Clustering *Acinetobacter* Strains by Optical Mapping

Barry G. Hall[1], Benjamin C. Kirkup[2], Mathew C. Riley[2], and Miriam Barlow[1,3,*]

[1]Bellingham Research Institute, Bellingham, Washington

[2]Walter Reed Army Institute of Research, Silver Spring, Maryland

[3]Molecular and Cell Biology, University of California, Merced

*Corresponding author: E-mail: miriam.barlow@gmail.com.

Accepted: May 23, 2013

Abstract

Optical mapping is a technique that produces an *ordered* restriction map of a bacterial or eukaryotic chromosome. We have developed a new method, the BOP method, to compare experimental optical maps with in silico optical maps of complete genomes to infer the presence/absence of short DNA sequences (bops) in each genome. The BOP method, as implemented by the Optical Mapping suite of four programs, circumvents the necessity of whole-genome multiple alignments and permits reliable strain typing and clustering on the basis of optical maps. We have applied the Optical Mapping Suite to 125 strains of *Acinetobacter* sp., including 11 completely sequenced genomes and 114 *Acinetobacter* complex from three US military hospitals. We found that optical mapping completely resolves all 125 strains. Signal to noise analysis showed that when the 125 strains were considered together almost 1/3 of the experimental fragments were misidentified. We found that the set of 125 genomes could be divided into three clusters, two of which included sequenced genomes. Signal to noise analysis after clustering showed that only 3.5% of the experimental restriction fragments were misidentified. Minimum spanning trees of the two clusters that included sequenced genomes are presented. The programs we have developed provide a more rigorous approach for analyzing optical map data than previously existed.

Key words: genome alignment, epidemiology, optical mapping.

Introduction

Molecular epidemiology involves tracking the spread of pathogens to determine the sources of disease outbreaks and to understand the dynamics of those outbreaks (van Belkum et al. 2001; Hall and Barlow 2006). The ability to follow and characterize outbreaks relies on strain typing (Olive and Bean 1999) and estimating the relationships among isolates by phylogenetic or clustering methods. One problem inherent in these methods is the trade-off between depth of information and having too much information to analyze. Many methods for whole-genome analysis are too computationally intensive for standard desktop computers. Additionally, they may require specialized knowledge of programming languages such as Perl or Python.

Comparisons of the genomes of multiple isolates within several species have led to the concept of the bacterial pangenome. For the majority of bacterial species, there is a set of genes that are present in all members of the species (the core genes) and an additional set of genes that are present in some, but not all, members of the species (the accessory or distributed genes) (Ehrlich et al. 2005; Tettelin et al. 2005; Lindsay et al. 2006; Hiller et al. 2007; Hogg et al. 2007; Lefebure and Stanhope 2007; Willenbrock et al. 2007; Lapierre and Gogarten 2009; Hall et al. 2010; Boissy et al. 2011). The great variability between isolates in the presence/absence of accessory genes has been shown to be a powerful tool for distinguishing among isolates. Analysis of complete genome sequences to determine the presence/absence of accessory genes provides much better resolving power than does MLST (Hall et al. 2010), but even that approach, being limited to accessory coding sequences, does not use all of the information in completely sequenced genomes.

Optical mapping is a powerful technique that is able to capture the presence/absence of accessory geneses without sequencing. The data produced from optical mapping are *ordered* restriction maps of bacterial or eukaryotic chromosomes (Cai et al. 1998; Jing et al. 1998). The restriction sites in those maps can be aligned with whole-genome sequence data to identify the physical location of sequence on a digested

chromosome. Optical maps are useful for generating correct assemblies of the whole genomes and also contain information about the similarities of whole genomes from multiple strains that are optically mapped. We cannot, however, directly compare those optical maps to infer the presence/absence of DNA sequences based on the presence/absence of restriction fragments because the restriction fragments are inherently degenerate in the same sense that the genetic code is degenerate. Just as multiple codons encode the same amino acid, in a set of maps multiple fragments may include the same DNA sequence. Consider homologous portions of two optical maps with restriction fragment lengths (fig. 1).

Direct comparison shows that the two maps have only four fragments in common, the 8, 12, 13 and 34-kb fragments. However, the inference that the sequences in the 20, and in the14- and 6-kb fragments are not shared would be incorrect. Those approximate sequences (not taking into account base substitutions and indels too small to be detected by optical mapping) can be inferred by comparing experimental optical maps with in silico restriction maps of completely sequenced genomes.

If an experimental optical map includes a *particular* restriction fragment, and we know, from an in silico map based on a sequenced genome of the same species, the DNA sequence of the homologous in silico fragment, then we can infer that the optical map includes a homologous sequence. There are, however, two major obstacles to inferring sequences from optical maps:

1. There are likely to be multiple restriction fragments that have identical lengths but are not homologous.
2. Among in silico maps of sequenced genomes, the same fragment may have slightly different lengths as the result of small indels.

