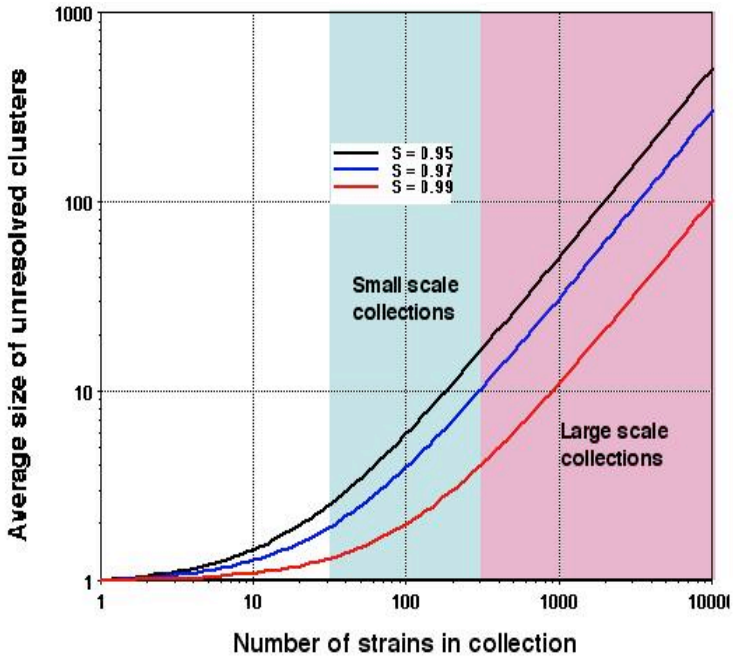

Figure 15. Average size of unresolved clusters of isolates as a function of collection size, for various values of the resolution index S. Small scale collections are defined as those containing fewer than 300 isolates, while large scale collections may have as many as 10,000 isolates.

If it is assumed that higher resolution typing methods are desirable, the next issue is how to systematically construct such methods. The model presented in section 6 suggests, for example, that a typing system that examined the first hundred or so fastest mutational loci might be nearly optimal, in the absence of a low error rate gene-chip technology. Thus, one obvious direction might be to develop more extensive sets of multi-locus PCR assays that examine a larger set of quickly mutating loci. (An additional advantage of the PCR approach is that such assays can work on trace samples, a feature often required in forensic applications.) While conventional PCR based typing approaches typically concentrate on a single type of mutation, e.g. VNTRs alone or IS elements alone, it is likely that "mixed" assays that combine the fastest mutating loci of all types would be superior. As long as the mutations are independent, combining different assays should produce additive improvements to α. Since routine assays with over 40 VNTR loci and nearly 30 IS loci have already been demonstrated, it does not seem out of the question that practical typing systems using as many as 100 different loci could be developed.

In spite of its apparent simplicity, this approach has a number of limitations as a prospective route to a "universal" strain typing system. One obvious challenge for this approach is how to identify reliably an exhaustive list of the *best* (i.e. fastest) such typing loci for a given pathogen. The limited literature on

mutational spectra provides some empirical guidance to this search, as outlined in section 3 above. Current practice involves performing bio-informatic searches through available genomic data to identify possible loci based on such prior empirical observations. This process usually results in a very large number of potential candidates. Some subset of these are chosen for further evaluation, based on a variety of (sometimes subjective) criteria. The assays that are designed for the loci and primer sites thus chosen then undergo extensive laboratory screening to verify that they work as expected. Thus, the rate at which such assays could be developed for all the select agent pathogens is ultimately limited by the rate of whole genome sequencing, and the rate at which such sequences can be annotated and analyzed for the presence of hypervariable loci. Without order-of-magnitude improvements in both sequencing rate and concomitant improvements in our understanding of sequence variation, it is likely that such an approach would require many years of intensive R&D to generate high-resolution assays for all select agents.

A second consideration is that the number of loci of a given type (VNTR, IS element, etc.) will almost certainly vary from bacterial species to species. Even within a species, the number and type of hyper-variable loci can vary. For example, some strains of M. *tuberculosis* are better typed by "spoligotyping" than by IS6110 element analysis, because they have relatively few IS6110 elements[23]. Therefore, while this general approach to typing assays is "universal," each organism would, in fact, require its own unique typing assay to achieve optimum strain discrimination.

In contrast to PCR based multi-locus assays, restriction fragment analysis *in principle* could examine a very large number of loci and does not require whole genome reference sequence information *per se*. Such techniques are truly universal, with the small caveat that slightly different restriction enzyme sets might be required to optimize the resolution for a given species. (Bio-informatics studies and simulations should prove useful in evaluating this issue.) While fragment size or mobility analysis does not reveal the exact nature of the mutations that differentiate two strains, it is generally not essential to do so in the forensic context. As was noted in section 4, these techniques are limited in practice by the number of fragments that can be resolved and characterized. Thus, a major improvement in the number of fragments and the size resolution would be necessary to make this approach attractive.

