Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing

Barry G. Hall,^{1,2} Garth D. Ehrlich^{2,3} and Fen Z. Hu^{2,3}

Correspondence

Barry G. Hall

barryhall@zeninternet.com

¹Bellingham Research Institute, 218 Chuckanut Point Rd, Bellingham, WA 98229, USA

²Center for Genomic Sciences, Allegheny-Singer Research Institute, 320 East North Ave, Pittsburgh, PA 15212, USA

³Departments of Microbiology and Immunology, and Otolaryngology-Head and Neck Surgery, Drexel University College of Medicine, Allegheny Campus, 320 East North Ave, Pittsburgh, PA 15212, USA

Received 2 October 2009

Revised 8 December 2009

Accepted 16 December 2009

The most widely used DNA-based method for bacterial strain typing, multi-locus sequence typing (MLST), lacks sufficient resolution to distinguish among many bacterial strains within a species. Here, we show that strain typing based on the presence or absence of distributed genes is able to resolve all completely sequenced genomes of six bacterial species. This was accomplished by the development of a clustering method, neighbour grouping, which is completely consistent with the lower-resolution MLST method, but provides far greater resolving power. Because the presence/absence of distributed genes can be determined by low-cost microarray analyses, it offers a practical, high-resolution alternative to MLST that could provide valuable diagnostic and prognostic information for pathogenic bacterial species.

INTRODUCTION

The risk of dying from infectious disease, even in first world countries, has a much higher heritability factor than that associated with any other type of disease, including heart attack and cancer (Sorensen *et al.*, 1988). Moreover, even in the 21st century, epidemic infectious disease is responsible for more mortality and morbidity worldwide than all other disease states combined. In spite of this overwhelming burden to humankind and society, we are still not able to rapidly and accurately distinguish among the many strains and substrains of even the most common pathogenic agents. This is particularly true for many bacterial pathogens, due in part to the fact that they often evolve rapidly through either horizontal gene transfer mechanisms or their ability to rapidly produce a cloud of related organisms through the use of highly mutable genes. Thus, there is an urgent need for improved molecular diagnostics to discriminate among the large numbers of related strains, and for better models for determining the relatedness among strains to aid in understanding the evolution and epidemiology of both

established and emerging pathogens. Together, the development of these tools will aid prognosis, treatment and our ability to track epidemics.

Microbial epidemiologists track the spread of pathogens associated with disease in order to determine the sources of outbreaks and to understand their dynamics (Hall & Barlow, 2006; van Belkum *et al.*, 2001). The ability to accurately characterize and follow epidemics is reliant on strain typing methods, sometimes called subtyping, to distinguish among isolates of the same species, and is usually accomplished using one or more DNA-based methods (Olive & Bean, 1999). The most widely used molecular strain typing method is multi-locus sequence typing (MLST), in which specific segments of seven or more housekeeping genes are sequenced. Each unique sequence of a locus is assigned an allele number, and an allele profile of a strain is defined as the set of allele numbers for that strain. Each unique allele profile is assigned a sequence type (ST) number. Strains that have the same ST number are identical at all of the sequenced loci and are considered to be members of the same clone because they cannot be distinguished from one another. At this time, there are 57 MLST schemes, representing 53 microbial species (<http://pubmlst.org/databases.shtml>). The MLST database spans a range from 12 798 isolates representing 7393 STs (*Neisseria* sp.) down to eight isolates representing eight STs (*Campylobacter helveticus*). MLST, unlike earlier molecular typing methods such as PFGE, is highly reproducible, is well suited for simple representation in databases and is relatively inexpensive.

Abbreviations: ANI, average nucleotide identity; FDG, fraction of distributed genes; MLST, multi-locus sequence typing; NG, neighbour grouping; ST, sequence type.

A supplementary table of the GenBank accession numbers of strains analysed in this study and 13 supplementary tables of the core and distributed gene distances and groupings for *Staphylococcus aureus*, *Escherichia coli*, *Bacillus cereus*, *Streptococcus pyogenes*, *Salmonella enterica* and *Streptococcus pneumoniae* are available with the online version of this paper.

Multiple outbreaks often result from infection by clonally related strains that are descended from a common ancestor and share biochemical and virulence properties. Understanding the dynamics of disease outbreaks requires estimating the relationships among isolates that are identified by strain typing; the most common approach to estimating those relationships is via phylogenetic analysis (Hall & Barlow, 2006; Olive & Bean, 1999).