To ameliorate the first problem, it is necessary to identify fragments with additional criteria besides length. The ordered arrangement of restriction fragments obtained by optical mapping provides the lengths of the fragments immediately flanking a fragment. This allows each segment to be uniquely identified by its own length and the length of its two neighbors. Although there may be multiple unrelated fragments of the same length, they are unlikely to be flanked by unrelated fragments of the same lengths. For instance, the 12-kb fragment in figure 1 would be named 8-12-20 in map A but be named 8-12-14 in map B. Although information about the flanking fragments is given in this scheme, its only purpose in this is for identifying the middle fragment.

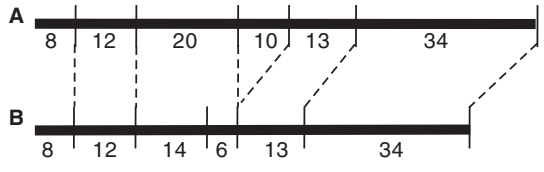

Fig. 1.—Illustration of degeneracy of restriction fragments.

The second problem is ameliorated with a method called "Fuzzy Matching." Optical mapping is inherently noisy, small fragments are detected inefficiently, and fragment sizes are calculated imprecisely (OpGen, Inc., personal communication). With that in mind, it is not practical to match fragment lengths precisely, instead, for fragments less than 20 kb, fragment lengths are considered to match if their lengths differ by less than 0.5 kb, and for fragments ≥20 kb, they are considered to match if their lengths differ by less than 2.5% of the longer fragment.

Because of the degeneracy of restriction fragments, it is necessary to conceptually break each in silico fragment down to short approximately 200-bp subfragments that we call "bops." It is then possible to determine that two strains may indeed have the same bops even though their restriction patterns may be quite different. To infer the presence of a particular bop in a strain that has been experimentally optically mapped, it is only necessary that that the experimental optical map contains a restriction fragment that fuzzy matches a fragment in the in silico map of a sequenced genome. If we can infer the presence/absence of each bop from an experimental optical map, we can then use that information to infer relationships among the optically mapped strains by clustering or to infer phenotypes by comparison with strains whose phenotypes are known (Hall BG, Cardenas H, Barlow M, unpublished data).

In this article, we propose an approach for optical map comparison called "The BOP Method." This method depends on identifying short (~200 bp) homologous sequences (bops) without regard to their positions in the genome. It can be used to estimate relationships among strains by minimum spanning trees (MSTs). We then apply this method to 114 optical maps generated from 114 clinical isolates of *Acinetobacter baumannii*.

Materials and Methods

Strains

Accession numbers of the sequenced genomes are in supplementary table S1, Supplementary Material online. The 114 *Acinetobacter* complex strains that were experimentally optically mapped were isolated from three US military hospitals over a period of 7.5 years.

DNA Sample and Whole-Genome Map Preparation

High-molecular-weight genomic DNA for each reference microbe was prepared directly from isolated colonies or broth culture using the OpGen Sample Preparation Kit (OpGen, Inc., MD) and Agencourt Genfind v2 Kit (Beckman Coulter, FL). In brief, cells were lysed using OpGen lysis buffer and the lysate diluted for direct use. To reduce DNA shearing, wide-bore pipette tips were used, and DNA-containing solutions were

not vortexed. Whole-genome maps were produced using the Argus Whole-Genome Mapping System (OpGen Inc., Gaithersburg, MD). An optimal restriction enzyme for *Acinetobacter* strains, *NcoI*, was chosen using a software program called Enzyme Chooser (OpGen Inc.) that identifies enzymes that result in a 6–12 kb average fragment size and no single restriction fragment larger than 80 kb across the genome.

DNA Sequence Contig Alignment Using MapSolver

DNA sequence contig data in Fasta formatted files were obtained for *A. baumannii* strains. Fasta files were imported into MapSolver software and converted into in silico maps using the same restriction enzyme as was used to generate the respective whole-genome map. DNA sequence contigs were aligned to the whole-genome maps using the sequence placement function of MapSolver, which uses a dynamic programming algorithm that finds the optimal alignment of two restriction maps according to a scoring model that incorporates fragment sizing errors, false and missing cuts, and missing small fragments. Therefore, longer alignments between more similar restriction patterns produced higher scores.

Map Alignment

The MapSolver alignment algorithm used the patterns of fragments and their sizes to generate a final alignment score. In brief, the number of aligned fragments in an alignment depends on the nearness of the match and the MapSolver alignment settings. By default, the initial minimum is determined by the advanced option of "minimum aligned cuts," which was set to 4 and corresponds to three fragments. The other limit is the alignment score, which for every pair of matched fragments can be up to a score of 1 if the fragments match in size perfectly. As the fragment sizes diverge in size, the scoring function awards less of a score.