Finding the best way to evaluate and screen new typing techniques is as important an issue as how to develop them. Currently, there is no valid a priori method to rank potential mutational loci with respect to mutation rate. Instead, the resolving power of an assay is judged by its ability to distinguish among some particular set of isolates that is available to the assay developers. This often has the effect of making the allelic diversity at a locus the criterion for ranking its contribution to resolving power. In some cases there will be a correlation between diversity and mutation rate. However, the general pre-disposition to

substitute diversity for high mutation rate should be regarded with caution. Diversity alone is simply a measure of the environmental "permissiveness" for the range of alleles possible at a given locus. The range of alleles that can be tolerated in a locus, and the rate by which a single clone will diversify to exhibit all possible such alleles are, in principle, completely independent quantities. By analogy with the theory of random walks, the time needed for a locus to evolve to its maximum diversity is given by:

$$T \approx n/\gamma \tag{37}$$

where n is the number of alleles that are permitted by fitness, and γ is the mutation rate at the locus. If the set of isolates used to screen for diversity diverged over times that are longer than T, the diversity present at a locus will provide no information about γ. Conversely, a locus that can tolerate only a small range of alleles may evolve very quickly to only limited diversity. Because high mutation rate, low diversity loci are subject to homoplasy, they are likely to be ranked low for discriminating among isolates in a typical diversity panel.

The result of substituting diversity for mutation rate and using an unconstrained (or simply opportunistic) diversity panel to rank loci is that there is no guarantee that the chosen set of loci is truly optimal for high resolution forensic microbial genotyping. *It would seem prudent, then, to construct an assay screening process that systematically includes sets of isolates that are known to be separated by a relatively small number of generations, such as those generated naturally in serial passage experiments.* This procedure would provide a needed check for high mutation rate, low diversity loci that might be neglected otherwise.

8. Concluding remarks

While this paper has focused on resolution, we note that there are other figures of merit that must be considered in any down selection process for forensic strain typing. Among the practical characteristics that must be established for a new system are reproducibility, ease of application, and cost. The proliferation of new typing techniques is not necessarily a substitute for a more deliberate and systematic approach towards universal typing methods. It is notable that the need for a more uniform approach to strain typing has also been pointed out within the epidemiological community[12]. While hybridization arrays are a rapidly emerging technology with high potential in this regard, their demonstrated utility has thus far been restricted to the discovery of only a limited set of mutational markers. Moreover, there is a concurrent rapid evolution of direct sequencing methods as well, for example the recent introduction of massively multiplexed pyrosequencing[92]. It is important to point out, however, that a rigorous and systematic determiniation of fundamental error rates is still lacking for both these technologies.

9. References

1. Cummings, C.A. and D.A. Relman, *Genomics and microbiology. Microbial forensics--"cross-examining pathogens"*. Science, 2002. **296**(5575): p. 1976-9.

2. Larkin, M., *Microbial forensics aims to link pathogen, crime, and perpetrator.* Lancet Infect Dis, 2003. **3**(4): p. 180.

3. Popovic, T. and M. Glass, *Laboratory aspects of bioterrorism-related anthrax--from identification to molecular subtyping to microbial forensics.* Croat Med J, 2003. **44**(3): p. 336-41.

4. Budowle, B., et al., *Public health. Building microbial forensics as a response to bioterrorism.* Science, 2003. **301**(5641): p. 1852-3.

5. Enserink, M. and D. Ferber, *Microbial forensics. Report spells out how to fight biocrimes.* Science, 2003. **299**(5610): p. 1164-5.

6. Murch, R.S., *Microbial forensics: building a national capacity to investigate bioterrorism.* Biosecur Bioterror, 2003. **1**(2): p. 117-22.

7. Keim, P., *Microbial Forensics: A Scientific Assessment*. 2003, American Academy of Microbiology.

8. Sayers, A.A., *Microbial Forensics*. 2004, ActionBioscience.org.

9. Maslow, J. and M.E. Mulligan, *Epidemiologic typing systems.* Infect Control Hosp Epidemiol, 1996. **17**(9): p. 595-604.

10. Tenover, F.C., R.D. Arbeit, and R.V. Goering, *How to select and interpret molecular strain typing methods for epidemiological studies of bacterial infections: a review for healthcare epidemiologists. Molecular Typing Working Group of the Society for Healthcare Epidemiology of America.* Infect Control Hosp Epidemiol, 1997. **18**(6): p. 426-39.

11. Struelens, M.J., *Molecular epidemiologic typing systems of bacterial pathogens: current issues and perspectives.* Mem Inst Oswaldo Cruz, 1998. **93**(5): p. 581-5.

12. van Belkum, A., et al., *Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology.* Clin Microbiol Rev, 2001. **14**(3): p. 547-60.

13. Foxman, B. and L. Riley, *Molecular epidemiology: focus on infection.* Am J

Epidemiol, 2001. **153**(12): p. 1135-41.

14. Zaidi, N., K. Konstantinou, and M. Zervos, *The role of molecular biology and nucleic Acid technology in the study of human infection and epidemiology.* Arch Pathol Lab Med, 2003. **127**(9): p. 1098-105.

15. Vaneechoutte, M., *DNA fingerprinting techniques for microorganisms. A proposal for classification and nomenclature.* Mol Biotechnol, 1996. **6**(2): p. 115-42.

16. Olive, D.M. and P. Bean, *Principles and applications of methods for DNA-based typing of microbial organisms.* J Clin Microbiol, 1999. **37**(6): p. 1661-9.