Phylogenetic analysis is a means of estimating the evolutionary history of a set of taxa (species, genes, individuals etc) that are descended from a common ancestor and depends absolutely on the assumption that the taxa are genetically isolated from one another. When the taxon analysed is a species, the validity of that assumption is implicit in the definition of a biological species. Although it is well understood that there is some genetic exchange among microbial species, the amount of that exchange accounts for only a minor fraction of the variation between species, and molecular sequence-based phylogenetic trees of microbial species are generally robust.

Over the last several years, it has become apparent from MLST studies that many species, including *Neisseria gonorrhoeae* (O'Rourke & Spratt, 1994), *Streptococcus pneumoniae* (Feil *et al.*, 2000), *Streptococcus pyogenes* (Feil & Spratt, 2001), *Helicobacter pylori* (Go *et al.*, 1996) and *Haemophilus influenzae* (Meats *et al.*, 2003), undergo considerable intra-species horizontal genetic exchange. In contrast, in *Escherichia coli* (Feil *et al.*, 2001) and *Staphylococcus aureus* (Feil *et al.*, 2003), genetic exchange was thought to be rare enough to be ignored for phylogenetic purposes, but a more recent study (Wirth *et al.*, 2006) contradicts that view for *E. coli*. In a recent study, Perez-Losada *et al.* (2006) attempted to estimate the population recombination parameters from MLST data for 13 species, and concluded that *H. pylori*, *N. gonorrhoeae* and *S. pneumoniae* populations experience high levels of recombination but that *Bacillus cereus*, *Haem. influenzae*, *Streptococcus agalactiae* and *S. pyogenes* only experience moderate levels of recombination; however, *Vibrio vulnificus*, *Campylobacter jejuni*, *Enterococcus faecium*, *E. coli*, *Staph. aureus* and *Moraxella catarrhalis* experience low levels of recombination. Again, other studies have contradicted some of those assessments (Guttman & Dykhuizen, 1994; Hughes & Friedman, 2005; Wirth *et al.*, 2006). The program ClonalFrame (Didelot & Falush, 2007) is reported to be able to extract sufficient phylogenetic signal from MLST data to permit estimation of phylogenetic relationships among some 58 isolates of various *Bacillus* species. That program is, however, computationally intensive and it is not clear that it would be practical to apply it to several hundred isolates of a single species.

Given the difficulty of assessing the extent of recombination within bacterial species, and in the absence of a clear understanding of how much recombination can be tolerated without obscuring meaningful phylogenetic signal, it is now well accepted that phylogenetic analysis

is generally inappropriate for estimating relationships among isolates within a bacterial species.

The problem of taking recombination into account when assessing relationships within a species has led to an alternative method for estimating relationships from MLST data. Instead of trees, which seek to represent the historical relationships among all isolates on a single diagram, attention has turned to clustering methods that group together the most closely related strains. Such clustering meets the needs of microbial epidemiology for understanding the dynamics of disease outbreaks. One clustering method, the eBURST analysis (Feil *et al.*, 2004), considers all MLST alleles to be equidistant from each other whether they differ by one or several nucleotides. As a consequence, recombination, which may introduce many substitutions from a single event, and mutation, which introduces a single substitution, are given equal weight. The eBURST method does not assume identity by descent; it only assumes, using identity by state, that two strains that share the same allele at several loci are more closely related than those that share the same allele at fewer loci. The word 'related' in this context does not imply descent from a common ancestor, and throughout this paper is simply used as a synonym for 'genetically similar'. A strain is evaluated as being directly related to another strain only if the two are identical at all but one of the loci. A cluster of strains, each of which is directly related to at least one other strain in the cluster, is termed a clonal complex. Unlike a phylogenetic tree, which attempts to link all of the isolates using identity by descent, eBURST links isolates into separate clonal complexes that may consist of as few as two isolates and can leave some isolates unlinked to any clonal complex. Thus, a pair of isolates that differ by a single nucleotide at each of four of seven loci are considered unrelated, while another pair that differ by 4 nt at one of the seven loci are considered directly related. eBURST provides no estimate of the genetic distance between two clonal complexes or between unrelated isolates. eBURST is, however, a powerful and elegant way of representing relationships among fairly closely related isolates and is thus an important tool for microbial epidemiology.

