Available online at www.sciencedirect.com
SCIENCE DIRECT®
Infection, Genetics and Evolution 4 (2004) 205–213
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Review

Anthrax molecular epidemiology and forensics: using the appropriate marker for different evolutionary scales

Paul Keim*, Matthew N. Van Ert, Talima Pearson, Amy J. Vogler, Lynn Y. Huynh, David M. Wagner

Keim Genetics Laboratory, Department of Biological Sciences, Northern Arizona University, Flagstaff AZ 86011-5640, USA

Received 22 October 2003; received in revised form 16 February 2004; accepted 26 February 2004

Available online 21 July 2004

Abstract

Precise identification of *Bacillus anthracis* isolates has aided forensic and epidemiological analyses of natural anthrax cases, bioterrorism acts and industrial scale accidents by state-sponsored bioweapons programs. Because there is little molecular variation among *B. anthracis* isolates, identifying and using rare variation is crucial for precise strain identification. We think that mutation is the primary diversifying force in a clonal, recently emerged pathogen, such as *B. anthracis*, since mutation rate is correlated with diversity on a per locus basis. While single nucleotide polymorphisms (SNPs) are rare, their detection is facilitated by whole genome discovery approaches. As highly stable phylogenetic markers, SNPs are useful for identifying long branches or key phylogenetic positions. Selection of single, diagnostic "Canonical SNPs" (canSNPs) for these phylogenetic positions allows for efficient and defining assays. We have taken a nested hierarchal strategy for subtyping *B. anthracis*, which is consistent with traditional diagnostics and applicable to a wide range of pathogens. Progressive hierarchical resolving assays using nucleic acids (PHRANA) uses a progression of diagnostic genomic loci that are initially highly stable but with low resolution and, ultimately, very unstable but with high resolution. This approach mitigates the need for data weighting and provides both a deeply rooted phylogenetic hypothesis and high resolution discrimination among closely related isolates.

Keywords: Anthrax; *Bacillus anthracis*; Bioterrorism; Canonical SNPs; canSNPs; Microbial forensics; MLVA; PHRANA; SNPs; SNR

1. Background and introduction

Molecular typing of *Bacillus anthracis* has provided important insights into bioterrorism or biowarfare related events. Recently, analyses performed by the Centers for Disease Control and Prevention (CDC) and the Keim Genetics Laboratory determined that the initial victim of the 2001 anthrax letter attacks had been infected with the Ames strain of anthrax (Hoffmaster et al., 2002; Read et al., 2002). As Ames is a commonly used laboratory strain and is rare in nature (Keim et al., 2000), the strain identity was a key factor in changing this case from an epidemiological to a criminal investigation. In 1979, there was an accident at an anthrax spore production facility in Sverdlosck (Yekaterinburg), USSR (Meselson et al., 1994), which resulted in the deaths of at least 64 people. Strain analyses performed on isolates from tissues of some of the dead suggested that the plume was a composite of multiple genotypes (Jackson et al., 1998; Price et al., 1999) and provided one of the few public insights into what happened in this tragedy. Finally, in 2001, retrospective molecular analyses (Keim et al., 2001) uncovered an attempted bioterrorism event in Kameido, Japan in 1993 that was carried out by members of the Aum Shinrikyo "dooms day" cult (Takahashi et al., 2004). The typing efforts determined that the anthrax spores used in the attack were from the highly attenuated Sterne strain, which explained the failure of the cult members to kill any of their neighbors.

Although the recent role of *B. anthracis* in bioterrorism events is probably a direct consequence of its development as a biological weapon by several countries (e.g., USSR, Britain, USA), it is also found naturally throughout the world. Molecular typing methods have been applied to isolates from natural outbreaks and have provided valuable insight into introduction and dispersal patterns in natural environments. For example, spatial and temporal analyses of anthrax-caused wildlife deaths were combined with strain subtyping analyses to reveal that two very distinct strains (A and B1 subtypes) are active in Kruger National Park (South

* Corresponding author. Tel.: +1-928-523-1078; fax: +1-928-523-0639.
E-mail address: paul.keim@nau.edu (P. Keim).

1567-1348/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2004.02.005

Africa), indicating separate introductions into this small area (Smith et al., 2000). The two subtypes rarely overlap but are temporally coordinated in the same outbreak years, leading to the conclusion that environmental cues coordinate emergence from the soil reservoir to cause outbreaks, rather than extensive animal to animal transmission from a single source.

Molecular subtyping of *B. anthracis* has been developed for law enforcement and national security applications, in contrast to other pathogens where public health and epidemiological applications are foremost. However, the essential components for a highly precise and robust subtyping system are the same for both purposes and may act as a paradigm for advancing the field overall. In all cases, it is essential to (1) identify diversity, (2) develop molecular typing assays, (3) populate a database, (4) establish a theoretical and probabilistic basis for the diversity, and (5) validate the system with studies of real disease outbreaks.

2. *B. anthracis* exhibits low genetic diversity

Pathogens that emerge through a population bottleneck (likely a single cell) initially form groups of genetically identical organisms or clones. Over evolutionary time, mutations will eventually generate genetic heterogeneity within these organisms and provide the basis for distinguishing among individual lineages. For this reason, when it is found that a pathogen has little or no genetic diversity it is assumed to have recently emerged. *B. anthracis* appears to fit this definition as it exhibits very low molecular diversity among widely distributed strains. For example, comparative sequencing of the *pagA* gene (Price et al., 1999) and MLST (Okinaka et al., 1999; Okinaka, unpublished data) revealed only a few substitution mutations among very diverse isolates. However, this characterization of *B. anthracis* as a recently emerged pathogen is problematic as it does not fully consider its complex life cycle.

Growth and reproduction in *B. anthracis* is unpredictable and slow in comparison to other bacterial pathogens. Outside of a suitable host, *B. anthracis* remains dormant in the spore stage, which adds a stochastic element to the population dynamics of this pathogen. Depending upon the regional ecology of the disease, years, decades or possibly even centuries may pass between infectious cycles. Evolution is greatly reduced during dormant periods and is minimal even during an infection-death-infection cycle, which typically involves only 20–40 generations (doublings). In a given area, an infection cycle might occur several times per year, but typically much less. Even at several times per year (a major anthrax outbreak), the number of annual generations for *B. anthracis* is likely much lower than *E. coli*, which is thought to undergo 300 generations per year (Guttman and Dykhuizen, 1994).

