Phylogenetic discovery bias in *Bacillus anthracis* using single-nucleotide polymorphisms from whole-genome sequencing

Talima Pearson*, Joseph D. Busch*, Jacques Ravel†, Timothy D. Read†, Shane D. Rhoton*, Jana M. U'Ren*, Tatum S. Simonson*, Sergey M. Kachur*, Rebecca R. Leadem*, Michelle L. Cardon*, Matthew N. Van Ert*, Lynn Y. Huynh*, Claire M. Fraser†, and Paul Keim*‡§

*Department of Biology, Northern Arizona University, Flagstaff, AZ 86011-5640; †Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850; and ‡Translational Genomics Research Institute, 400 Fifth Street, Suite 1600, Phoenix, AZ 85004

Edited by Stanley Falkow, Stanford University, Stanford, CA, and approved July 29, 2004 (received for review May 29, 2004)

Phylogenetic reconstruction using molecular data is often subject to homoplasy, leading to inaccurate conclusions about phylogenetic relationships among operational taxonomic units. Compared with other molecular markers, single-nucleotide polymorphisms (SNPs) exhibit extremely low mutation rates, making them rare in recently emerged pathogens, but they are less prone to homoplasy and thus extremely valuable for phylogenetic analyses. Despite their phylogenetic potential, ascertainment bias occurs when SNP characters are discovered through biased taxonomic sampling; by using whole-genome comparisons of five diverse strains of *Bacillus anthracis* to facilitate SNP discovery, we show that only polymorphisms lying along the evolutionary pathway between reference strains will be observed. We illustrate this in theoretical and simulated data sets in which complex phylogenetic topologies are reduced to linear evolutionary models. Using a set of 990 SNP markers, we also show how divergent branches in our topologies collapse to single points but provide accurate information on internodal distances and points of origin for ancestral clades. These data allowed us to determine the ancestral root of *B. anthracis*, showing that it lies closer to a newly described "C" branch than to either of two previously described "A" or "B" branches. In addition, subclade rooting of the C branch revealed unequal evolutionary rates that seem to be correlated with ecological parameters and strain attributes. Our use of nonhomoplastic whole-genome SNP characters allows branch points and clade membership to be estimated with great precision, providing greater insight into epidemiological, ecological, and forensic questions.

The discovery method used to find nonhomoplastic molecular characters can have a great impact on subsequent phylogenetic analyses (1). "Discovery bias" will most likely be encountered when polymorphic characters are rare and extraordinary means (i.e., whole-genome sequencing) are necessary to find them. In such cases, only a small subset of the target organisms can be exhaustively examined for genetic differences. When these polymorphisms are assayed across a more extensive and diverse panel of strains, the phylogenetic topology will be impacted greatly by the original choices of strains used for character discovery. Consider a six-taxon dendrogram (Fig. 1) in which nonhomoplastic polymorphic characters are discovered by exhaustive comparisons of two taxa. Mutational events that create character differences on the evolutionary path connecting the two taxa will be discovered, but characters on tangential (i.e., secondary) branches will not. Subsequent character-state determination in all taxa will accurately place them onto a linear phylogeny but will underrepresent the taxonomic complexity by eliminating secondary branching. If the initial character discovery is between two closely related taxa, large diverse clades will be reduced to a single point on a short linear dendrogram (Fig. 1*B*). This theoretical model assumes that no additional character variation is fortuitously discovered when all taxa are assayed for character variation. This will be true when taxonomic groups are evolutionarily very young and, hence, have not had enough time to accumulate differences. In older taxonomic groups and nonclonal organisms, conflicting character states (i.e., homoplasies) may be discovered in subsequent analyses that reflect evolutionary reversals, convergence, parallelism, multiple step mutations, or genetic recombination. By contrast, young evolutionary groups and clonal bacteria typically have less homoplasy than older groups or nonrecombining organisms. As a result, finding molecular characters that exhibit low homoplasy will be more challenging in older evolutionary groups or with groups that exhibit horizontal gene transfer.

Bacillus anthracis, the causative bacterium of anthrax, is a recently emerged pathogen with characteristics of a young evolutionary group. It occurs primarily as quiescent spores that exist for long periods of time in the soil between host infections. When inside a host, *B. anthracis* cells emerge from dormancy, undergo rapid proliferation for the remainder of the life of their host, and then return to the soil as spores. This episodic reproductive cycle results in massive numbers of spores produced over a short evolutionary time (measured in generations). Consequently, opportunity for accumulating DNA mutations is limited, resulting in relatively little genetic variation within the species (2–4). *B. anthracis* has received considerable attention because of its potential as a biological weapon and difficulties associated with forensic tracking of this genetically homogeneous species (4, 5). The threat of *B. anthracis* and the difficulties in strain differentiation underscore the importance of fully understanding the ecology, evolution, and epidemiology of this species.