The major limitation of MLST is its inability to resolve many isolates from each other. For instance, the *Staph. aureus* MLST database includes 2425 isolates that fall into 958 STs. However, 142 of those isolates are ST8 and thus are indistinguishable from each other, as are the 120 ST239 isolates. MLST involves sequencing portions of 7–10 housekeeping genes, and thus samples only about 0.1–0.2 % of a microbial genome, so it should not be terribly surprising that 142 isolates are identical when using that fraction of their genomes. The question is whether all, or many, of those isolates are actually different from each other. It seems likely that all ST8 isolates are not, in fact, identical and that higher resolution methods can usefully distinguish them from each other. We suggest that whole-genome information can provide that desired increase in resolution.

Comparisons of multiple whole-genome sequences of *S. agalactiae* ($n=8$) (Tettelin *et al.*, 2005), *Haem. influenzae* ($n=13$) (Hogg *et al.*, 2007) and *S. pneumoniae* ($n=17$) (Hiller *et al.*, 2007) has led to the concepts of the 'pan-genome' and the distributed genome hypothesis (Ehrlich *et al.*, 2005). More recent studies have extended the concept to *E. coli* (Willenbrock *et al.*, 2007), *Staph. aureus* (Lindsay *et al.*, 2006), *S. pyogenes* (Lefebure & Stanhope, 2007), and even to the set of all bacteria (Lapierre & Gogarten, 2009). For each of those species, there is a set of genes that are present in each member 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 distributed genes).

The core genes afford an opportunity to assess clonality on the basis of sequence differences among thousands of genes, rather than on the basis of differences among fragments of 7–10 genes. The distributed genes afford the opportunity to assess clonality on the basis of the presence or absence of individual distributed genes in each isolate.

In this study, we consider six bacterial species for which both extensive MLST databases and the complete genome sequences of at least 10 isolates exist. For each species, we ask whether multiple sequenced genomes have the same MLST sequence type, and if so, whether those MLST-identical strains can be resolved by either core gene or distributed gene differences. We propose a clustering method that permits estimating the relationships among isolates based on whole-genome information. Finally, we suggest a means by which sufficient information about distributed genes can be acquired with an ease and cost comparable to that of MLST to permit practical application of these findings to microbial epidemiology.

METHODS

Accession numbers. The GenBank accession numbers of the complete genome sequences that were analysed in this study are given in Supplementary Table S1, available in Microbiology Online.

MLST analysis. For each species, the gene fragment sequences for each MLST locus were extracted from the genome sequences and the appropriate MLST website was used to identify the allele numbers of the locus sequences and the ST number. In some cases, allele numbers could not be assigned because of ambiguity codes in a sequence, and some novel alleles precluded assigning genomes to known STs. For the majority of genomes, however, sequence types could be assigned.

Analysis of pan-genomes. For each species, we identified orthologous genes by conducting BLAST (Altschul *et al.*, 1990, 1997) searches of each sequence identified as a gene in the GenBank annotations against a database of all such genes. A set of Perl scripts was used to determine which sets of orthologues constituted core genes and distributed genes in all genomes. Genes were classified as orthologous if they shared at least 70 % nt sequence identity over at least 70 % of the length of the query sequence (Hogg *et al.*, 2007). Genes that were present in each of the genomes were classified as core genes, while those that were present in at least one, but not all, genomes were classified as distributed genes.

The core gene distances between all possible pairs of genomes of each species were calculated by a Perl program, C-group, that used the BLAST program Blast2seq to calculate the nucleotide sequence identities between all possible pairs of each set of core orthologues. Those identities were averaged over all core genes and the distance between a pair of genomes was defined as $1 - \text{the average nucleotide identity (ANI)}$ (Snel *et al.*, 2005). A Perl program, S-group, was used to determine for each genome which distributed genes were present (scored as 1) and which were absent (scored as 0) in the genome. All possible pairs of genomes were compared and scored as 1 if a distributed gene was either present in both genomes or absent in each genome, and scored as 0 if the distributed gene was present in one genome but absent in the other genome. The scores for each of the distributed genes were summed and divided by the number of distributed genes to calculate the fraction of distributed genes (FDG) that were the same in both genomes. The distributed gene distance between the two genomes is $1 - \text{FDG}$ (Hiller *et al.*, 2007). Both distance measures are frequently used. A recent review discusses these and other measures of distance and similarity among genomes (Snel *et al.*, 2005).