Although we are unsure of the specific causal mechanisms, there is a distinct lack of diversity within *B. anthracis* at genes and in common genetic markers, such as AFLPs (Harrell et al., 1995; Keim et al., 1997). This lack of genetic diversity has hindered efforts to discriminate among strains for phylogenetic and forensic purposes. To address this problem, we have focused much of our attention over the last few years on developing novel typing systems for *B. anthracis* by locating areas of molecular diversity within the genome. Our efforts led us to three different molecular markers: single nucleotide polymorphisms (SNPs), variable number tandem repeats (VNTRs) and single nucleotide repeats (SNRs), which are a special case of VNTRs. Across a given set of *B. anthracis* strains, these three types of markers exhibit very different diversity values, which are primarily a function of different mutation rates in the cases we have studied.

3. Genetic diversity and mutation rates

Genetic diversity, a compound measure that includes the number of allelic states as well as their frequency distribution within the population, is affected by four processes: mutation, selection, genetic drift and recombination. Over evolutionary time, mutations provide the raw material of diversity by creating additional alleles. However, this process eventually approaches a maximum value, where novel mutations are offset by recurrent mutations (Fig. 1). Selection, genetic drift and recombination can influence genetic diversity by affecting the distribution of the different allele states created by mutation. The more evenly distributed the allelic states are, the greater the diversity (Fig. 2). All four of these factors may be shaping genetic diversity in *B. anthracis*.

Selection, drift and recombination may all affect the distribution of alleles and, therefore, genetic diversity in *B. anthracis*. Selection likely eliminates certain SNPs from the population and limits maximum array size in VNTRs. Phenotypic selection on VNTRs is quite possible, though only one VNTR in *B. anthracis* has been shown to have a phenotypic effect (Sylvestre et al., 2003). Despite the paucity of evidence, selection may be important in constraining genetic diversity in *B. anthracis* at particular loci. Founders effect and, therefore, genetic drift may strongly influence regional patterns of diversity in *B. anthracis*, as a single spore can cause a local anthrax case. Indeed, these types of population bottlenecks must have been common in the evolution and dispersal of regional *B. anthracis* populations. However, averaged over the entire global population, the effects of drift are probably small. Recombination within *B. anthracis* may have had little or no effect on genetic diversity, as the only recognized point for horizontal transfer of genetic material is during the rapid disease phase in an animal, when there would be little time for mixed infections. Growth of *B. anthracis* in the environment (e.g., soil) might also allow for horizontal gene movement, although there is no evidence for this scenario.

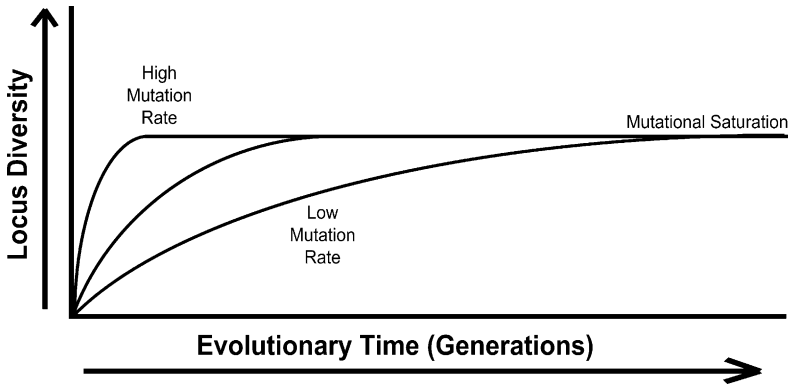

Fig. 1. Mutational saturation curve. Mutational saturation occurs at a locus or nucleotide position when additional mutations do not increase diversity [$D = 1 - \Sigma(\text{allele frequencies})^2$]. At this state, the influence of additional novel mutations is offset by recurrent mutations. Each locus will reach mutational saturation at a unique point in evolutionary time due to locus-specific mutation rates and other factors. In general, loci that evolve rapidly will reach saturation more quickly then loci that evolve more slowly.

Mutation, more specifically mutation rate, is probably the most important factor affecting diversity in *B. anthracis*. For example, diversity of individual VNTR markers in *B. anthracis* is highly correlated with the mutation rate of those markers, which suggests that mutation rates drive genetic diversity within VNTR loci (Vogler and Keim, unpublished data). Although mutation rates at an individual locus are partially intrinsic to that locus, they are also a function of more global mutational processes. These global processes and, therefore, mutation rates are different for SNPs and VNTRs.

SNPs are generated by nucleotide substitutions, probably DNA replication errors that are not subsequently repaired. These events are rare in *B. anthracis*, occurring at estimated rates of approximately 10^{-10} changes per nucleotide per generation (Vogler et al., 2002). The actual observed mutation rates of SNPs in *B. anthracis* are likely much lower than this estimate, as mutations in coding regions that significantly altered gene function would be selected against and never observed. With the exception of coding versus non-coding regions, there is little available information concerning the effects of local, locus-specific properties on SNP mutation rates.

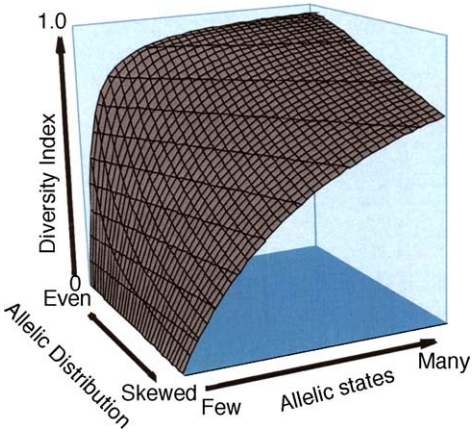

Fig. 2. Allelic distribution and number of states affect the diversity index. Diversity at a locus is a function of both the number of alleles (states) and their frequency distribution. As the number of alleles increases, the potential diversity also increases. However, maximum diversity within a population is only obtained if the individual allele states are equalized in frequency.