Recently, vast resources have been allocated toward developing diagnostic characters for *B. anthracis* through whole-genome sequencing (5, 6). In this study, we present an example of exhaustive but highly biased character discovery from the pathogenic bacterium *B. anthracis*. Whole-genome sequences were compared to find rare single-nucleotide polymorphism (SNP) characters that define relationships of *B. anthracis* strains with great precision but in a highly biased fashion. Characteristics of a young evolutionary age for this species are illustrated by the rarity of SNPs and very low character homoplasy. Using appropriate outgroups, we were also able to determine the evolutionary root of our *B. anthracis* phylogeny. Our work represents a model for future whole-genome analyses of any organism for phylogenetic purposes but especially for bacterial pathogens with genomes that are now readily available.

This paper was submitted directly (Track II) to the PNAS office.

Freely available online through the PNAS open access option.

Abbreviations: SNP, single-nucleotide polymorphism; VNTR, variable-number tandem repeat; MLVA, multiple-locus VNTR analysis; OTU, operational taxonomic unit.

[§] To whom correspondence should be sent at the * address. E-mail: paul.keim@nau.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0403844101

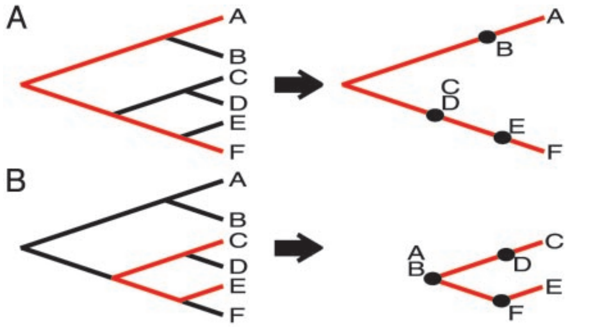


Fig. 1. Evolutionary model showing the consequences of biased character discovery for nonhomoplastic molecular markers. (*Left*) The ''true'' path structure of OTUs A–F is shown. (*A*) When OTUs A and F are used for comparative character (i.e., SNPs) discovery, only mutations on the connecting evolutionary path (red) will be discovered, resulting in the disappearance of all secondary branches but showing accurate node positions of all other OTUs. (*B*) Similarly, if C and E are used for character discovery, only mutations on the connecting path will be discovered, causing A and B to collapse at a single point. Again, accurate node positions are retained.

Materials and Methods

Five *B. anthracis* strains were selected for whole-genome sequencing based on relationships obtained from previous work using multiple-locus variable-number tandem repeat (VNTR) analysis (MLVA) (4) (Fig. 2). This sequencing effort was pursued by closure and finishing for one of the genomes [Ames: operational taxonomic unit (OTU) 26] and to 12× draft coverage for the other four genomes (OTUs 1, 2, 5, and 16). Whole-genome comparisons of these strains to the Ames (OTU 26) reference revealed SNPs, which were converted into a high-throughput assay based on the SNaPshot primer-extension protocol (Applied Biosystems, Foster City, CA). A large subset of the SNPs were verified on these five strains and used to determine character states of 26 diverse *B. anthracis* strains (Table 1) and three ''near neighbors'' (*Bacillus cereus* and *Bacillus thuringiensis*) selected from amplified fragment length polymorphism profiles (7). The near neighbors were included as an outgroup that allowed the root of the *B. anthracis* phylogeny to be determined.