17. Soll, D.R., *The ins and outs of DNA fingerprinting the infectious fungi.* Clin Microbiol Rev, 2000. **13**(2): p. 332-70.

18. Gurtler, V. and B.C. Mayall, *Genomic approaches to typing, taxonomy and evolution of bacterial isolates.* Int J Syst Evol Microbiol, 2001. **51**(Pt 1): p. 3-16.

19. Kristensen, V.N., et al., *High-throughput methods for detection of genetic variation.* Biotechniques, 2001. **30**(2): p. 318-22, 324, 326 passim.

20. *Public Health Security and Bioterrorism Preparedness and Response Act of 2002*. 2002. p. 637-647.

21. *H.R. Conference Report 107-481*. 2002.

22. Kremer, K., et al., *Comparison of methods based on different molecular epidemiological markers for typing of Mycobacterium tuberculosis complex strains: interlaboratory study of discriminatory power and reproducibility.* J Clin Microbiol, 1999. **37**(8): p. 2607-18.

23. van Soolingen, D., et al., *Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis.* J Clin Microbiol, 1993. **31**(8): p. 1987-95.

24. Schouls, L.M., et al., *Comparative genotyping of Campylobacter jejuni by amplified fragment length polymorphism, multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination.* J Clin Microbiol, 2003. **41**(1): p. 15-26.

25. D'Agata, E.M., et al., *Comparison of pulsed-field gel electrophoresis and amplified fragment-length polymorphism for epidemiological investigations of common nosocomial pathogens.* Infect Control Hosp Epidemiol, 2001. **22**(9): p. 550-4.

26. Huang, X.Z., et al., *Genotyping of a homogeneous group of Yersinia pestis strains isolated in the United States.* J Clin Microbiol, 2002. **40**(4): p. 1164-73.

27. Kotetishvili, M., et al., *Multilocus sequence typing has better discriminatory ability for typing Vibrio cholerae than does pulsed-field gel electrophoresis and provides a measure of phylogenetic relatedness.* J Clin Microbiol, 2003. **41**(5): p. 2191-6.

28. Hahm, B.K., et al., *Subtyping of foodborne and environmental isolates of Escherichia coli by multiplex-PCR, rep-PCR, PFGE, ribotyping and AFLP.* J Microbiol Methods, 2003. **53**(3): p. 387-99.

29. Soldati, L. and J.C. Piffaretti, *Molecular typing of Shigella strains using pulsed field gel electrophoresis and genome hybridization with insertion sequences.* Res Microbiol, 1991. **142**(5): p. 489-98.

30. Garcia Del Blanco, N., et al., *Genotyping of Francisella tularensis strains by pulsed-field gel electrophoresis, amplified fragment length polymorphism fingerprinting, and 16S rRNA gene sequencing.* J Clin Microbiol, 2002. **40**(8): p. 2964-72.

31. de la Puente-Redondo, V.A., et al., *Comparison of different PCR approaches for typing of Francisella tularensis strains.* J Clin Microbiol, 2000. **38**(3): p. 1016-22.

32. Babbette L. Marrone, R.K.a.P.K., *BDAP:User access to CBNP-Developed Bioforensics Technologies*, in *Chemical and Biological National Security Program FY02 Technical Appendix*. 2002, U.S. Department of Energy National Nuclear Security Administration. p. 195-199.

33. Orskov, F. and I. Orskov, *From the national institutes of health. Summary of a workshop on the clone concept in the epidemiology, taxonomy, and evolution of the enterobacteriaceae and other bacteria.* J Infect Dis, 1983. **148**(2): p. 346-57.

34. Research, C.f.C., *Anthrax In America: A Chronology and Analysis of the Fall 2001 Attacks*. 2002, Center for Counterproliferation Research, National Defense University.

35. *Imported plague--New York City, 2002.* MMWR Morb Mortal Wkly Rep, 2003. **52**(31): p. 725-8.

36. Budowle, B., *Defining a New Forensic Discipline: Microbial Forensics*. 2003, Laboratory Division of the Federal Bureau of Investigation.

37. *Possession, Use, and Transfer of Select Agents and Toxins:Interim Final Rule 42 CFR Part 73*, in *Federal Register*. 2002. p. 240:67:76886-905.

38. Warrick, S.F.a.J., *Army Sent anthrax strain to only 5 labs*, in *Washington Post*. November 30, 2001: Washington, DC.

39. Metzker, M.L., et al., *Molecular evidence of HIV-1 transmission in a criminal case.* Proc Natl Acad Sci U S A, 2002. **99**(22): p. 14292-7.

40. Halliday, J.A. and B.W. Glickman, *Mechanisms of spontaneous mutation in DNA repair-proficient Escherichia coli.* Mutat Res, 1991. **250**(1-2): p. 55-71.

41. Schaaper, R.M. and R.L. Dunn, *Spontaneous mutation in the Escherichia coli lacI gene.* Genetics, 1991. **129**(2): p. 317-26.

42. van Belkum, A., et al., *Short-sequence DNA repeats in prokaryotic genomes.* Microbiol Mol Biol Rev, 1998. **62**(2): p. 275-93.