The global mechanism that likely generates mutations in VNTR loci is slipped strand mispairing (Levinson and Gutman, 1987; Eisen, 1999). These insertion-deletion (INDEL) mutations occur at varied rates in VNTR loci, ranging from $<10^{-5}$ to $>10^{-4}$ in *B. anthracis* and exceeding 10^{-3} in other bacteria (Vogler and Keim, unpublished data). This wide range of mutation rates across VNTR loci is probably the result of intrinsic, locus-specific properties, including length of the tandem array (Levinson and Gutman, 1987; Andreassen et al., 1996; Rosche et al., 1996; Brinkmann et al., 1998; Buard et al., 1998; Parniewski et al., 2000), existence of imperfect or interrupted repeat sequences within the array (Brinkmann et al., 1998; Buard et al., 1998; Eisen, 1999; Henderson and Petes, 1992), ability of VNTR sequences to form stable secondary structures (Freudenreich et al., 1997; Hartenstine et al., 2000), functionality of the mismatch repair system (Levinson and Gutman, 1987; Strand et al., 1993), size of the individual repeat unit (Henderson and Petes, 1992; Weber and Wong, 1993; Yang and Masker, 1996; Eckert and Yan, 2000) and strand orientation (Henderson and Petes, 1992; Rosche et al., 1996; Freudenreich et al., 1997). The wide range of mutation rates among VNTR loci produces a similarly wide range of diversity values for these markers, which makes them useful for examining genetic patterns at several evolutionary scales.

The increased level of genetic diversity in VNTRs compared to SNPs is not only a function of differences in mutation rates, but also the maximum number of possible allelic states for each type of marker. Diversity (D) is a bounded measure with values ranging from 0 to <1.0 [$D = 1 - \Sigma(\text{allele frequencies})^2$ throughout this manuscript]. As a SNP can take just one of four possible allelic states, the maximum diversity value for any SNP locus is 0.75. In

reality, however, such high diversity values for SNP loci are seldom observed within bacterial populations. As single nucleotide substitutions are rare, most SNP loci have only two allelic states within populations (maximum $D = 0.5$). In contrast, differences in VNTR allelic states are due to length polymorphisms, so a greater number of alleles are possible compared to SNPs. For example, the maximum number of allelic states observed for each of the 15 VNTR loci that we use to examine *B. anthracis* ranges from 2 to 10. Thus, there is more potential diversity at any given VNTR locus compared to any given SNP locus for two reasons: (1) VNTRs have higher mutation rates, and (2) VNTRs have a larger number of possible allelic states.

When analyzing a set of *B. anthracis* isolates for phylogenetic or forensic purposes, care should be exercised to select an appropriate type of marker for both the set of strains and the questions being asked. It is not appropriate to simply utilize the most highly diverse markers for every analysis, as high levels of genetic diversity can also confound phylogenetic patterns. Thus, to facilitate selection of a marker system, following is a brief synopsis of SNPs, VNTRs and SNRs and their respective utility for molecular typing of *B. anthracis*.

4. Canonical markers: single nucleotide polymorphisms (SNPs)

SNPs occur at very low frequencies in the *B. anthracis* genome but they can be readily discovered using intensive sampling methods. As we describe above, the genome of *B. anthracis* is relatively homogenous and point mutations that form SNPs occur at very low rates. In comparison to other genetic markers, SNPs are exceedingly rare among even distantly related *B. anthracis* isolates and, therefore, would seem to have limited subtyping capacity. Fortunately, despite their rarity, thousands of SNPs have been discovered in *B. anthracis* by interrogating large portions of the genome. Indeed, SNPs can even be found among closely related isolates if large enough portions of their genomes are compared. SNP discovery is currently cumbersome and expensive but, as whole-genome sequencing becomes more widespread (Cole et al., 1998; Perna et al., 2001; Fleischmann et al., 2002; Read et al., 2002), genomic comparisons will maximize the efficiency of SNP discovery (Cummings and Relman, 2002; Read et al., 2002). These types of techniques are capable of identifying hundreds or even thousands of SNPs within a single species.

The rarity of SNPs makes them important diagnostic markers in *B. anthracis*. SNP evolution suggests unique origins, as it is likely that each point mutation occurred just once in the phylogenetic history of the species. Thus, SNP markers are evolutionarily stable and unlikely to mutate again to either a novel or ancestral state. This stability can be invaluable for broadly defining strain groups, such as major phylogenetic divisions, as well as specifically defining a particular terminal branch strain (e.g., Ames group). If SNP discovery procedures are intensive and yield relatively large numbers of SNPs (hundreds to thousands), multiple loci will be identified along individual phylogenetic branches. As these SNPs provide the same phylogenetic information, they are diagnostically redundant and optimized subtyping assays can be developed by eliminating this redundancy. Following optimization, a relatively small number of SNPs can be used to define major genetic groups (one per group) in a bacterium like *B. anthracis*.

Canonical characters are diagnostic features that can be used for identifying a particular phylogenetic point in the evolutionary history of a species or set of organisms. In the case of long-branch lengths, there may be multiple phylogenetic characters that could be used, but designating one as canonical greatly reduces diagnostic redundancy. It is highly desirable from a diagnostic standpoint that the character be stable and not prone to homoplasy. In the evolution of *B. anthracis*, there are many SNPs that meet these conditions. Hence, we have developed a set of *canonical SNPs* (canSNPs) that identify the deeper nodes in our phylogenetic hypotheses for *B. anthracis*. To develop these canSNPs, we first had to discover rare SNPs and then determine their phylogenetic positions. Identification of canSNPs is more efficient if there is (1) an independent initial phylogenetic hypothesis, (2) discovery of a large number of SNPs, and (3) an extensive strain collection against which the discovered SNPs can be queried. An initial phylogenetic hypothesis is useful for guiding SNP discovery, as this process is based upon available genomic sequence and, therefore, may be highly biased (Pearson and Keim, unpublished data). In the case of *B. anthracis*, both AFLP (Keim et al., 1997) and VNTR (Keim et al., 2000) markers provided the initial hypothesis, dividing the *B. anthracis* phylogeny into several major clades (Keim et al., 2000; Van Ert and Keim, unpublished data). Our goal, therefore, in identifying canSNPs in *B. anthracis* was to find SNPs that defined major clades. We achieved this goal by first querying a large number of SNPs against a diverse set of 26 strains from our existing collection of over 1300 *B. anthracis* isolates. We then mapped the position of each SNP on our existing, initial phylogenetic hypothesis. One representative SNP marker from each of the major evolutionary branches was then selected as the defining SNP for that clade. Finally, each canSNP was tested against our strain collection of over 1300 isolates to ensure the validity of its canonical designation. The specific details of these procedures will be published elsewhere.