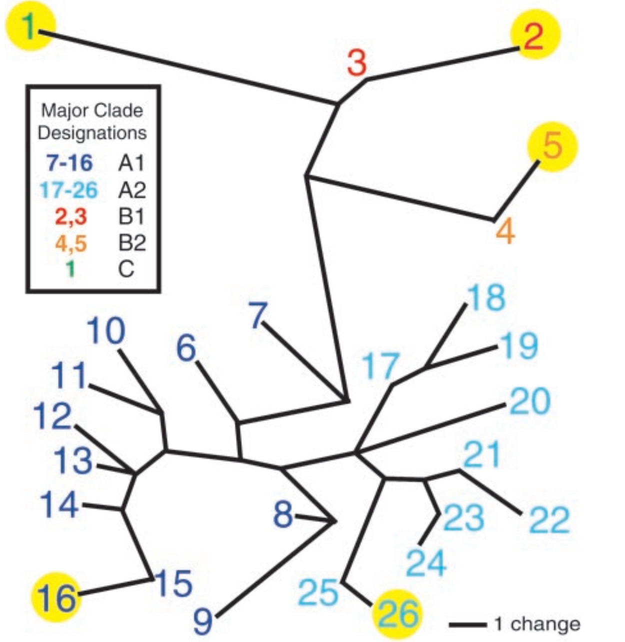


Fig. 2. Unrooted phylogenetic tree of 26 diverse *B. anthracis* strains based on variation at 15 VNTR loci (M.N.V.E., unpublished data). No near relatives are used because the VNTR primers are specific to *B. anthracis*. The homoplasy index = 0.3721, and branch lengths are proportional to VNTR mutational steps. Numbers refer to different strains that are grouped by color into major clade designations (for more strain information, see Table 1, which is published as supporting information on the PNAS web site). Reference strains used for whole-genome comparisons are indicated by a yellow circle. Strain 1, branch C (A1055); strain 2, Kruger B1; strain 5, CNEVA 9066 B2; strain 16, North America 1; strain 26, Ames.

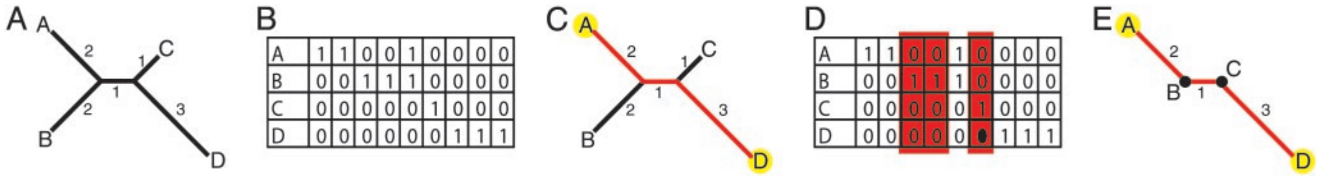
A model binary data set was used to simulate discovery bias when SNPs detected in different reference genomes were used for phylogenetic reconstruction. This simulation was based on MLVA branch distances among diverse *B. anthracis* strains (Fig. 2). The MLVA tree shown in Fig. 2 was selected because it was the most similar to other phylogenetic hypotheses (4). Working backward from this tree, we generated model binary data based on MLVA branch distances (see Fig. 3 for a simplified example). To simulate discovery bias, we eliminated all loci with allelic states that did not contrast between reference strains (Fig. 3). Such loci are not polymorphic between reference strains and, thus, would not be discovered by comparing the two genomes. We identified contrasting allelic states for strain 2 vs. strain 26 (Fig. 4*A*) and strain 16 vs. strain 26 (Fig. 4*B*) and then used those differences for phylogenetic reconstructions of a diverse set of strains.

The software package PAUP 4.0 b10 (8) was used to reconstruct unrooted phylogenies based on MLVA and binary model data. Heuristic searches saving a maximum of 100 trees produced 100 MLVA trees and 1 tree based on model data. WINCLADA 1.00.08 (9) was used to reconstruct SNP-based phylogenies. Heuristic analyses that saved a maximum of 1,000 trees with 10 replications were used. A consensus tree was used when more than one tree was equally parsimonious. Neighbor joining using MEGA 2.1 (10) was used to confirm SNP-based reconstructions and was based on a simple matching coefficient.

Results

Binary Character Modeling. Consistent with the theory depicted in Fig. 1, analysis of biased-discovery SNPs resulted in linear phylogenies with no secondary branching and with reference strains located at the termini. Complex evolutionary branches peripheral to the direct connecting pathway between reference OTUs collapsed onto single precise points. However, the arrangement and positions of these points on the linear evolutionary path were maintained correctly. Our results show that binary modeling of complex phylogenetic information is a simple method for predicting the effects of discovery bias in organisms such as *B. anthracis*, that exhibit characteristics of a young evolutionary age (see Fig. 3 and Table 2, which is published as supporting information on the PNAS web site).

SNPs from Whole-Genome Sequencing. Whole-genome sequencing predicted ≈3,500 SNPs among five diverse strains of *B. anthracis* (J.R., unpublished data). MLVA-based phylogenetic reconstructions show the positions of these five strains (OTUs 1, 2, 5, 16, and 26) relative to 21 other diverse strains (Fig. 5*A*). The number of SNPs found by comparing each genome with OTU 26 and the number of SNPs shared between comparisons (Fig. 5*B*) can be used to construct a basic dendrogram that depicts the number of characters on different branches of a tree (Fig. 5*C*). Consistent with previous MLVA, strain 1 from the C branch was the most distantly related of the *B. anthracis* genomes (Figs. 5*C* and 6*E*). Although other relationships were topologically similar to MLVA predictions, small discrepancies existed between MLVA and SNP dendrograms. For example, in the VNTR (MLVA) dendrogram (Fig. 2), strain 7, not strain 6, is the first A1 strain to branch off the main backbone, as depicted in the SNP results