Initial Clustering with *StructureOptMaps*

StructureOptMaps uses a set of optical maps, both experimental and in silico, to generate a list of the unique restriction fragments contained in the set. It then describes each strain as a binary string that indicates the presence or absence of each of those fragments. It then calls *Structure 2.3.1*, a widely used Bayesian inference population structure program (Pritchard et al. 2000), to cluster strains on the basis of the presence/absence of those restriction fragments. The user specifies a maximum number of clusters, Kmax. ***StructureOptMaps*** assigns individuals to $k = 1, 2, \ldots$ Kmax clusters by a Bayesian algorithm. For each number of clusters, k, a log likelihood (ln L) is reported. The number of clusters with the highest ln L is the most likely number of natural clusters of the data. It is typical to observe that ln L increases sharply with increasing k until a plateau is reached for a few values of k, then starts to decline. ***StructureOptMaps*** automatically selects the number of clusters with the highest ln L. When there is more than one k with very similar ln L, it is more realistic to choose the lower value of k as representing the most likely natural clustering. Indeed, in such circumstances, it is quite common to see that for the higher k, one cluster is empty. ***StructureOptMaps*** allows the user to select a value of k different from the automatically chosen maximum ln L. It also reports for each strain the probability that the strain belongs to each cluster. Strains are assigned to their most probable cluster.

The BOP Method

The BOP method is applied to a set of finished (completely sequenced and closed) genomes as follows: The genomes are digested in silico with one of several restriction enzymes to produce an *ordered* restriction map. Each fragment is given an ID that consists of its length (rounded to the nearest 100 bp) and the lengths of the fragments that flank it. Thus, a 16,542-bp fragment flanked by a 4,320 bp to its left and a 1,680 fragment to its right would be identified as 4.3-16.5-1.7. That particular 16.5-kb fragment is distinguished from all other 16.5-kb fragments by the lengths of its flanking fragments. At this point, that system appears sufficient to uniquely identify restriction fragments. Cases of multiple occurrences of the same fragment turn out to be duplicated regions that contain the same internal restriction fragments. Usually such regions represent multiple copies of mobile elements or phages. The BOP method is implemented with the program ***OptMapsIS***.

OptMapsIS makes a list of all unique restriction fragments; that is, as each of the sequenced genomes is considered, a fragment is added to the list only if it not already in the list. Restriction fragments cannot be used directly to assess genome content because restriction fragments are degenerate; that is, multiple restriction fragments can include the same homologous sequence. For instance, the appearance of a new restriction site destroys an existing restriction fragment and creates in its place two fragments whose lengths sum to the length of the original fragment. To deal with restriction fragment degeneracy, each fragment is divided into approximately 200-bp sections called "bops." Most bops are exactly 200 bp, but bops at the end of a fragment may be less than 200 bp. If a bop is less than 100 bp, it is joined to the previous bop, thus generating a bop of up to 300 bp.

After introducing the restriction sites and creating the bops, the program lists all unique bops. As each bop is considered for addition to the list, it is aligned against each of the bops already in the list by the blast2seq program (National Center for Biotechnology Information). If a bop shares $\geq$80% sequence identity over >50% of its length with a bop already in the list, it is not added to the list. At the same time, lists of each bop in each restriction fragment are maintained. Thus, from the ordered restriction map of a genome, we know

which bops are present. Because we know the sequence of each bop, we know the sequence information that is present in each genome.

After characterizing the sequenced genomes, the restriction fragments defined by experimental optical maps are used to infer the presence/absence of each bop in the experimentally mapped genomes. The presence of a bop from an experimental optical map is determined when the bop is present in a restriction fragment that matches a restriction fragment present in one of the sequenced strains.

Finally, each genome is described by a binary string in which the *i*th character indicates the presence of bop number *i* by a 1 and its absence by a 0. Note that homologous bops (those that share >80% sequence identity) are considered equivalent. Minor variation in sequence is lost to this analysis, as is the position of a sequence in the genome. The binary strings that describe the presence/absence of each bop in each of the genomes are contained in the ***OptMapsIS*** output file with the extension **".scores."**

Fuzzy Matching

Although in silico optical mapping of a sequenced strain is absolutely accurate (or at least as accurate as is the genome sequence), for two reasons experimental optical mapping is not: first, small fragments are detected inefficiently and sized unreliably (OptGen, personal communication). With that in mind, we define *valid* fragments as fragments ≥5 kb. Only valid in silico restriction fragments are recorded and listed, and only valid experimental fragments are considered.