Routine analysis of SNPs for molecular typing requires an assay that is robust, as well as high capacity. Current techniques available for scoring SNPs include: allele-specific hybridization, oligonucleotide ligation assays, mini sequencing and real-time fluorescent PCR methods. A number of factors should be considered when deciding upon a SNP scoring system, including time required for assay development, cost, accuracy and desired through-put. Our preferred system for scoring known SNPs is an allelic discrimination as-

say using real-time PCR in conjunction with TaqMan-minor groove binding (MGB) probes. This assay is particularly attractive for high-throughput genotyping efforts (large numbers of samples coupled with a relatively small number of SNPs) since it combines the PCR amplification and detection steps and is also amenable to automation. TaqMan-MGB allelic discrimination assays can be rapidly designed around canSNP markers and allow thousands of samples to be analyzed in a single workday. An additional advantage of applying this technology to SNP scoring is that it can be performed on samples with very low DNA levels (sub nanogram), making it effective for environmental sampling. The ability to simultaneously perform low-level detection and genotyping of *B. anthracis* may prove to be an invaluable tool in forensic and biodefense applications.

5. High resolution analysis: multiple locus VNTR analysis (MLVA)

Amplified fragment length polymorphism (AFLP) analysis was an important first step in the molecular characterization of *B. anthracis* and led to the discovery and use of VNTRs in this pathogen (Keim et al., 1997). Rare variation was observed within the AFLP markers, which allowed successful differentiation among some *B. anthracis* isolates and identification of novel genetic lineages. Even this limited diversity in the AFLP markers was an improvement over other molecular typing methods (Harrell et al., 1995; Henderson et al., 1995), which failed to locate areas of genetic diversity within *B. anthracis*. Sequence analyses of AFLP markers revealed that, in many cases, VNTRs were responsible for the observed fragment length polymorphisms (Schupp et al., 2000); an arbitrarily primed PCR polymorphism (Henderson et al., 1995) was also shown to be a VNTR (Andersen et al., 1996). As VNTRs appeared to account for a significant portion of the distinguishing genomic features in *B. anthracis*, they were successfully developed as molecular markers (Keim et al., 2000; Read et al., 2003).

The discriminatory power of VNTRs is greatly enhanced if multiple loci are examined concurrently. Individual VNTR loci provide only a small amount of information, which can be confusing in some cases due to independently derived variants or reversal mutations. Fortunately, the methods used to detect length polymorphisms in VNTR regions, PCR amplification and fragment sizing, are amenable to multiplexing. This facilitates the creation of multiple-locus VNTR analysis (MLVA) systems, which increase the precision of estimating genetic relationships and mitigate, somewhat, the confounding effects of multiple mutations at a single locus. Indeed, the original eight-locus MLVA system for *B. anthracis* exhibits good genetic resolution and defines phylogenetic relationships consistent with the previous AFLP study (Keim et al., 2000).

One powerful feature of a MLVA system is the ability to simultaneously employ multiple VNTR markers that exhibit varying levels of diversity and, therefore, resolving power. Our current MLVA system for *B. anthracis* uses 15 VNTR loci, which have diversity values ranging from 0.32 to 0.85. VNTR loci with lower diversity values are useful for establishing deeper phylogenetic relationships, whereas markers with higher diversity values provide greater discriminatory power among closely related isolates. Although our MLVA system contains VNTR loci with very high diversity values, it is occasionally incapable of differentiating among *B. anthracis* isolates that are very closely related, such as those from a natural outbreak or a bioterrorism event. In these situations, additional genetic resolution is required for differentiation.

6. Highest resolution analysis: single nucleotide repeats (SNR)

Genetic differentiation among very closely related individuals requires the use of molecular markers that exhibit very high diversity and, therefore, have very high mutation rates. Single nucleotide repeats are a class of VNTRs that have been well characterized in the human genome, where they have been shown to exhibit extreme mutability (Mori et al., 2001; Zhang et al., 2001). These regions are mutational hot spots due to high occurrences of slipped-strand mispairing, which reduces the fidelity of DNA replication (Chung et al., 2003). Occurrence of these mispairings is positively correlated with the length of mononucleotide sequences, with longer arrays having higher mutation rates (Chung et al., 2003). SNRs have been identified in a number of bacterial species (Tomb et al., 1997; Gur-Arie et al., 2000; Josenhans et al., 2000; Read et al., 2002), including *B. anthracis* (Huynh and Keim, unpublished data), which provided an opportunity to develop these areas as very fine-scale markers.

We identified and developed a set of four very diverse SNR markers (SNR-4) for high-resolution molecular typing of *B. anthracis*. To locate potential SNR markers, we surveyed the *B. anthracis* genome for SNR regions that were at least 9 bp in length. We focused on areas with longer repeats to identify SNR regions with the highest potential mutation rates. Based on this criterion, our survey revealed more than 50 potential SNR markers, which were all poly-A/T tracts (Huynh and Keim, unpublished data). We evaluated the diversity of these potential markers against a set of dissimilar *B. anthracis* isolates and selected the four most diverse loci ($D = 0.80$–0.94). We estimate the mutation rates of these four markers to be as high as 6.0×10^{-4}, based on in vitro parallel serial passage experiments (Vogler and Keim, unpublished data).