A	1	1	0	0	1	0	0	0	0
B	0	0	1	1	1	0	0	0	0
C	0	0	0	0	0	1	0	0	0
D	0	0	0	0	0	0	1	1	1

A	1	1	0	0	1	0	0	0	0
B	0	0	1	1	1	0	0	0	0
C	0	0	0	0	0	1	0	0	0
D	0	0	0	0	0	●	1	1	1

Fig. 3. Binary modeling of a four-taxon phylogeny. Conversion of a multiple character-state (but unbiased) phylogeny involves the following steps. First, converting a phylogenetic tree into binary data requires that the branch lengths be known (*A*). Next, a ''1'' is assigned to a terminal OTU, and a ''0'' is assigned to all other OTUs. Data then are created for multiple identical ''markers'' corresponding to branch length in the unbiased phylogeny. For example, the branch length leading to OTU ''A'' is 2, therefore data from two markers should assign a ''1'' to the terminal OTU and a ''0'' to all others. For internal branches that lead to multiple OTUs, one character state was assigned to each OTU within the group and the other state to each OTU outside the group. Once again, the number of markers created should correspond to branch lengths (*B*). To simulate biased character discovery (i.e., between OTUs A and D), (*C*) markers with character states that show no difference between reference strains (circled) would not be discovered and thus were deleted (highlighted in red) (*D*) before phylogenetic analysis. Phylogenetic reconstruction on simulated and, subsequently, bias-sorted data sets shows branch collapse caused by biased character discovery (*E*).

(Fig. 6*E*). Furthermore, the VNTR hypothesis does not group strain 7 with strains 8 and 9. These small discrepancies are undoubtedly the result of homoplasy coupled with weak character support in the VNTR system. By contrast, because of their evolutionary stability, SNP data show low homoplasy and therefore a more accurately defined phylogeny.

Of the predicted SNPs chosen based on sequence quality, ≈20% (990) were developed into high-capacity genotyping assays for defining evolutionary relationships among a larger set of 26 diverse strains (Fig. 2) and near neighbors (see Table 1), and ≈17% of the verified SNPs were located in intergenic regions, with the remaining genic SNPs being a mixture of synonymous (24%) and nonsynonymous (58%) mutations. Only one confirmed phylogenetic character-state conflict was observed among the *B. anthracis* strains assayed (1 of >25,000 data points) (see Fig. 7, which is published as supporting information on the PNAS web site). This homoplastic locus was a nonsynonymous mutation that caused an amino acid change from valine to alanine in a hypothetical protein-coding sequence, which could have selective consequences.

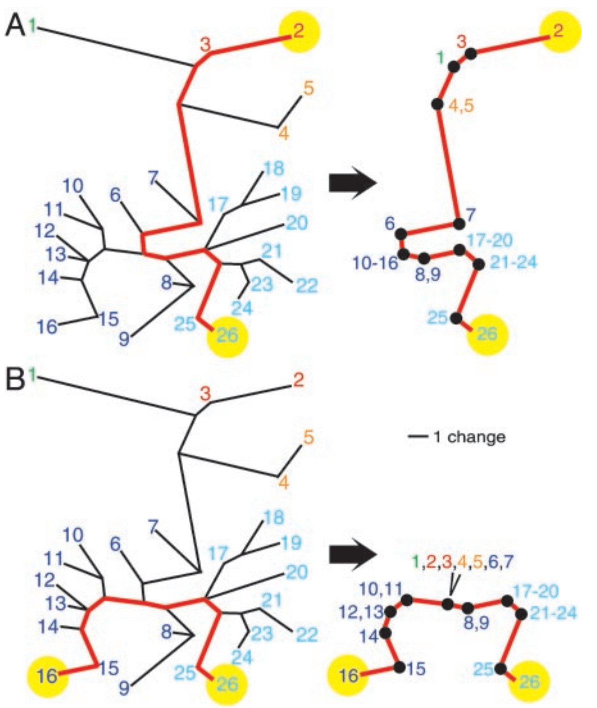


Fig. 4. Consequences of biased character discovery on a model phylogeny of *B. anthracis* with no homoplasy (see Table 2). Phylogenetic results using strains ''2'' and ''26'' for discovery (*A*) and ''16'' and ''26'' as discovery strains (*B*) exhibit the expected collapse of secondary branching and the retention of node positions (*Right*). (*Left*) The original tree is shown, with the discovery pathways designated in red. Reference strains are denoted with a yellow circle, and major clade designations are indicated with color as for Fig. 2.