RESULTS

MLST

MLST-based analyses were able to resolve all 16 *Salmonella enterica*, all 11 *S. pneumoniae* and all 10 of the *B. cereus* strains that have been fully sequenced.

In contrast, of 14 sequenced *Staph. aureus* genomes, three strains (NCTC8325, USA300_FPR3757 and USA300_TCH1516) were ST8, three (Mu3, Mu50 and N315) were ST5, and two (JH1 and JH9) were ST105. It is perhaps not that surprising that there should be so many strains that cannot be distinguished by MLST among the 14 completely sequenced *Staph. aureus* genomes. There are 142 ST8, 137 ST5 and 41 ST105 isolates in the *Staph. aureus* database.

Similarly, among 22 *E. coli* genomes, three strains (APEC01, UT189 and S88) were ST95, two (O157:H7_Sakai and O157:H7_EC4115) were ST11 and two (MG1655 and W3110) were ST10. Among 12 *S. pyogenes* genomes, two (MGAS315 and SST-1) were ST15, two (MGAS9429 and MGAS2096) were ST36 and two (MGAS5005 and M1_GAS) were ST7.

The question, then, is whether pan-genome information using allelic data from all core genes or the presence/absence of distributed genes is able to resolve those strains that MLST cannot.

Properties of the pan-genomes

Table 1 and Fig. 1 describe the detailed properties of the *Staph. aureus* pan-genome, while Table 2 summarizes the pan-genomes of *E. coli*, *S. pyogenes*, *B. cereus*, *S. pneumoniae* and *Salm. enterica*. These properties are consistent with earlier descriptions of the pan-genomes of these and other species (Hiller *et al.*, 2007; Hogg *et al.*, 2007; Lefebure & Stanhope, 2007; Lindsay *et al.*, 2006; Tettelin *et al.*, 2005; Willenbrock *et al.*, 2007).

Table 1. Distributed genes of the *Staph. aureus* pan-genome

The *Staph. aureus* pan-genome includes 4221 genes: 1923 core genes (in all strains, data not shown in table) and 2298 distributed genes. On average, 73 % of a *Staph. aureus* genome is core genes.

Strain	Distributed genes
RF122	652
COL	778
JH1	814
JH9	752
MRSA252	781
MSSA476	683
MW2	695
Mu3	749
Mu50	745
N315	624
NCTC8325	953
USA300_FPR3757	636
USA300_TCH1516	783
Newman	622

Resolution of genomes within a species by allelic analyses of core gene sequences

Table 3 shows the core gene distances among the 14 *Staph. aureus* genomes. All genomes were resolved from each other, i.e. distances were >0 , including those that could not be resolved by MLST. Notice that whereas the average distance between all genome pairs is 0.0073, the distance between genome pairs that could not be resolved by MLST is always ≤ 0.0002 . Similarly, all 22 *E. coli* strains and all 12 *S. pyogenes* strains were resolved and, as was the case for *Staph. aureus*, the distances among genomes of the same ST were much smaller than the average distance among genomes. All 16 *Salm. enterica* strains, all 10 *B. cereus* strains and all 11 *S. pneumoniae* strains were likewise resolved. Tables of the core gene distance matrices, similar to Table 3, are given in Supplementary Tables S2–S13.

Neighbour grouping (NG) based on core gene distances

We propose a clustering approach, NG, that, like eBURST, is based on identity by state but that is suitable for application to thousands of loci. The NG method is a variation on classical hierarchical clustering methods that use single linkage. The distances in Table 3 describe strictly differences of state. Those distance measures make no assumptions about whether the relatedness, in the sense of genetic similarity, is through identity by descent or whether it derives from recombination and/or horizontal gene transfer; they simply provide a metric of the degree of relatedness. While population biologists might have an interest in the historical relationships among groups of related isolates, for epidemiological purposes, it is sufficient to group isolates and perhaps to estimate whether or not any of the groups are related to each other. Note that we do

Fig. 1. Distribution of genes among *Staph. aureus* genomes.

not call such groups 'clonal complexes' because clonality implies identity by descent.