The utility of SNR markers for determining phylogenetic relationships among *B. anthracis* isolates is a function of the evolutionary scale at which these markers are applied. The high mutation rates of SNR markers prohibit their use in determining phylogenetic relationships among diverse iso-

lates, as these analyses will likely be confounded by homoplasy. In contrast, SNR markers are extremely powerful for distinguishing among isolates in an outbreak scenario, where the predicted genetic diversity of the isolates is exceedingly low. Observed mutations in our SNR-4 system may even allow epidemiological insights into individual anthrax outbreaks, such as patterns of movement across the landscape. From a forensics perspective, our SNR-4 system provides unparalleled resolution for the investigation of both previous and future bioterrorism events that involve *B. anthracis*.

7. Maximizing phylogenetic accuracy across evolutionary scales in *B. anthracis* using PHRANA (progressive hierarchical resolving assays using nucleic acids)

The ultimate utility of SNPs, VNTRs and SNRs as informative genetic markers is largely dependent on their mutation rates, as well as the population being examined (Fig. 3). If marker mutation rates are extremely low, then polymorphisms in those markers will only be detected in highly diverse populations and less diverse populations will appear relatively monomorphic. Conversely, if mutation rates are high, diversity will quickly arise at these loci, enabling differentiation among even closely related strains. A high mutation rate, however, also increases the likelihood of homoplasy, which is also more likely to be found in populations with high genetic diversity. Thus, if markers with high mutation rates are used to analyze a diverse population, incorrect assignments will be inevitable.

When constructing a phylogenetic hypothesis, there are several methods that can be used to address the issue of homoplasy but none seem completely satisfactory. One strategy is to simply assume that the effects of homoplasy will be swamped out by other, more homologous characters. However, the degree to which homologous characters can counter the effects of homoplastic characters depends on the ratio of homologous to homoplastic characters. As this ratio decreases, so does the accuracy of the resulting phylogenetic hypothesis. Another strategy is to differentially weight characters to reduce homoplastic noise while at the same time preserving phylogenetic information (Farris, 1969; Goloboff, 1993). Unfortunately, determination of weights can sometimes be difficult or subjective, and errors introduced by improper weights can lead to misleading hypotheses.

Identification of an unknown bacterium is usually performed in a progressive and hierarchal fashion. For example, a diagnostic test (e.g., Gram staining) is initially carried out to determine the gross classification of the unknown, which is then followed by a series of other, more specific tests (e.g., lactose fermentation) to identify the unknown to genus or species. At each level of the identification process, the diagnostic test becomes more specific and the unknown is placed into an increasingly smaller group until, finally, the species is determined. The use of nested hierarchal analyses for traditional subtyping of bacteria is both practical and intuitive. This approach should also work well for subtyping on a finer scale, such as determining phylogenetic relationships within a single bacterial species.

The variable mutation rates and, therefore, discriminatory power among different genomic regions fit well with a

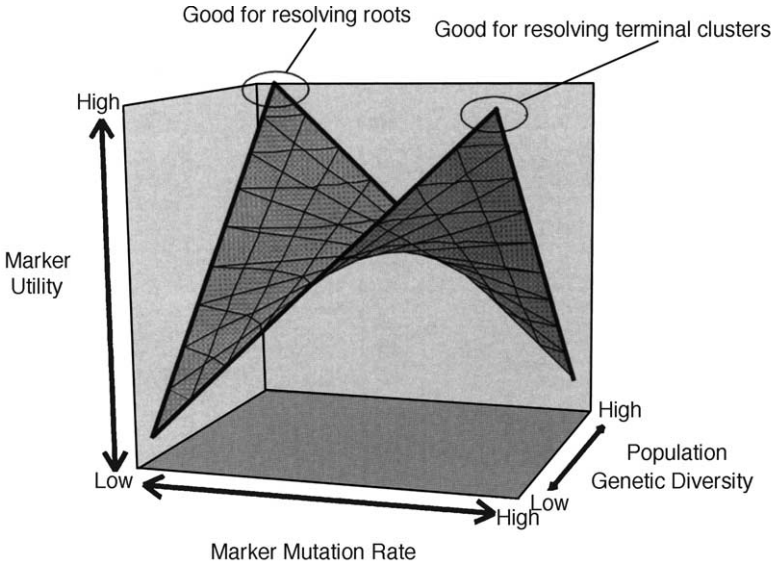

Fig. 3. Marker utility. The utility of a genetic marker for determining phylogenetic relationships within a given population is a function of (1) the mutation rate of the marker, and (2) the overall genetic diversity of the examined population. When genetic diversity of the population is low (e.g., a young, terminal phylogenetic cluster), only markers with high mutation rates, such as SNRs, will be able to differentiate among individuals in the population. Conversely, when genetic diversity of the population is high (e.g., an older and deeper phylogenetic group), only markers with low mutation rates, such as SNPs, will yield accurate phylogenetic patterns, as information obtained from markers with higher mutation rates is likely to be distorted by homoplasy.