Second, experimental sizing of restriction fragments is far from precise. As a result, it is very unlikely that a restriction fragment, identified by its size and the sizes of its flanking fragments, will precisely match a fragment in the list derived from the sequenced genomes. Furthermore, absolute sizing accuracy varies with the fragment size. Because it is not possible to precisely match experimental fragments with in silico fragments, *fuzzy matching* is required to determine whether an experimental fragment matches an in silico fragment.

Fuzzy matching not only brings with it the possibility of failing to match an experimental fragment with one of the in silico fragments, but it also bring the possibility of incorrectly matching an in silico fragment. When that happens, we incorrectly infer the presence of bops that may not be present in that genome. We consider experimental fragments that correctly match in silico fragments as **"**signal**"** and those that incorrectly match as **"**noise.**"** The signal to noise ratio indicates the degree to which we can trust the inference that a bop is present in an experimentally mapped strain. To estimate the signal to noise ratio, the program ***S2N*** is used to compare the fragment sequences that are inferred to be present by ***OptMapsIS*** with the fragment sequences that are actually present in a sequenced strain.

Efficiency

The efficiency with which restriction fragments in experimental optical maps are found among the in silico optical map fragments in a set of sequenced genomes was determined by the program ***Efficiency***, and the signal to noise ratio (ratio of correctly detected fragments to falsely detected fragments) was determined by ***S2N***.

Some information is lost from an experimental map as the result of experimental fragments that fail to match any of the in silico fragments. That failure can have two sources: 1) the experimental fragment is genuinely not present in any of the sequenced strains or 2) it is present but the size reported falls outside the fuzzy matching criteria. (Only fragments ≥5 kb are considered.) In either case, we are unable to infer the presence of any of the bops that are actually in that fragment.

StructureOptMaps, OptMapsIS, Efficiency, and ***S2N*** for Macintosh OS X are available as part of the **Optical Mapping Suite** at no cost upon request to barryghall@gmail.com.

Estimation of MSTs

MSTs were estimated by ***MSTgold*** as described in Salipante and Hall (2011). ***MSTgold*** for Macintosh is available at no cost at bellinghamresearchinstitute.com/software/.

Results

The 114 *Acinetobacter* sp. Isolates used for this analysis were obtained from three US Army hospitals between March 2003 and October 2010. In figures 3 and 4, isolate IDs are shown in boldface next to nodes that are colored according to the hospital from which the isolate was obtained. The isolation dates, where known, are indicated above and slightly offset from the isolate ID in lightface. To determine the similarity structure of the isolates, we performed the following analysis.

Efficiency and Accuracy

We used the BOP method to analyze and perform clustering analysis on 125 strains of *Acinetobacter* sp., including 11 completely sequenced genomes and 114 *Acinetobacter* strains from three US military hospitals. Approximately 15% of the optical mapping information was lost as the result of discarding small fragments. For the 114 experimental maps, efficiency (the fraction of an experimental optical map that could be matched with fragments from an in silico map of a sequenced genome) ranged from 36.4% to 70.4% with a mean of 48.9 ± 0.6%.

The accuracy with which this process matched an experimental optical map fragments to the correct in silico fragment was then determined using the program S2N. Only one experimental optical map (1311) was available for one sequenced strain, AB0057. For that comparison, the signal to noise ratio was 94:47, meaning that an experimental

fragment was only about twice as likely to be correctly matched as incorrectly matched. A 66.6% level of confidence in inferring the presence of sequences is not sufficient to make a method useful, so we refined our method to increase the signal while decreasing the noise. We found that performing preliminary clustering analyses improve the accuracy of our results by making correct matches between optical mapping data and sequence data more likely.

Preclustering with *StructureOptMaps*

To perform preliminary clustering before implementing the BOP method, we used the following approach: An MST based on the combined experimental and in silico optical maps showed that the sequenced genomes were not well dispersed among the experimental optical maps (result not shown). We reasoned that the more distantly related an experimental genome is to a sequenced genome the more likely is the sequenced genome to include, by chance, a fragment that falsely matches a fragment in the experimental genome. By extension, we reasoned that the signal to noise ratio might be improved if we could subdivide the set of genomes, then investigate each subset on the basis of sequenced genomes contained in that subset. We decided, despite the degeneracy of restriction fragments, to cluster all the genomes on the basis of the presence/absence of restriction fragments. For the 125 *Acinetobacter* sp. Strains, there were 5,145 unique restriction fragments.

There are a number of methods, including the famous K-means algorithm, to assign individuals to a specific number of clusters on the basis of overall similarity. The problem is that we do not know in advance the number of clusters that best reflects the natural clustering of a population of individuals.