A large number of nonhomoplastic evolutionary characters has allowed very precise phylogenetic estimations of branch lengths and positions of last common ancestors (Fig. 6 and Table 3, which is published as supporting information on the PNAS web site) among diverse strains of *B. anthracis*. As expected from theory (see the Introduction and Fig. 1) and modeling (Figs. 3 and 4), branches secondary to the primary evolutionary lineage connecting reference strains were absent. The lack of secondary branching argues strongly for the stability of these characters and the clonality of the lineages. Our simulated data sets also show that as homoplasy increases, so does secondary branching (see Fig. 8, which is published as supporting information on the PNAS web site). Overall, the large number of nonhomoplastic SNP characters reveal a precise evolutionary arrangement that can be used to reconstruct accurate phylogenetic topologies for diverse strains (6).

Phylogenetic Rooting. Phylogenetic rooting has been problematic in *B. anthracis* because of the lack of characters that could be polarized into ancestral and derived states relative to an appropriate outgroup (VNTRs, which vary within *B. anthracis*, can be highly homoplasious and generally reach mutational saturation in the outgroups). The large number of highly stable SNP loci and the identification of close *B. cereus*/*B. thuringiensis* relatives (7) make outgroup rooting possible. For each of the four SNP phylogenies (Fig. 6 *A–D*), the root can be determined as the node containing the outgroup strains (OTUs 27–29; Fig. 6). When *B. anthracis* reference strains are chosen with a connecting evolutionary path that passes through the root (Fig. 6*C*), the outgroup will be contained by itself at a node, indicating the precise location of the root for the *B. anthracis* phylogeny. On the other hand, when reference strains with a connecting evolutionary path that does not pass through the root are chosen (Fig. 6 *A*, *B*, and *D*), other strains including the outgroup will occupy the ancestral node. The position of this node is highly accurate because of the numerous characters, but it indicates the root of only a subclade and not the root to the entire *B. anthracis* phylogeny. When the four linear dendrograms are combined and branch lengths are adjusted for sampling differences (Fig. 6*E*), a more traditional topology with branches emerges, although this is still very much determined by the genomes used for character discovery. Care must be taken when combining loci discovered between different pairwise comparisons, because branch lengths depend on the percentage of SNPs assayed out of the total number that were discovered.

Rooting the phylogeny and determining major groups within *B. anthracis* along with accurate branch-length measurements

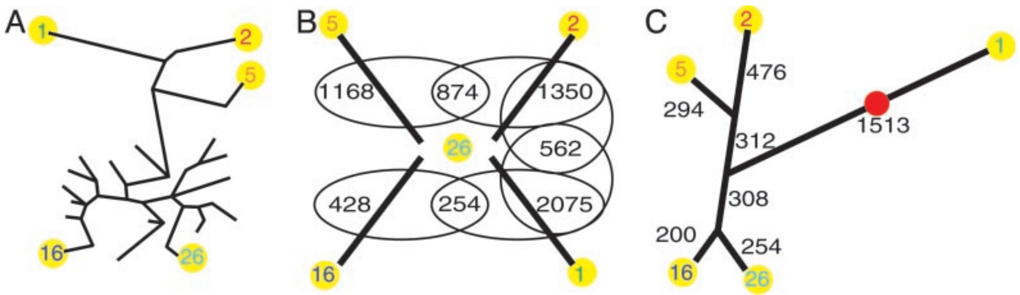


Fig. 5. Five diverse strains selected for whole-genome sequencing (yellow circles). (*A*) Phylogenetic locations of these five strains are shown with 21 other diverse *B. anthracis* strains (see Fig. 2). (*B*) High-quality sequences of four strains were compared to strain 26, the Ames strain. (*C*) The number of SNPs detected from these comparisons and the number of SNPs shared between comparisons (circled) are shown and can be used to accurately estimate evolutionary distances and phylogenetic topology. The approximate position of the phylogenetic root (from Fig. 6*E*) is denoted with a red circle.

lend to comparisons of historical evolutionary rates. The large character set and lack of homoplasy allow highly precise branch-length comparisons to detect unequal rates of evolution. Most notable is the C branch, which is much shorter than the A+B branches (Fig. 6). Also, the B2 branch leading to strain 5 is significantly shorter than the B1 branch that terminates with strain 2. Likewise, the A1 branch ending with strain 16 is shorter than the A2 branch of strain 26 (Fig. 6). These differences are significant (see Table 4, which is published as supporting information on the PNAS web site) and suggest greatly differing rates of evolution among *B*. *anthracis* lineages.