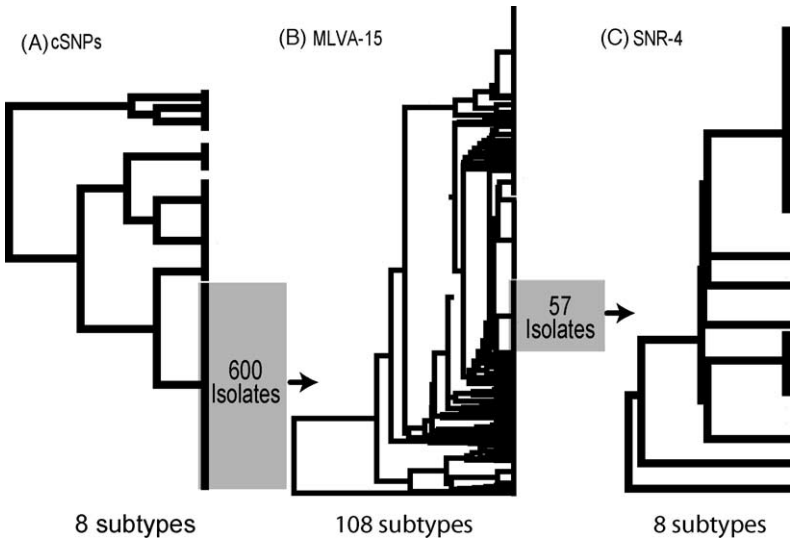

Fig. 4. PHRANA. (A) Eight canonical SNPs separated 1067 *B. anthracis* isolates into eight major phylogenetic groups. (B) MLVA-15 analysis of the 600 isolates in the largest cluster identified 108 unique types. (C) SNR-4 analysis of one of the MLVA types containing 57 isolates yielded eight unique PHRANA genotypes. In all, the 1067 isolates were broken into 209 MLVA-15 and 476 PHRANA genotypes (Van Ert et al. unpublished data).

nested hierarchal approach that overcomes the problem of marker weighting. Clearly, some types of genetic data are so central to determining major phylogenetic patterns that they should not be overwhelmed by large volumes of more peripheral data. This weighting problem can be overcome by using a nested hierarchal approach. In many organisms, phylogenetically stable markers, such as 16S rRNA, can be used initially to define major groups, while more mutable loci, such as house-keeping genes in MLST, can then be used to progressively define individual strains and isolates within those major groups. While this approach will be appropriate for many other species, the 16S and MLST regions provide little or no phylogenetic resolution within *B. anthracis*. Fortunately, canSNPs, VNTRs and SNRs do provide varying levels of resolution within *B. anthracis* and, therefore, are suitable for inclusion in this type of analysis.

Progressive hierarchical resolving assays using nucleic acids (PHRANA) is a nested hierarchal approach that employs canSNPs, MLVA-15 and, finally, the SNR-4 system to accurately characterize phylogenetic relationships among *B. anthracis* isolates. The step-wise hierarchical nature of this assay is vital because it progressively reduces the genetic diversity of the focused population, maximizing the utility of each marker type. PHRANA also alleviates the need for weighting among the three marker types, as they are all applied independently of one another. Thus, the phylogenetic hypothesis that results from PHRANA will be highly resolving, as well as accurate, since the effects of homoplasy will be minimized.

To demonstrate the use of PHRANA, consider the following example involving 1067 diverse *B. anthracis* isolates (Fig. 4). Initially, PHARANA uses eight canSNP markers to separate the 1067 isolates into eight major phylogenetic groups (Keim et al., 2000), the largest of which contains 600 isolates. At this stage of the analysis, there are just eight different genotypes and, as a result, there is relatively low genotype diversity ($D = 0.64$). The MLVA-15 system, a set of 15 VNTR loci combined into a multiplexed assay, is then applied independently to the isolates within each of eight groups. At this step, the 600 isolates in the largest group are subdivided into 108 unique MLVA genotypes (Fig. 4). Overall, the 1067 isolates are now separated into a total of 209 unique genotypes across the eight major groups (Van Ert et al. unpublished data), increasing the genotype diversity greatly ($D = 0.96$). The last step in PHRANA is to apply the high resolution SNR-4 assay to each of the 209 MLVA genotypes, which subdivides the 1067 isolates into 476 unique PHRANA genotypes (overall genotype diversity $= 0.98$). In our example (Fig. 4), 57 isolates with an identical MLVA genotype are further subdivided into eight unique PHRANA genotypes by the SNR-4 assay.

The use of canSNPs, MLVA-15 and SNR-4 in a nested hierarchical fashion in PHRANA has more resolving power and more accurately determines phylogenetic patterns within *B. anthracis* than any of these three assays used independently. In the example above, PHRANA resolved a total of 476 unique genotypes among 1067 *B. anthracis* isolates. In comparison, when applied individually to these same isolates, canSNPs, MLVA-15 and SNR-4 resolve 8, 209 and 418 genotypes, respectively. Although the SNR-4 system is capable of resolving a large number of genotypes when applied alone to this group of *B. anthracis* isolates, the resulting phylogenetic hypothesis would be plagued by homoplasy and the deeper phylogenetic patterns would be impossible to discern. In contrast, canSNPs or MLVA-15 are better able to discern deeper phylogenetic patterns, but will lack the resolution of SNR-4. Thus, use of PHRANA for phylogenetic analysis of *B. anthracis* is superior to the independent

use of any of its three component analyses as it both identifies more unique genotypes and also preserves phylogenetic patterns across multiple evolutionary scales.

8. Conclusion

The genomes of pathogenic bacteria are relatively large, providing the opportunity to selectively choose optimal strategies for molecular subtyping. This involves the inevitable trade-off between selecting markers with high genetic resolution and selecting markers that accurately describes large evolutionary relationships. A common solution for trying to overcome this trade-off is to combine data from several marker types that exhibit variable levels of genetic diversity and, therefore, discriminatory power. Unfortunately, combining data from markers with very different levels of genetic diversity requires that these data be weighted prior to phylogenetic analyses, which can be subjective and problematic. The PHRANA system described here overcomes these problems because it is (1) nested and, therefore, does not require weighting, and (2) hierarchical, so it captures phylogenetic patterns at multiple evolutionary scales while maintaining high genetic resolution. The applicability of PHRANA is not limited to *B. anthracis*. For example, PHRANA could be applied to *Mycobacterium tuberculosis*, as both SNP and VNTR markers have been developed for this pathogen (Gutacker et al., 2002; Alland et al., 2003; Spurgiesz et al., 2003). Indeed, these methods can easily be applied to many other organisms as markers with variable diversity values are discovered.

Acknowledgements

This work was supported by grants from National Institutes of Health—General Medical Sciences, Department of Energy—Chem Bio Non Proliferation program, National Science Foundation and the Cowden Endowment for Microbiology. We also thank our collaborators at TIGR (Jacques Ravel and Timothy Read) and NAU (Joseph Busch, Ryan Easterday, Sergey Kachur, Rebecca Leadem, Shane Rhoton, Tatum Simonson, Daniel Solomon, Jana U'Ren and Shaylan Zanecki) for allowing us to discuss the implications of unpublished data.