Structure (Pritchard et al. 2000) offers two alternatives for the user to consider. If recombination is not allowed, ***Structure*** assumes that each strain belongs to only one cluster. If recombination is allowed, it assumes that a strain may be drawn partially from one cluster and partially from one or more other clusters, that is, that there is exchange between strains of different clusters. Unless the recombination situation is well understood from other sources, it is a good practice to run ***StructureOptMaps*** both with and without recombination and to choose whichever gives the higher maximum ln *L*. In both cases, the most likely number of clusters for the 125 *Acinetobacter* sp. strains was $k = 3$. The log likelihood (ln *L*) of the clustering can be used to assess which clustering approach is preferred; the approach with the higher ln *L* is preferred. With recombination ln *L* was −148,531.5 and without recombination ln *L* was −147,402.7. Further analyses were based on clustering with $k = 3$ and without recombination.

Without recombination, ***StructureOptMaps*** sorted the 125 strains into three clusters (fig. 2). Cluster 1 contained 66 genomes including eight sequenced genome. The eight sequenced genomes contained 3,598 unique *NcoI* fragments

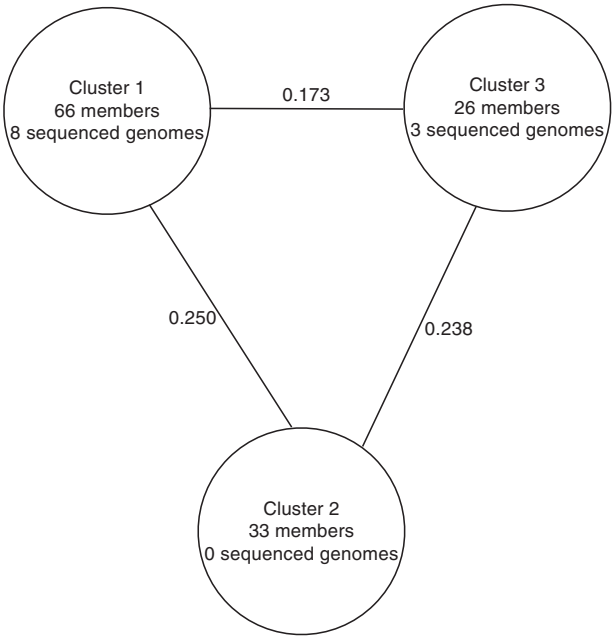

Fig. 2.—Clustering of *Acinetobacter* strains by ***StructureOptMaps***.

that included 39,568 unique bops. Cluster 2 contained 33 genomes but no sequenced genomes. Cluster 3 contained 26 genomes including three sequenced genomes. The three sequenced genomes contained 425 unique *NcoI* fragments that included 18,488 bops.

The mean efficiency (percent of the total genome represented by *NcoI* fragments that matched an in silico fragment) for the 58 experimental genomes in Cluster 1 was $43.8 \pm 0.9\%$, with the range being 32.2–62.4%. That efficiency is slightly worse than when all of the genomes were analyzed together. The mean efficiency for the 23 experimental genomes in Cluster 3 was $44.8 \pm 1.5\%$, with the range being 34.7–63.1%. That efficiency is, again, slightly worse than when all of the genomes were analyzed together.

Cluster 3 includes the sequenced genome AB0057 and its corresponding experimental genome 1311 and thus permits calculation of the signal to noise ratio. For that comparison, the signal to noise ratio was 138:5, meaning that an experimental fragment had a 96.5% probability of being correctly matched with an in silico fragment. That dramatic increase in signal:noise results from preclustering the genomes using ***StructureOptMaps***. Without the availability of an experimental map of a sequenced genome in cluster 1, we are forced to assume that cluster 1 had a similar increase in signal:noise.

Minimum Spanning Trees

MSTs based on the presence or absence of each bop and calculated by ***MSTgold*** are shown in figures 3 and 4. Just as a phylogenetic tree is a graph that illustrates the