Discussion

Our data reveal an accurate method for defining major phylogenetic lineages of *B. anthracis*. We identify three major lineages (A–C) with the ancestral root located between the A+B and C

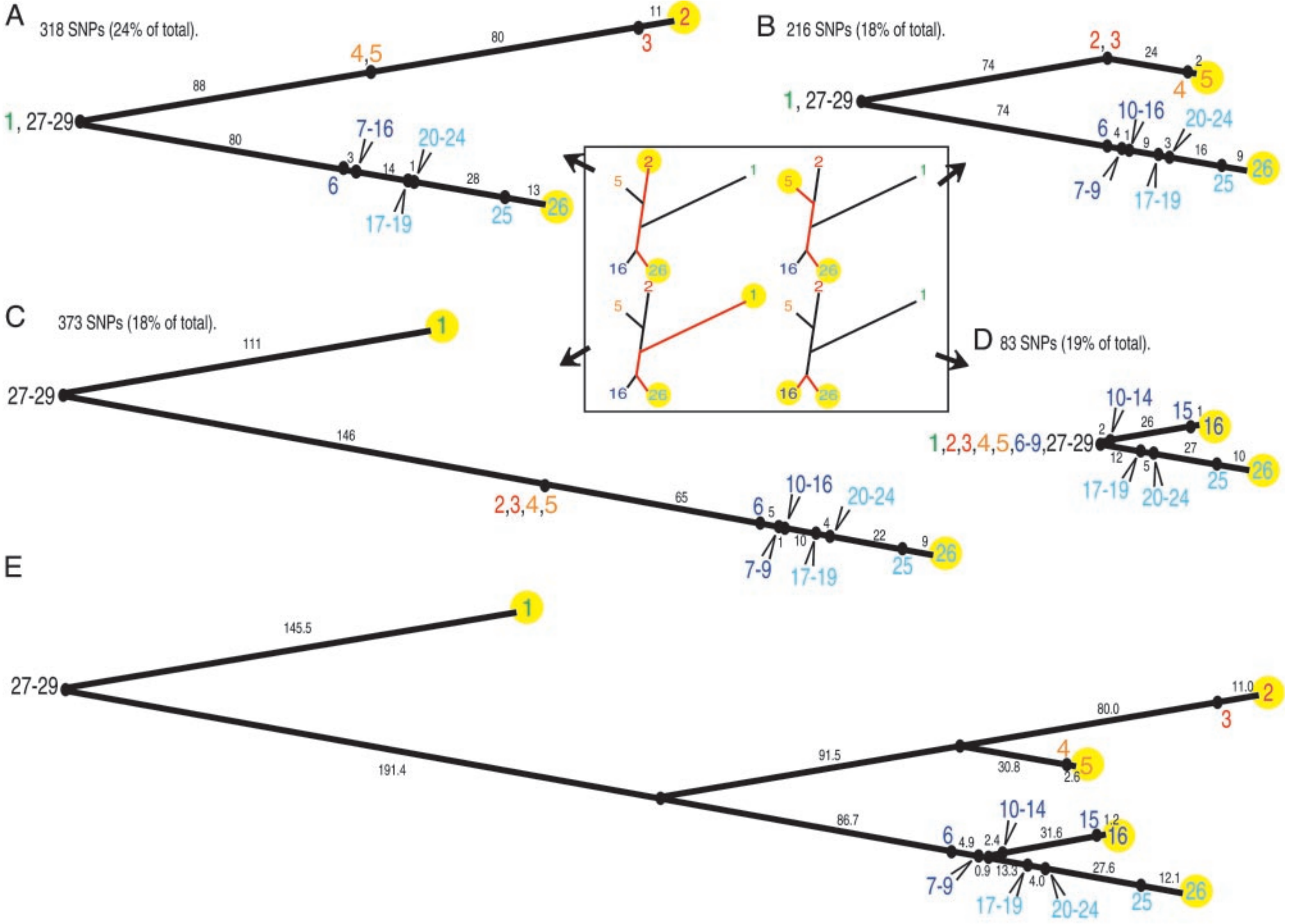


Fig. 6. Four linear phylogenies and a composite dendrogram. Phylogenetic reconstructions using SNP loci discovered between strains 2 and 26 (*A*), 5 and 26 (*B*), 1 and 26 (*C*), and 16 and 26 (*D*) are shown. As predicted, only characters lying on the connecting evolutionary pathway between reference strains (*Inset*; denoted by yellow circles) were discovered. (*E*) A combined tree made by merging the four previous trees and standardizing branch lengths by using weighted average lengths of shared branches (branch lengths were weighted by the percentage of SNPs assayed out of the total number of SNPs discovered). Each phylogenetic tree is rooted with the outgroup (OTUs 27–29). Note the consistent grouping of strains and relative internodal distances across trees. Only one character in this study was homoplastic with one character state that was incompatible with all other loci.