References

Alland, D., Whittam, T.S., Murray, M.B., Cave, M.D., Hazbon, M.H., Dix, K., Kokoris, M., Duesterhoeft, A., Eisen, J.A., Fraser, C.M., Fleischmann, R.D., 2003. Modeling bacterial evolution with comparative-genome-based marker systems: application to *Mycobacterium tuberculosis* evolution and pathogenesis. J. Bacteriol. 185, 3392–3399.

Andersen, G.L., Simchock, J.M., Wilson, K.H., 1996. Identification of a region of genetic variability among *Bacillus anthracis* strains and related species. J. Bacteriol. 178, 377–384.

Andreassen, R., Egeland, T., Olaisen, B., 1996. Mutation rate in the hypervariable vntr g3 (d7s22) is affected by allele length and a flanking DNA sequence polymorphism near the repeat array. Am. J. Hum. Genet. 59, 360–367.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., Rolf, B., 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am. J. Hum. Genet. 62, 1408–1415.

Buard, J., Bourdet, A., Yardley, J., Dubrova, Y., Jeffreys, A.J., 1998. Influences of array size and homogeneity on minisatellite mutation. EMBO J. 17, 3495–3502.

Chung, K.Y., Kim, N.G., Li, L.S., Kim, H., Kim, H., Nam, C.M., Kim, H., Shin, D.H., 2003. Clinicopathologic characteristics related to the high variability of coding mononucleotide repeat sequences in tumors with high-microsatellite instability. Oncol. Rep. 10, 439–444.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, L., Gas, S., Barry III, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Barrell, B.G., et al., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature 393, 537–544.

Cummings, C.A., Relman, D.A., 2002. Genomics and microbiology. Microbial forensics—"cross-examining pathogens". Science 296, 1976–1979.

Eckert, K.A., Yan, G., 2000. Mutational analyses of dinucleotide and tetranucleotide microsatellites in *Escherichia coli*: influence of sequence on expansion mutagenesis. Nucleic Acids Res. 28, 2831–2838.

Eisen, J.A., 1999. Mechanistic basis for microsatellite instability. In: Goldstein, D.B., Schlötterer, C. (Eds.), Microsatellites Evolution and Applications. Oxford University Press, New York, pp. 34–48.

Farris, J.S., 1969. A successive approximations approach to character weighting. Syst. Zool. 18, 374–385.

Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., Deboy, R., Dodson, R., Gwinn, M., Haft, D., Hickey, E., Kolonay, J.F., Nelson, W.C., Umayam, L.A., Ermolaeva, M., Salzberg, S.L., Delcher, A., Utterback, T., Weidman, J., Khouri, H., Gill, J., Mikula, A., Bishai, W., Jacobs Jr., W.R., Venter, J.C., Fraser, C.M., 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol. 184, 5479–5490.

Freudenreich, C.H., Stavenhagen, J.B., Zakian, V.A., 1997. Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol. Cell. Biol. 17, 2090–2098.

Goloboff, P.A., 1993. Estimating character weights during tree search. Cladistics 9, 83–91.

Gur-Arie, R., Cohen, C.J., Eitan, Y., Shelef, L., Hallerman, E.M., Kashi, Y., 2000. Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition and polymorphism. Genome Res. 10, 62–71.

Gutacker, M.M., Smoot, J.C., Lux Migliaccio, C.A., Ricklefs, S.M., Hua, S., Cousins, D.V., Graviss, E.A., Shashkina, E., Kreiswirth, B.N., Musser, J.M., 2002. Genome-wide analysis of synonymous single nucleotide polymorphisms in *Mycobacterium tuberculosis* complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics 162, 1533–1543.

Guttman, D.S., Dykhuizen, D.E., 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266, 1380–1383.

Harrell, L.J., Andersen, G.L., Wilson, K.H., 1995. Genetic variability of *Bacillus anthracis* and related species. J. Clin. Microbiol. 33, 1847–1850.

Hartenstine, M.J., Goodman, M.F., Petruska, J., 2000. Base stacking and even/odd behavior of hairpin loops in DNA triplet repeat slippage and expansion with DNA polymerase. J. Biol. Chem. 275, 18382–18390.

Henderson, S.T., Petes, T.D., 1992. Instability of simple sequence DNA in *Saccharomyces cerevisiae*. Mol. Cell. Biol. 12, 2749–2757.

Henderson, I., Yu, D., Turnbull, P.C., 1995. Differentiation of *Bacillus anthracis* and other '*Bacillus cereus* group' bacteria using IS231-derived sequences. FEMS Microbiol. Lett. 128, 113–118.

Hoffmaster, A.R., Fitzgerald, C.C., Ribot, E., Mayer, L.W., Popovic, T., 2002. Molecular subtyping of *Bacillus anthracis* and the 2001 bioterrorism-associated anthrax outbreak United States. Emerg. Infect. Dis. 8, 1111–1116.

Jackson, P.J., Hugh-Jones, M.E., Adair, D.M., Green, G., Hill, K.K., Kuske, C.R., Grinberg, L.M., Abramova, F.A., Keim, P., 1998. PCR analysis of tissue samples from the 1979 Sverdlovsk anthrax victims: the presence of multiple *Bacillus anthracis* strains in different victims. Proc. Natl. Acad. Sci. U.S.A. 95, 1224–1229.

Josenhans, C., Eaton, K.A., Thevenot, T., Suerbaum, S., 2000. Switching of flagellar motility in *Helicobacter pylori* by reversible length variation of a short homopolymeric repeat in *flip*, a gene encoding a basal body protein. Infect. Immun. 68, 4598–4603.

Keim, P., Kalif, A., Schupp, J., Hill, K., Travis, S.E., Richmond, K., Adair, D.M., Hugh-Jones, M., Kuske, C.R., Jackson, P., 1997. Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers. J. Bacteriol. 179, 818–824.