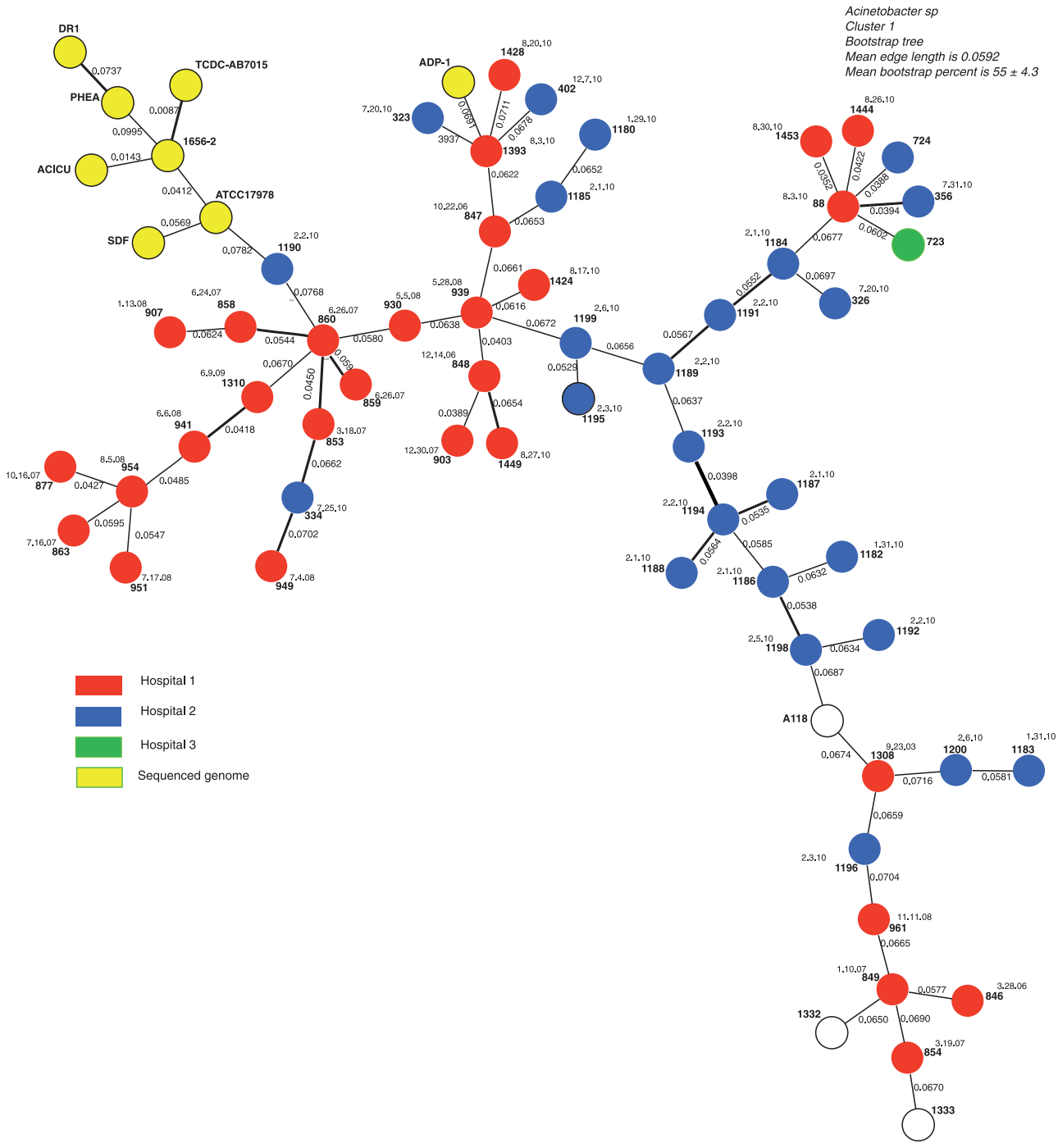

Fig. 3.—MST of cluster 1 from figure 2. Strain IDs are in boldface, and edge lengths are in lightface in a smaller font.

relationships among individuals and their hypothetical ancestors based on identity by descent, an MST is a graph that illustrates the relationships among individuals based on identity by state. On an MST, each node represents an individual, and nodes are connected by edges whose lengths reflect the distance between the nodes. In this case, the distance between a pair of genomes is shown as the number of differences in the state of the bop (0 or 1) divided by the number of bops. A *spanning tree* is a subset of a fully connected graph in which there is a single path from any node to any other node. An MST is the shortest spanning tree of all the possible spanning trees. Depending on the order in which the nodes are considered, it is possible for there to be more than one MST (Salipante and Hall 2011). For Cluster 1 (fig. 3), there