Based upon core gene sequence identity, the mean distance among *Staph. aureus* core genomes is 0.00733 ± 0.00042 (Table 3). NG defines a pair of genomes as valid neighbours if they are significantly more closely related than the average pair of genomes; in the case of *Staph. aureus*, if their distance is less than 0.00691. On that basis, neither RF122 nor MRSA252 has a valid neighbour. A pair of strains are assigned to the same group if they are nearest neighbours of each other. As subsequent strains are sequentially considered, one is assigned to a group only if it is the nearest neighbour of a member of that group, thus all members of a group are nearest neighbours of at least one other genome in that group. On that basis, there are seven groups: group 1, RF122; group 2, COL, Newman; group 3, JH1, JH9; group 4, MRSA252; group 5, MSSA476, MW2; group 6, Mu3, Mu50, N315; and group 7, NCTC8325, USA300_FPR3757, USA300_TCH1516.

We further define a pair of groups as related if at least one member of a group has at least one valid neighbour in the other group. On that basis, group 2 is related to groups 3, 5, 6 and 7; thus, groups 2, 3, 5, 6 and 7 form a complex, but groups 1 and 4 are not related either to that complex or to each other (Fig. 2a).

Supplementary Tables S4–S13 give details of the groupings for *E. coli*, *S. pyogenes*, *B. cereus*, *S. pneumoniae* and *Salm. enterica*, and the groupings for each of the six species are summarized in Table 4.

NG based on core gene sequence differences clusters strains in a way that is consistent with eBURST clustering of

Table 2. Summary of the pan-genomes of six bacterial species

Species	<i>n</i> *	Genes in pan-genome			Fraction of genome that is core genes		Does MLST resolve all genomes?
		Core genes	Distributed genes	Total	Pan-genome	Typical genome	
<i>B. cereus</i>	10	1993	10218	12211	0.1632	0.3842	Yes
<i>E. coli</i>	22	2610	10489	13099	0.1992	0.5569	No
<i>Salm. enterica</i>	16	2055	7911	9966	0.2062	0.4654	Yes
<i>Staph. aureus</i>	14	1923	2298	4221	0.4556	0.7200	No
<i>S. pneumoniae</i>	11	1455	2479	3934	0.4004	0.6843	Yes
<i>S. pyogenes</i>	12	1311	1972	3283	0.3993	0.6941	No

**n*, the number of completely sequenced genomes that were analysed.

MLST data for the same strains, while permitting far greater resolution among strains than that provided by MLST. NG also permits estimation of deeper relationships than eBURST analysis, by clustering related groups into complexes. Despite the higher resolution afforded by analysis of core gene sequences, the cost, time and effort required for whole genome sequencing precludes practical application of this approach to microbial epidemiology.

NG based on the presence or absence of distributed genes

While it is both expensive and time consuming to determine the sequences of the core genes of a strain, it is relatively inexpensive and quick to determine the

presence or absence of each of the distributed genes in the pan-genome of a species by performing comparative genome hybridization to microarrays imprinted with each of the distributed genes. Thus, we determined for each of the six species identified above, whether the presence or absence of distributed genes provides as much resolution among strains as core gene sequences and, if so, whether clustering by NG based on distributed genes is consistent with both MLST clustering by eBURST and with NG clustering based on core gene sequences.

Table 3 also shows the distributed gene distances for *Staph. aureus*. These were calculated on the basis of the presence or absence of each of the distributed genes from the *Staph. aureus* pan-genome in each genome. The distance between two genomes is defined as the fraction of distributed genes