43. Le Fleche, P., et al., *A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis.* BMC Microbiol, 2001. **1**(1): p. 2.

44. Moxon, E.R., R.E. Lenski, and P.B. Rainey, *Adaptive evolution of highly mutable loci in pathogenic bacteria.* Perspect Biol Med, 1998. **42**(1): p. 154-5.

45. Moxon, E.R., et al., *Adaptive evolution of highly mutable loci in pathogenic bacteria.* Curr Biol, 1994. **4**(1): p. 24-33.

46. Bayliss, C.D., D. Field, and E.R. Moxon, *The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis.* J Clin Invest, 2001. **107**(6): p. 657-62.

47. Metzgar, D. and C. Wills, *Evolutionary changes in mutation rates and spectra and their influence on the adaptation of pathogens.* Microbes Infect, 2000. **2**(12): p. 1513-22.

48. Mahillon, J. and M. Chandler, *Insertion sequences.* Microbiol Mol Biol Rev, 1998. **62**(3): p. 725-74.

49. Mahillon, J., C. Leonard, and M. Chandler, *IS elements as constituents of bacterial genomes.* Res Microbiol, 1999. **150**(9-10): p. 675-87.

50. Bennett, P.M., *Genome plasticity: insertion sequence elements, transposons and integrons, and DNA rearrangement.* Methods Mol Biol, 2004. **266**: p. 71-114.

51. Achaz, G., et al., *Origin and fate of repeats in bacteria.* Nucleic Acids Res, 2002. **30**(13): p. 2987-94.

52. Bzymek, M. and S.T. Lovett, *Instability of repetitive DNA sequences: the role*

of replication in multiple mechanisms. Proc Natl Acad Sci U S A, 2001. **98**(15): p. 8319-25.

53. Versalovic, J., T. Koeuth, and J.R. Lupski, *Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes.* Nucleic Acids Res, 1991. **19**(24): p. 6823-31.

54. Groenen, P.M., et al., *Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method.* Mol Microbiol, 1993. **10**(5): p. 1057-65.

55. Goyal, M., et al., *Differentiation of Mycobacterium tuberculosis isolates by spoligotyping and IS6110 restriction fragment length polymorphism.* J Clin Microbiol, 1997. **35**(3): p. 647-51.

56. Rocha, E.P., *An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction.* Genome Res, 2003. **13**(6A): p. 1123-32.

57. Hare, J.M. and K.A. McDonough, *High-frequency RecA-dependent and -independent mechanisms of Congo red binding mutations in Yersinia pestis.* J Bacteriol, 1999. **181**(16): p. 4896-904.

58. Henderson, I.R., P. Owen, and J.P. Nataro, *Molecular switches--the ON and OFF of bacterial phase variation.* Mol Microbiol, 1999. **33**(5): p. 919-32.

59. Sohanpal, B.K., et al., *Orientational control of fimE expression in Escherichia coli.* Mol Microbiol, 2001. **42**(2): p. 483-94.

60. Grogan, D.W., G.T. Carver, and J.W. Drake, *Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon Sulfolobus acidocaldarius.* Proc Natl Acad Sci U S A, 2001. **98**(14): p. 7928-33.

61. Read, T.D., et al., *Comparative genome sequencing for discovery of novel polymorphisms in Bacillus anthracis.* Science, 2002. **296**(5575): p. 2028-33.

62. Drake J.W. *The Distribution of Rates of Spontaneous Mutation over Viruses, Prokaryotes, and Eukaryotes*, Annals of the New York Academy of Sciences (1999), **870**, pp. 100 – 107; Drake J.W. et. al. *Rates of Spontaneous Mutation*, Genetics (1998), **148**, pp. 1667-1686.

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

64. Girard J.M, et. al. *Differential plague-transmission dynamics determine*

Yersinia pestis population genetic structure on local, regional, and global scales, Proc. Nat. Acad. Sci. 2004, **101**, pp.8408-8413.

65. Papadopoulos, D. et. al. *Genomic evolution during a 10,000-generation experiment with bacteria*, Proc. Natl. Acad. Sci. 1999, 96, pp.3807-3812.

66. Cooper, V.S. *Mechanisms Causing Rapid and Parallel Losses of Ribose Catabolism in Evolving Populations of Escherichia coli B*, J. Bacteriol. 2001, 183, pp. 2834-2841.

67. Tanaka, M.M. *Optimal estimation of transposition rates of insertion sequences for molecular epidemiology*, Statist. Med. 2001, **20**, pp.2409-2420.

68. Martusewitch, E. *High Spontaneous Mutation Rate in the Hyperthermophilic Archaeon Sulfolobus solfataricus Is Mediated byTransposable Elements*, J. Bacteriol. 2000, **182,** pp. 2574-2581.

69. Sinden, R.R. *On the Deletion of Inverted Repeated DNA in Escherichia coli: Effects of length, Stability, and Cruciform Formation in Vivo*, Genetics 1991, **129**, pp. 991-1005.