branches (Fig. 6). The importance of these lineages to worldwide anthrax varies greatly, because isolates from the C branch are rare and known from only two cases; the B and A branches represent ≈15% and 85%, respectively, of recent worldwide anthrax cases (4). Although the A and B branches are well known from previous studies, the very unusual C branch isolates have been described only recently from archival collection (A. Hoffmaster, personal communication). The C branch discovery suggests that more diverse *B. anthracis* genotypes may exist in nature but may be rarely seen in anthrax outbreaks. In contrast, the A branch radiation represents the most successful of the major diversity groups, perhaps because of fortuitous global dispersal or an unknown adaptive advantage. A fit genotype can be successful because of either stochastic processes or phenotypic characteristics that create a reproductive advantage. No such biological differences are known among these groups; however, adaptive phenotypes could be subtle or only relevant in a particular environmental context. As such, adaptive phenotypes would be difficult to discern and thus cannot be ruled out as an explanation for the A group success. Molecular clock estimates are problematic in this species because of the difficulty in estimating the number of generations per year, but human activity in the Holocene epoch is likely correlated with the phylogenetically shallow radiation of the A group. The last 10,000 years have been dramatic in human diasporas and the emergence of infectious diseases; fitness of the A group is possibly a consequence.

As previously stated, calibrating absolute evolutionary rates is especially problematic for *B. anthracis*, but the comparison of relative rates among branches is straightforward and very precise in this data set. C branch evolution has occurred at one third the rate seen in the A+B branch, and the B2 branch rate is one third the B1 rate (Fig. 4*E* and Table 4). The variable evolutionary rates may be a reflection of erratic epidemiological patterns associated with anthrax in different ecological situations. Anthrax transmission frequencies can vary dramatically, with large outbreaks being scattered temporally and spatially. Because the frequency of animal transmission greatly influences the evolutionary rates, higher mutation accumulation should be associated with higher transmission frequencies. Host availability or density may be the driving ecological force behind transmission frequency and, thus, evolutionary rates. The B1 branch, for example, is common in southern Africa where there are large herds of susceptible ungulate hosts (11). In contrast, the B2 branch is primarily found in southern France (12) where the hosts may be mostly domesticated animals with many fewer opportunities for infection.

On the other hand, isolates on the A branch may have diverse hosts. For example, the large radiation before the A1+A2 (Fig. 6*E*) bifurcation (indicated by the cluster of nodes) could be linked to an expansion into many novel environments and reflected by the geographic diversity of A branch strains. We know very little about strains for the C branch, but their rarity and relatively slow evolutionary rates suggest low fitness. Indeed, members of the C branch do not possess the pX01 plasmid, which encodes toxin virulence factors, suggesting that the associated loss of virulence may be the causative factor behind the demography and retarded relative rates of evolution. Alternatively, the evolutionary pace of each branch may reflect adaptive differences, such as spore-germination rates, thus balancing selection for long-term survival against prolific reproduction.

A natural analytical outcome of intensive reference-based character discovery in clonal organisms is the generation of linear phylogenetic hypotheses. This method of discovery has complicated several previous studies (e.g., refs. 1 and 13), and its ramifications have not been fully explored. Because only characters on direct-connecting evolutionary paths will be discovered by whole-genome-based comparisons of a few strains, branch collapse should be expected. As shown in this study, phylogenetic hypotheses that depict secondary or tertiary branching are indicative of homoplasy (see Fig. 8). The information gleaned from each comparison depends on reference strains chosen for SNP discovery. When two closely related reference strains are compared, information that separates distantly related strains will not be found (Fig. 6*D*). However, precise locations of the last common ancestor for the reference strains as well as node locations for nested clades will be obtained. Thus, the questions being asked should dictate the strains chosen for comparisons. For resolution of deep phylogenetic patterns, reference strains should be as diverse as possible. Alternatively, for resolution along a specific branch, strains with a connecting evolutionary path that is predicted to lie on the branch of interest should be chosen. In both cases, maximum resolution in the area of interest will be gained by selecting strains with a connecting evolutionary path that bisects the largest number of phylogenetic bifurcation points. Placement of the phylogenetic root can be particularly confusing if the connecting evolutionary path between reference strains does not pass directly through the root (Fig. 6 *A*, *B*, and *D*). In such cases, branch collapse causes the outgroup to be grouped with members of the ingroup. This rooting is highly accurate and represents the position of the last common ancestor to the reference strains. Efficient reference strain selection and correct interpretation of the resulting phylogenetic hypothesis depends on a thorough understanding of the bias involved in character discovery (Fig. 1).