Keim, P., Price, L.B., Klevytska, A.M., Smith, K.L., Schupp, J.M., Okinaka, R., Jackson, P.J., Hugh-Jones, M.E., 2000. Multiple-locus variable-number tandem repeat analysis reveals genetic relationships within *Bacillus anthracis*. J. Bacteriol. 182, 2928–2936.

Keim, P., Smith, K.L., Keys, C., Takahashi, H., Kurata, T., Kaufmann, A., 2001. Molecular investigation of the Aum Shinrikyo anthrax release in Kameido. Japan. J. Clin. Microbiol. 39, 4566–4567.

Levinson, G., Gutman, G.A., 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. 4, 203–221.

Meselson, M., Guillemin, J., Hugh-Jones, M., Langmuir, A., Popova, I., Shelokov, A., Yampolskaya, O., 1994. The Sverdlovsk anthrax outbreak of 1979. Science 266, 1202–1208.

Mori, Y., Yin, J., Rashid, A., Leggett, B.A., Young, J., Simms, L., Kuehl, P.M., Langenberg, P., Meltzer, S.J., Stine, O.C., 2001. Instabilotyping: comprehensive identification of frameshift mutations caused by coding region microsatellite instability. Cancer Res. 61, 6046–6049.

Okinaka, R.T., Cloud, K., Hampton, O., Hoffmaster, A.R., Hill, K.K., Keim, P., Koehler, T.M., Lamke, G., Kumano, S., Mahillon, J., Manter, D., Martinez, Y., Ricke, D., Svensson, R., Jackson, P.J., 1999. Sequence and organization of pXO1, the large *Bacillus anthracis* plasmid harboring the anthrax toxin genes. J. Bacteriol. 181, 6509–6515.

Parniewski, P., Jaworski, A., Wells, R.D., Bowater, R.P., 2000. Length of CTG.CAG repeats determines the influence of mismatch repair on genetic instability. J. Mol. Biol. 299, 865–874.

Perna, N.T., Plunkett III, G., Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., Posfai, G., Hackett, J., Klink, S., Boutin, A., Shao, Y., Miller, L., Grotbeck, E.J., Davis, N.W., Lim, A., Dimalanta, E.T., Potamousis, K.D., Apodaca, J., Anantharaman, T.S., Lin, J., Yen, G., Schwartz, D.C., Welch, R.A., Blattner, F.R., 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409, 529–533.

Price, L.B., Hugh-Jones, M., Jackson, P.J., Keim, P., 1999. Genetic diversity in the protective antigen gene of *Bacillus anthracis*. J. Bacteriol. 181, 2358–2362.

Read, T.D., Salzberg, S.L., Pop, M., Shumway, M., Umayam, L., Jiang, L., Holtzapple, E., Busch, J.D., Smith, K.L., Schupp, J.M., Solomon, D., Keim, P., Fraser, C.M., 2002. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 296, 2028–2033.

Read, T.D., Peterson, S.N., Tourasse, N., Baillie, L.W., Paulsen, I.T., Nelson, K.E., Tettelin, H., Fouts, D.E., Eisen, J.A., Gill, S.R., Holtzapple, E.K., Okstad, O.A., Helgason, E., Rilstone, J., Wu, M., Kolonay, J.F., Beanan, M.J., Dodson, R.J., Brinkac, L.M., Gwinn, M., Deboy, R.T., Madpu, R., Daugherty, S.C., Durkin, A.S., Haft, D.H., et al., 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. Nature 423, 81–86.

Rosche, W.A., Jaworski, A., Kang, S., Kramer, S.F., Larson, J.E., Geidroc, D.P., Wells, R.D., Sinden, R.R., 1996. Single-stranded DNA-binding protein enhances the stability of CTG triplet repeats in *Escherichia coli*. J. Bacteriol. 178, 5042–5044.

Schupp, J.M., Klevytska, A.M., Zinser, G., Price, L.B., Keim, P., 2000. vrrb, a hypervariable open reading frame in *Bacillus anthracis*. J. Bacteriol. 182, 3989–3997.

Smith, K.L., Devos, V., Bryden, H., Price, L.B., Hugh-Jones, M.E., Keim, P., 2000. *Bacillus anthracis* diversity in Kruger National Park. J. Clin. Microbiol. 38, 3780–3784.

Spurgiesz, R.S., Quitugua, T.N., Smith, K.L., Schupp, J., Palmer, E.G., Cox, R.A., Keim, P., 2003. Molecular typing of *Mycobacterium tuberculosis* by using nine novel variable-number tandem repeats across the Beijing family and low-copy-number IS6110 isolates. J. Clin. Microbiol. 41, 4224–4230.

Strand, M., Prolla, T.A., Liskay, R.M., Petes, T.D., 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365, 274–276.

Sylvestre, P., Couture-Tosi, E., Mock, M., 2003. Polymorphism in the collagen-like region of the *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. J. Bacteriol. 185, 1555–1563.

Takahashi, H., Keim, P., Kaufmann, A.F., Smith, K.L., Keys, C., Taniguchi, K., Inouye, S., Kurata, T., 2004. Epidemiological and laboratory investigation of a *Bacillus anthracis* bioterrorism incident, Kameido, Tokyo, 1993. Emerg. Infect. Dis. 10, 117–120.

Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Venter, J.C., et al., 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature 388, 539–547.

Vogler, A.J., Busch, J.D., Percy-Fine, S., Tipton-Hunton, C., Smith, K.L., Keim, P., 2002. Molecular analysis of rifampin resistance in *Bacillus anthracis* and *Bacillus cereus*. Antimicrob. Agents Chemother. 46, 511–513.

Weber, J.L., Wong, C., 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. 2, 1123–1128.

Yang, Y., Masker, W., 1996. Instability of repeated dinucleotides in bacteriophage T7 genomes. Mutat. Res. 354, 113–127.

Zhang, L., Yu, J., Wilson, J.K.V., Markowitz, S.D., Kinzler, K.W., Vogelstein, B., 2001. Short mononucleotide repeat sequence variability in mismatch repair-deficient cancers. Cancer Res. 61, 3801–3805.