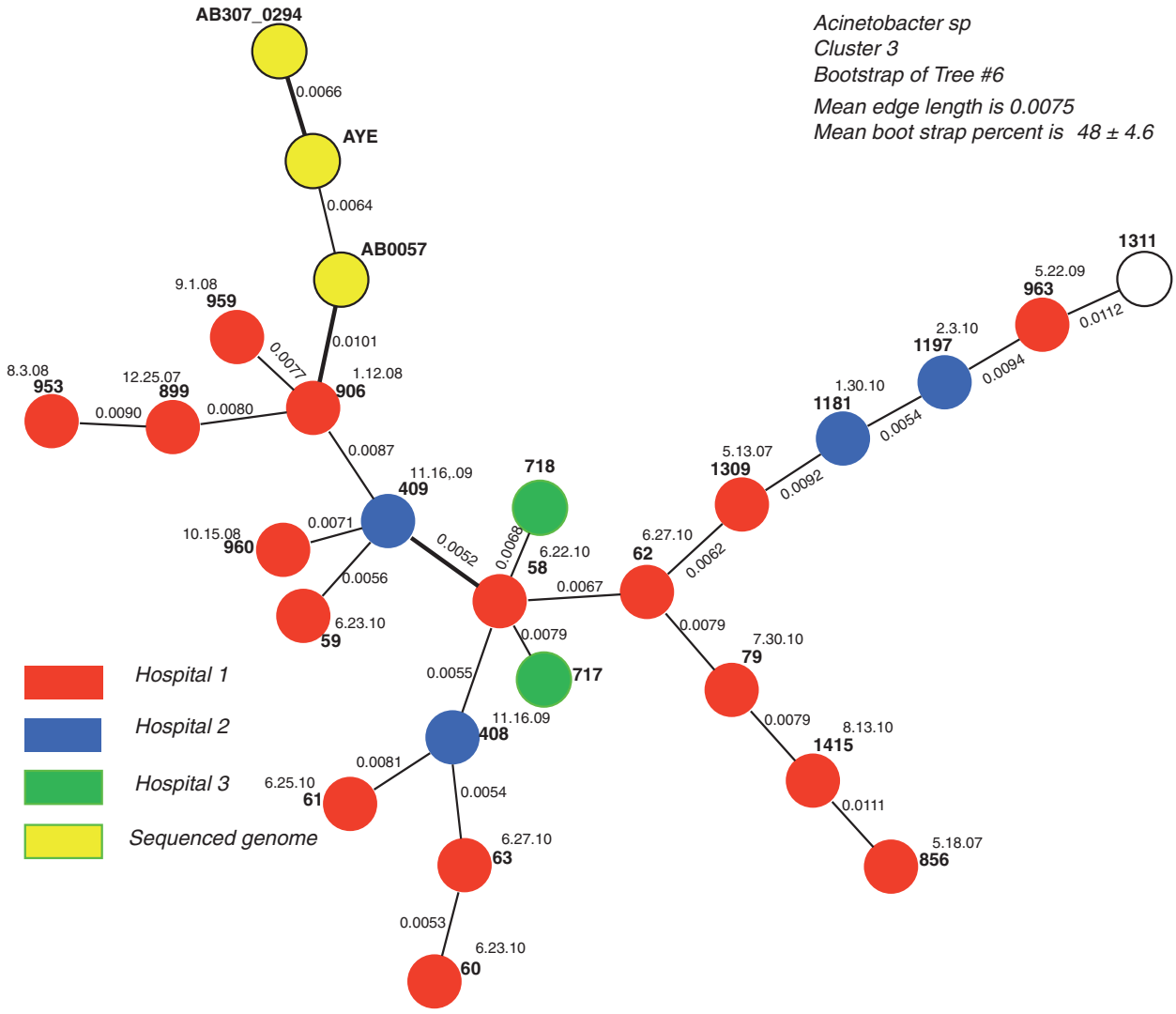

Fig. 4.—MST of cluster 3 from figure 2. Strain IDs are in boldface, and edge lengths are in lightface in a smaller font.

was only one MST, but for Cluster 3, there were 16 MSTs that had the same total length. The reliability of the MSTs was evaluated by bootstrapping with 100 replicates. Cluster 3 tree number 6 had the highest average bootstrap value and is shown in figure 4.

In figures 3 and 4, edges that were present in ≥90% of the bootstrap replicates are drawn as thick lines, and those that were present in less than 90% of the replicates are drawn as thin lines. Although in phylogenetic analyses, bootstrap percentages more than 70% are generally considered to be trustworthy, for MSTs bootstrap probabilities are typically lower and must be considered relative measures of reliability, rather than a hard and fast cutoff of believability (Salipante and Hall 2011). Indeed, the average bootstrap percentages for the MSTs in figures 2 and 3 are typical of the maximum average bootstrap percentages observed by Salipante and Hall (2011) in their study.

To determine the consistency of our results with other accepted methods, we compared our MST with a published *Acinetobacter* UPGMA cladogram (McQueary et al. 2012) that had been assembled from a pairwise distance matrix of optical maps. Although it is impossible to directly compare these methods because UPGMA creates a bifurcating tree and MSTs are presented as clusters, we did find that the groupings of most similar strains was largely consistent between methods.

Discussion

The BOP method as implemented by ***OptMapsIS*** permits the analysis of optical maps for strain typing and for estimating strain relationships by clustering. That implementation requires comparing experimental optical maps with in silico optical maps based on completely sequenced genomes of the

same organism. Signal to noise ratio analysis showed that comparing the undifferentiated set of 114 experimental maps with the 11 in silico maps results in misidentifying one-third of the experimental restriction fragments. Population structure analysis using ***Structure*** showed that the set of 125 genomes could be divided into three clusters, two of which included sequenced genomes (fig. 2). Signal to noise analysis after clustering showed that only 3.5% of the experimental restriction fragments were misidentified, confirming the importance of comparing experimental maps with in silico maps that are as closely related as possible.