70. Sohanpal, B.K. *Orientational control of fimE expression in Escherichia coli*, Mol. Microb. 2001, 42, pp.483-494.

71. Drake J.W. *The Distribution of Rates of Spontaneous Mutation over Viruses, Prokaryotes, and Eukaryotes*, Annals of the New York Academy of Sciences (1999), **870**, pp. 100 – 107

72. Drake, J.W. and Holland, J.J., *Mutation rates among RNA viruses*, Proc. Natl. Acad. Sci. 1999, **96**, pp. 13910-13913.

73. Stumpf M.P.H. and Pybus, O.G., *Genetic diversity and models of viral evolution for the hepatitis C virus*, FEMS Microbiology Letters, 2002, **214**, pp.143-152.

74. Haydon, D. T. *The generation and persistence of genetic variation in foot-and-mouth disease virus*, Preventive Veterinary Medicine 2001, **51**, pp. 111-124.

75. Drake J.W. *A constant rate of spontaneous mutation in DNA based microbes*, Proceedings of the National Academy of Science (1991) **88**, pp. 7160-7164

76. Bridges, B.A. *DNA turnover and mutation in resting cells*, BioEssays 1997, **19**, pp.347-352.

77. Naas, T. et. al. *Dynamics of IS-related Genetic Rearrangements in Resting Escherichia coli K-12*, Mol. Biol. Evol. 1995, **12**, pp. 198-207.

78. Hamming, R.W., *The Art of Probability for Scientists and Engineers*, (Addison-Wesley, Inc., New York, 1991)

79. Maiden, M.C.J., *Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms*, Proc. Natl. Acad. Sci. 1998, **95**, pp.3140-3145.

80. Unwin, R. and Maiden, M.C.J., *Multi-locus sequence typing:a tool for global epidemiology*, TRENDS in Microbiology 2003, **11**, pp. 479-487.

81. Vos, P. et. al. *AFLP: a new technique for DNA fingerprinting*, Nucleic Acids Research, 1995, **23**, pp.4407-4414.

82. Savelkoul, P.H.M. et. al. *Amplified-Fragment Length Analysis:the State of an Art*, J. Clin. Microbiol. 1999, **37**, pp.3083-3091.

83. Mockler, T.C. and Ecker, J.R., *Applications of DNA tiling arrays for whole-genome analysis*, Genomics 2005, **85**, pp. 1-15.

84. Gunderson, K.L. et. al. *A genome-wide scalable SNP genotyping assay using microarray technology*, Nature Genetics 2005, **37**, pp. 549-554.

85. Karaman, M.W. et. al. *Comparisons of substitution, insertion, and deletion probes for resequencing and mutational analysis using oligonucleotide microarrays*, Nucleic Acids Research 2005, **33**, online publication.

86. Cherkosova, E. et. al. *Microarray analysis of evolution of RNA viruses: Evidence of circulation of virulent highly divergent vaccine-derived polioviruses*, Proc. Natl. Acad. Sci. 2003, **100**, pp. 9398-9403; Wang, D. et. al., *Microarray-based detection and genotyping of viral pathogens*, Proc. Natl. Acad. Sci. 2002, **99**, pp. 15687-15692.

87 Zwick, M.E. et. al. *Microarray-based resequencing of multiple Bacillus anthracis isolates*, Genome Biology 2004, **6**, R10.

88. McDonough R.N. and Whalen, A.D. *Detection of Signals in Noise* 2nd Ed. (Academic Press, New York, 1995.)

89. Kato-Maeda, M. et. al. *Comparing Genomes within the Species Mycobacterium tuberculosis*, Genome Research 2001, **11**,pp. 547-554; Liu, S. et. al. *Analysis of the factors affecting the accuracy of detection for single base alterations by*

oligonucleotide microarray, Exp. Mol. Medicine 2005, **37**, pp. 71-77.

90. Hacia, J.G. *Resequencing and mutational analysis using oligonucleotide microarrays*, Nature Genetics Supp. 1999, **21**, pp. 42-47; Cutler, D.J. et. al. *High-Throughput Variation Detection and Genotyping Using Microarrays*, Genome Research 2001, **11**, pp.1913-1925.

91. Hunter, P.C. *Reproducibility and Indices of Discriminatory Power of Microbial Typing Methods*, J. Clin. Microbiol. 1990, **28**, pp. 1903-1905.

92. Margulies M., et. al. *Genome sequencing in microfabricated high density picoliter reactors*, Nature 2005, electronic publication ahead of print.

Appendix 1: A model for the rate constants γ_j.

This model attempts to capture several observations about mutation rates:

1) There must be some mutation whose rate value is the highest among all the possible mutations on all G typable loci within the genome. This may be, for example, insertion/deletion at a particular monomeric repeat or VNTR. We will call this rate γ_0 and assume that $\gamma_0 \approx 10^{-4}$ per generation.

2) If we order the mutational rates of all possible mutations at all possible loci within a genome, we find that there is a relatively rapid fall off (i.e. within some small fraction of the total number of loci, G) to a baseline mutation rate γ_{base}, somewhere around 3×10^{-10} per generation. This appears to be supported by (unpublished) experimental observations that the mutation rates of particular mutations, say VNTRs.

3) The sum of all the rate constants, which is the overall genomic mutation rate Γ_g, is some value close (within an order of magnitude) to Drake's "universal" value of 0.0034 per generation.