Regardless of the evolutionary age and mutation saturation level, recombination will result in homoplasious characters in a data set. Conceptually, this situation is caused by mixed phylogenetic paths across the genome that result in character-state conflicts in dispersed loci. The strict bifurcating evolutionary model invoked by cladistic analysis is violated, and a more phenetic or statistical approach to classification will be needed. This compromise does not result in accurate phylogenetic reconstruction but rather a "major rule" estimation of the most dominant evolutionary path. However, extremely biased character discovery will not exist to the same extent in organisms with high genetic recombination, because any particular individual is a mosaic of genomic sampling from the entire species. If genomes are continually mixed, combining characters from dispersed genomic regions will increase homoplasy in a data set because of conflicts in the real phylogenetic history of different loci. Because linear phylogenies indicate low homoplasy, which in turn is consistent with short evolutionary histories and a lack of recombination, *B. anthracis* seems to be lacking detectable recombination events across even the most diverse subtypes. In contrast, the much older and more diverse *B. cereus* group has detectable levels of recombination (R. Okinaka, personal communication).

In humans and other sexual populations, molecular characters discovered in autosomal regions will likely be polymorphic in a wide range of populations (14), whereas the mitochondrial and Y chromosome characters are more likely to be fixed because of their clonal propagation. Phylogenetic hypotheses based on subgenomic autosomal regions with great disequilibrium, however, may be impacted by discovery bias. Despite the complications associated with biased discovery and linear phylogenetic hypotheses, the benefits of intensive reference-based discovery are tremendous, allowing for the creation of extremely accurate phylogenies, the likes of which can only be surpassed by entire genome sequencing of all operational taxonomic units.

We thank G. J. Allan, David Wagner, Jonathan Eisen, Mark Miller, and two anonymous referees for critical reviews of the manuscript; Karen Hill and Paul Jackson (Los Alamos National Laboratory, Los Alamos, NM) for *B. cereus* and *B. thuringiensis* strains and associated information; and Alex Hoffmaster (Centers for Disease Control and Prevention, Atlanta) for strain A1055 from the "C" group. We also thank James Schupp and Amy Vogler for technical support and critical reviews of the ideas presented here. This work was supported by National Institute of General Medical Sciences Grants R01-GM060795, National Institute of Allergy and Infectious Disease Contract N01-AI15447, and a grant from the U.S. Department of Homeland Security.

1. 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., *et al*. (2003) *J. Bacteriol*. **185,** 3392–3399.
2. Harrell, L., Andersen, G. & Wilson, K. (1995) *J. Clin. Microbiol*. **33,** 1847–1850.
3. Keim, P., Kalif, A., Schupp, J., Hill, K., Travis, S., Richmond, K., Adair, D., Hugh-Jones, M., Kuske, C. & Jackson, P. (1997) *J. Bacteriol*. **179,** 818–824.
4. Keim, P., Price, L. B., Klevytska, A. M., Smith, K. L., Schupp, J. M., Okinaka, R., Jackson, P. J. & Hugh-Jones, M. E. (2000) *J. Bacteriol*. **182,** 2928–2936.
5. 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., *et al*. (2002) *Science* **296,** 2028–2033.
6. Keim, P., Van Ert, M. N., Pearson, T., Vogler, A. J., Hyunh, L. Y. & Wagner, D. M. (2004) *Infect. Genet. Evol*. **4,** 205–213.
7. Hill, K. K., Ticknor, L. O., Okinaka, R. T., Asay, M., Blair, H., Bliss, K. A., Laker, M., Pardington, P. E., Richardson, A. P., Tonks, M., *et al*. (2004) *Appl. Environ. Microbiol*. **70,** 1068–1080.
8. Swofford, D. (1999) PAUP, Phylogenetic Analysis Using Parsimony (and Other Methods) (Sinauer, Sunderland, MA), Version 4.0 Beta.
9. Nixon, K. C. (2002) WINCLADA (K. C. Nixon, Ithaca, NY), Version 1.00.08.
10. Kumar, S., Tamura, K., Jakobsen, I. & Nei, M. (2001) MEGA, Molecular Evolutionary Genetics Analysis (Pennsylvania State University, University Park, PA), Version 2.1.
11. Smith, K. L., DeVos, V., Bryden, H., Price, L. B., Hugh-Jones, M. E. & Keim, P. (2000) *J. Clin. Microbiol*. **38,** 3780–3784.
12. Fouet, A., Smith, K. L., Keys, C., Vaissaire, J., Le Doujet, C., Levy, M., Mock, M. & Keim, P. (2002) *J. Clin. Microbiol*. **40,** 4732–4734.
13. Gutacker, M. M., Smoot, J. C., Migliaccio, C. A. L., Ricklefs, S. M., Hua, S., Cousins, D. V., Graviss, E. A., Shashkina, E., Kreiswirth, B. N. & Musser, J. M. (2002) *Genetics* **162,** 1533–1543.
14. Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., *et al*. (2001) *Science* **294,** 1719–1723.