MSTs of the genomes in clusters 1 and 3 showed that, although there are obvious instances of migration between hospitals, genomes clustered well according to the hospital of origin and the date of isolation. It is noteworthy that the genomes in Cluster 3 are about seven times more similar to each other than are the genomes in Cluster 1.

In this study, the relationships of only 71% of the experimental genomes could be estimated because Cluster 2 included no sequenced genomes. For those strains whose relationships could be estimated, less than half of the information in each genome could be utilized for those estimates, that is, the average efficiency was about 44%. The efficiency is a function of how well the experimental restriction fragments are represented among the sequenced strains. In these cases, the genotypes of the sequenced strains were not well dispersed among those of the experimental strains. The efficiency could probably be greatly improved by sequencing a few additional strains in each cluster, for example, 1197 and 408 in Cluster 3 and 1198, 88, and 939 in Cluster 1. Similarly, an MST based on the presence/absence of restriction fragments in Cluster 2 could be used to identify 4 or 5 well-dispersed genomes for sequencing. That additional sequencing, although time consuming and expensive, would permit estimating relationships among all the strains and would probably significantly increase efficiency, and thus reliability of the resulting MSTs. Unfortunately, the resources to sequence those genomes in this study are not presently available.

Efficiency and accuracy, in the sense of signal to noise ratio, are also affected by the fuzzy matching algorithm. In this case, we have only one strain, AB0057, that has been sequenced and has also been optically mapped (as 1311). Optical mapping of several other sequenced genomes, particularly in Cluster 1, would permit better evaluation of accuracy.

Recent versions of the optical mapping software from OpGen calculate the standard errors on the size of each fragment. Those standard errors could very possibly allow the development of better fuzzy matching criteria that would improve accuracy and perhaps efficiency as well. Unfortunately, that software cannot be applied retrospectively to the current data sets. The preservation of the original images is data storage intensive and as a mapping service, OpGen Inc. provided only the assembled consensus maps.

These results indicate that optical mapping has the potential to serve as a viable alternative to whole-genome sequencing for elucidating relationships among strains on the basis of almost-complete genome sequence information.

Supplementary Material

Supplementary table S1 is available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by grant number 1R15GM090164-01A1 from the Institute of General Medical Sciences of the National Institutes of Health to M.B. and by the Department of Defense and the Department of the Army to M.C.R. and B.C.K. Material has been reviewed by the Walter Reed Army Institute of Research.

Literature Cited

Boissy R, et al. 2011. Comparative supragenomic analyses among the pathogens *Staphylococcus aureus, Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. BMC Genomics 12:187.

Cai W, et al. 1998. High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. Proc Natl Acad Sci U S A. 95:3390–3395.

Ehrlich GD, Hu FZ, Shen K, Stoodley P, Post JC. 2005. Bacterial plurality as a general mechanism driving persistence in chronic infections. Clin Orthop Relat Res. 20–24.

Hall BG, Barlow M. 2006. Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. Ann Epidemiol. 16:157–169.

Hall BG, Ehrlich GD, Hu FZ. 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156:1060–1068.

Hiller NL, et al. 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol. 189:8186–8195.

Hogg JS, et al. 2007. Characterization and modeling of the *Haemophilus influenzae* core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol. 8: R103.

Jing J, et al. 1998. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. Proc Natl Acad Sci U S A. 95: 8046–8051.

Lapierre P, Gogarten JP. 2009. Estimating the size of the bacterial pan-genome. Trends Genet. 25:107–110.

Lefebure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. Genome Biol. 8:R71.

Lindsay JA, et al. 2006. Microarrays reveal that each of the ten dominant lineages of *Staphylococcus aureus* has a unique combination of surface-associated and regulatory genes. J Bacteriol. 188: 669–676.

McQueary CN, et al. 2012. Extracellular stress and lipopolysaccharide modulate *Acinetobacter baumannii* surface-associated motility. J Microbiol. 50:434–443.

Olive DM, Bean P. 1999. Principles and applications of methods for DNA-based typing of microbial organisms. J Clin Microbiol. 37:1661–1669.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
Salipante SJ, Hall BG. 2011. Inadequacies of minimum spanning trees in molecular epidemiology. J Clin Microbiol. 49:3568–3575.
Tettelin H, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci U S A. 102:13950–13955.
van Belkum A, Struelens M, de Visser A, Verbrugh H, Tibayrenc M. 2001. Role of genomic typing in taxonomy, evolutionary genetics, and microbial epidemiology. Clin Microbiol Rev. 14:547–560.
Willenbrock H, Hallin PF, Wassenaar TM, Ussery DW. 2007. Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. Genome Biol. 8:R267.

Associate editor: Bill Martin