The third condition determines a constrained trade-off between the highest value that γ_0 can attain and the number of loci over which the mutation rates decay to the baseline value. That is, γ_0 can be larger than 10^{-4}, but this means that there will be fewer loci that have mutation rates that are not significantly lower than γ_0.

With these assumptions, a relatively general representation of the mutation rate for locus j is:

$$\gamma_j = Ae^{-(j/Nfast)} + \gamma_{base} \qquad \text{(A1.1)}$$

The values of γ_0 and γ_{base} determine $A = \gamma_0 - \gamma_{base}$; N_{fast} is roughly the number of loci with very fast rates, i.e. rates close to γ_0. It can be determined from the relation

$$\Gamma_g = G\gamma_{base} + A(1 - e^{-G/Nfast})/(1 - e^{-1/Nfast}) \qquad \text{(A1.2)}$$

Which is derived by summing (A1.1) over all G loci.

A typing system that examines m loci has a figure of merit given by $\alpha = \Gamma_m/\Gamma_g$, where:

$$\Gamma_m = m\gamma_{base} + A(1 - e^{-m/Nfast})/(1 - e^{-1/Nfast}) \qquad (A1.3)$$

With the values of γ_0, γ_{base}, and Γ_g given above, and assuming $G = 3 \times 10^6$, we derive the dependence of α on m shown in figure A1.1.

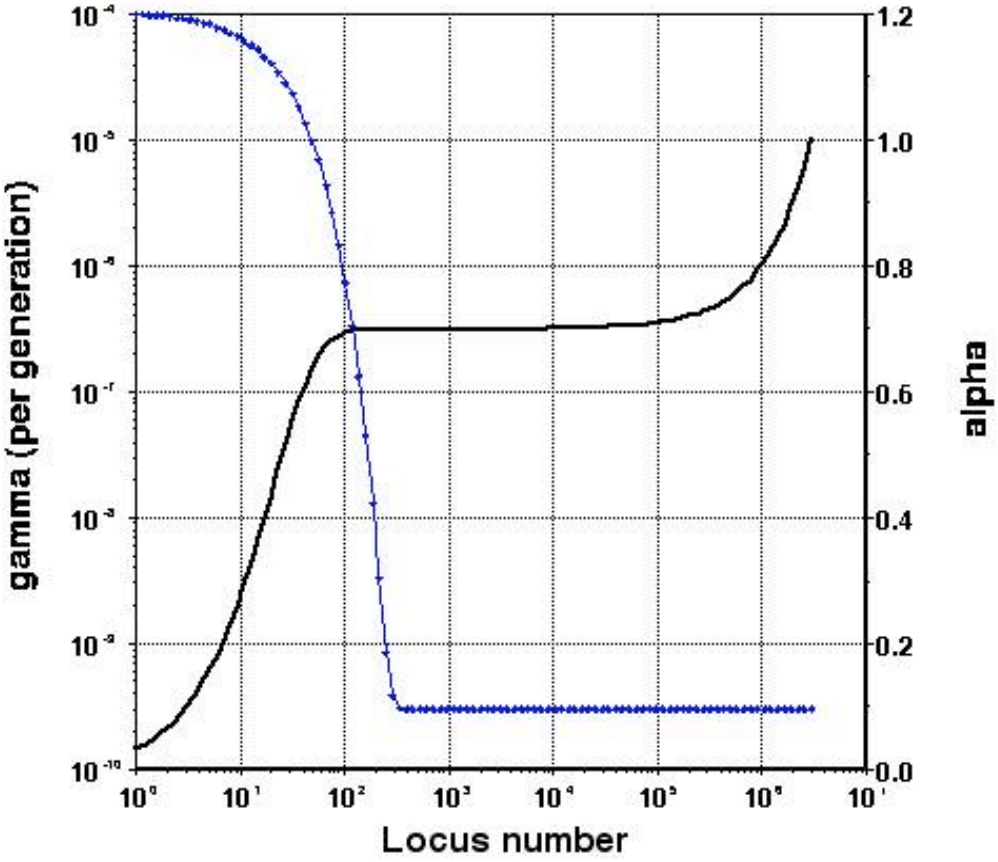

Figure A1.1. Rate constants as a function of locus number and the α value resulting from integration of the rates up to locus number m. Note the logarithmic scales. A total of 3 million loci were assumed, and the highest rate constant (m=1) was 1×10^{-4}.

Appendix 2. A figure-of-merit for the phylogenetic depth of a typing system

Two typing systems for a microbe could have the same α value, but not be equivalent in utility. In particular, a system that uses just one very quickly changing marker clearly conveys much less information about the relatedness of two isolates than a typing system that consists of many more slowly changing markers. In order to derive a figure-of-merit that reflects the amount of information that is associated with the marker set, it is convenient to introduce the notion of "surprise" (see, for example, R. W. Hamming *The Art of Probability for Scientists and Engineers*, 1991.) The less likely an event is to happen, the more surprise is associated with its observation.

Consider a typing system that interrogates m_s loci. By analogy with equations (2) and (10), the conditional probability that a mutation will be observed at a particular interrogated locus j *given* that at least one mutation has occurred somewhere among the m_s loci is $(1 - e^{-\gamma_j N})/(1 - e^{-\Gamma_s N})$. The "surprise" S in observing that a mutation has occurred at locus j is given by:

$$S_j = -\ln\{(1 - e^{-\gamma_j N})/(1 - e^{-\Gamma_s N})\} \qquad \text{(A2.1)}$$

We can define an average surprise for the typing system by weighting S_j by its associated probability, and summing over the m_s loci that are interrogated by the system:

$$\langle S \rangle = -\Sigma[(1 - e^{-\gamma_j N})/(1 - e^{-\Gamma_s N})]S_j \qquad \text{(A2.2)}$$

In the limit that $N \approx 0$, this expression simplifies to :

$$S_0(m_s) = -\Sigma\,(\gamma_j/\Gamma_s)\ln(\gamma_j/\Gamma_s) \qquad \text{(A2.3)}$$

Since the normalization $\Sigma\,(\gamma_j/\Gamma_s) = 1$ holds (the sum is over m_s loci), S_0 is an entropy-like measure which is maximized by having equal γ_j values at each locus. If we compare two typing systems with the same α value, but consisting of two distinct groups of equal valued rate constants, $\{m_{s1}, \gamma_1\}$ and $\{m_{s2},\gamma_2\}$, it is easy to show that $S_0(m_s) = \ln(m_s)$ so that the system with the larger value of m_s (and hence smaller value of γ_j) will have a larger S_0 value.

A normalized "surprise" figure of merit β can be defined by:

$$\beta = S_0(m_s)/S_0(G) \qquad \text{(A2.4)}$$

where G represents the totality of loci in the genome. β thus represents the fraction of "surprise" that is captured by the typing system, just as α represents the fraction of the total genomic mutation rate that is captured. Like α, β is additive over independent sets of mutational loci.

S_0 can be interpreted as a measure of how closely the typing system approaches an ideal evolutionary clock, in which each marker changes at the same constant rate. In addition, this quantity is a measure of the phylogenetic depth of the typing system. This follows from the fact that the "surprise" at seeing a single marker j change after N generations decreases after $N > 1/\gamma_j$. Thus, a typing system that incorporates many slower markers could be said to maintain its ability to surprise over a larger number of generations separating two isolates. Finally, a system with a high β figure-of-merit would clearly be less sensitive to homoplasy error.

To estimate β values for some common bacterial typing systems, we utilize the same kinds of crude approximations used to estimate α. We use average values of mutation rates for a given class of mutations , and assign a representative number of mutational loci to each typing system. Since all the markers within a set of common mutational loci have the same value, S_0 values for each marker system are then simply given by $S_0(m_s) \approx m_s \bullet \ln(m_s)$. Table A2.1 provides the values of the parameters used in the computation of Γ and S_0,

Table A2.1 Parameters used in the calculation of Γ and β.

Mutational marker	Average mutation rate γ (per generation)	Number of markers m_s	Γ_s (per generation)	S_0
Single nucleotide substitution (SNS)	5×10^{-10}	2×10^{6}	1×10^{-3}	2.9×10^{7}
Short Indel	5×10^{-10}	1×10^{5}	5×10^{-5}	1.2×10^{6}
VNTR	3×10^{-5}	30	1×10^{-3}	1×10^{2}
SNR	3×10^{-4}	10	3×10^{-3}	2.3×10^{1}
IS	1×10^{-5}	30	3×10^{-4}	1×10^{2}
Large deletion or duplication	1×10^{-6}	30	3×10^{-5}	1×10^{2}
Inversion	1×10^{-5}	10	1×10^{-4}	2.3×10^{1}
Total	-	-	$\Gamma_g = 5.5 \times 10^{-3}$	$S_0(G) = 3 \times 10^{7}$

Note that the while the whole genome mutation rate is dominated by VNTRss and single nucleotide repeats, the whole genome value $S_0(G)$ is dominated by the contribution of the single nucleotide substitutions and short insertion/deletions.

For PCR based typing assays that examine one type of mutational locus (e.g. VNTRs), the calculation of α and β are straightforward, using equations 13 and A2.4. For MLST, we assume that 7 gene fragments totaling 3500 loci are probed for SNS and short indels. For AFLP, we assume conditions similar to those assumed in section 5, so that $m_s \approx 3 \times 10^4$ and $\alpha = 6 \times 10^{-3}$. Finally, we include an estimate for a hybridization microarray (e.g. a tiling array) that can find SNPs over large fractions of the entire genome. The results of these approximate calculations are given in Table A2.2 and illustrated graphically in Figure A2.1.

Table A2.2 Approximate α and β values for some common typing systems.

Typing system	Effective # of loci	α	β
MLVA - 30	30	0.2	3×10^{-6}
SNR - 3	3	0.2	1×10^{-7}
IS - 30	30	6×10^{-2}	3×10^{-6}
MLST - 7	3500	1×10^{-4}	1×10^{-3}
AFLP - 1	3×10^4	6×10^{-3}	1×10^{-2}
SNP microarray	1×10^3 to 2×10^6	0.2	0.97

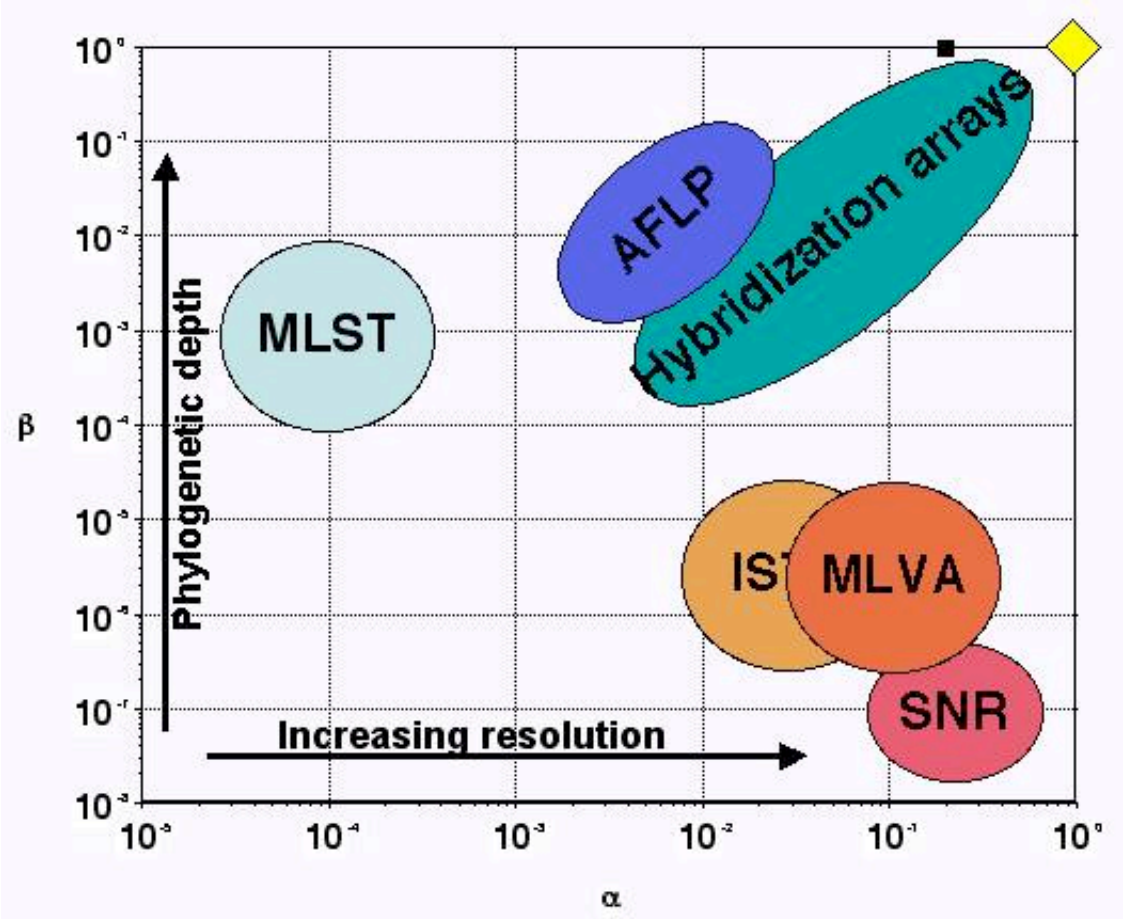

Figure A2.1 Mapping of various typing systems onto the space defined by the figures of merit a and b. The sizes of the various regions attempt to account for variances in marker number and marker mutation rate in practical systems. The yellow diamond marks the position occupied by ideal (error-free) whole genome sequencing.

Appendix 3. Probability of missed detection

The conditional probability that no mutations will be reported by the typing system, given h_1, i.e. that at least one mutation has occurred among the m_s examined loci, is not simple to calculate. However, consider the case that exactly one mutation has occurred among the examined loci, (which we will denote hypothesis $h_{=1}$.) If p_j is the probability that the mutation occurred at locus j, then

$$P(D_0 \mid h_{=1}) = \Sigma p_j \varepsilon_{2j} , \qquad \text{(A3.1)}$$

where the sum is over the m_s loci examined by the typing system. It is clear that the conditional probabilities that correspond to the hypotheses of exactly 2, 3 etc. mutations will involve products of 2, 3 etc. e_{2j} values. In any reasonable typing system, we should expect that $e_{2j} << 1$, and it should be reasonable to neglect these higher order terms. Thus,

$$P(D_1 \mid h_1) \approx P(D_1 \mid h_{=1}). \quad \text{(A3.2)}$$

Finally, assuming that ε_{2j} is approximately the same for each locus j, we can replace it by the constant average value ε_2. Then, noting that, given the hypothesis $h_{=1}$, $\Sigma p_j = 1$, we can approximate

$$P(D_0 \mid h_1) \approx \varepsilon_2. \quad \text{(A3.3)}$$

This accords with the intuition that the probability of missing a mutation among the examined loci should not be a function of the number of loci, and should be much less probable when there are two or more mutations. By contrast, the probability of *falsely* reporting mutations does depend on the number of loci examined, because, in order to report D_0 the typing system must get every locus correct, thus:

$$P(D_0 \mid h_0) = \Pi(1 - \varepsilon_{1j}), \quad \text{(A3.4)}$$

Where the product is over all m_s examined loci.