Table 3. Distances between *Staph. aureus* strains

Above the diagonal: core gene distances expressed as $1 - \text{ANI}$ in core genes. The mean $\pm$ SEM core gene distance between genomes is 0.00733 ± 0.00042 . The maximum core gene distance between valid neighbours is 0.00691. Below the diagonal: distributed gene distances expressed as the fraction of the distributed genes in the pan-genome at which the two strains differ with respect to presence or absence of the gene. The mean $\pm$ SEM distributed gene distance between genomes is 0.28418 ± 0.00585 . The maximum distributed gene distance between valid neighbours is 0.27833.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. RF122	0	0.0141	0.0141	0.0141	0.0161	0.0139	0.0139	0.0141	0.0141	0.0141	0.0141	0.0141	0.0141	0.0141
2. COL	0.358	0	0.0063	0.0063	0.0172	0.0054	0.0054	0.0063	0.0063	0.0063	0.0002	0.0002	0.0002	0.0001
3. JH1	0.3371	0.3054	0	1×10^{-5}	0.0171	0.0062	0.0062	0.0002	0.0002	0.0001	0.0063	0.0063	0.0063	0.0063
4. JH9	0.3235	0.2963	0.0553	0	0.0171	0.0062	0.0062	0.0002	0.0002	0.0001	0.0063	0.0064	0.0063	0.0063
5. MRSA252	0.3774	0.3675	0.3058	0.3022	0	0.0171	0.0171	0.0171	0.0171	0.0171	0.0171	0.0172	0.0172	0.0171
6. MSSA476	0.3086	0.2805	0.266	0.2605	0.2628	0	0.0001	0.0062	0.0062	0.0062	0.0054	0.0054	0.0054	0.0054
7. MW2	0.314	0.2886	0.2768	0.2678	0.2918	0.0942	0	0.0062	0.0062	0.0061	0.0054	0.0054	0.0054	0.0054
8. Mu3	0.3249	0.3493	0.208	0.1908	0.3172	0.2818	0.2728	0	6×10^{-5}	0.0001	0.0063	0.0063	0.0063	0.0063
9. Mu50	0.324	0.3512	0.2107	0.1935	0.3172	0.2827	0.2737	0.0045	0	0.0001	0.0063	0.0064	0.0063	0.0063
10. N315	0.2891	0.3063	0.1876	0.1749	0.2587	0.2361	0.2243	0.0847	0.0856	0	0.0063	0.0063	0.0063	0.0063
11. NCTC8325	0.4173	0.2442	0.3276	0.3276	0.4341	0.3099	0.3172	0.396	0.396	0.3729	0	0.0002	0.0002	0.0002
12. USA300 FPR3757	0.338	0.2021	0.2655	0.2411	0.304	0.2333	0.2261	0.2968	0.2968	0.2447	0.2895	0	2×10^{-5}	0.0002
13. USA300 TCH1516	0.4599	0.256	0.3548	0.3521	0.4005	0.328	0.3335	0.4024	0.4042	0.3502	0.309	0.1953	0	0.0002
14. Newman	0.3036	0.2583	0.28	0.2619	0.3448	0.2605	0.2633	0.2705	0.2723	0.2383	0.2778	0.1994	0.3131	0

Fig. 2. The relationships among *Staph. aureus* genomes estimated by the NG method. The members of a group are enclosed in a box. Solid arrows are drawn from a genome to its nearest neighbour, distances are indicated next to the arrows. The distance units are as described in Methods and elsewhere in the text. Dashed lines connect members of different groups that are valid neighbours, associating those groups into complexes. Only a single dashed line is drawn between any two groups. (a) Groupings estimated from distances calculated from core gene similarities. (b) Groupings estimated from distances calculated from the presence or absence of distributed genes.

from the pan-genome in which the two genomes differ with respect to the presence or absence of a distributed gene. Thus, for *Staph. aureus*, whose pan-genome includes 2298 distributed genes, a pair of strains that differ in the presence/absence of 150 distributed genes would have a distance of 0.065.

Table 3 shows that all 14 *Staph. aureus* genomes were resolved on the basis of distributed gene distances. The NG groups are very similar to those estimated on the basis of core gene sequences and are consistent with MLST eBURST clustering. There are five groups: group 1, RF122; group 2, COL, USA300_FPR3757, NCTC8325, USA300_TCH1516, Newman; group 3, JH1, JH9; group 4, MRSA252, N315, Mu3, Mu50; and group 5, MSSA476, MW2. The differences between clustering on the basis of core gene sequences and on the basis of the presence or absence of distributed genes is that the latter combines the core gene groups 2 and 7 into a single group and it includes MRSA252 in the group with N315, Mu3 and Mu50. Groups 2–5 are linked into a single complex (Fig. 2b). Similarly, all genomes of the other five species under consideration were resolved by distributed gene distances (Supplementary Tables S2–S13).

In an NG analysis, each pair of genomes has one of two relationships: they are in the same group or they are in different groups. Two groupings can be compared to

determine what fraction of the relationships are the same in the two analyses. That fraction is the similarity of the groupings. For *Staph. aureus*, the grouping similarity was 0.857. Table 4 shows the core and distributed gene groupings, and the similarities of those groupings for all six species.

DISCUSSION

Whole genome sequences provide two distinct ways to distinguish, or type, strains within a species: on the basis of core gene similarities and on the basis of the presence or absence of distributed genes. Both approaches sample an enormously greater proportion of the genome than MLST. Core genes sample a greater fraction of the genome (56 % for *E. coli* and 72 % for *Staph. aureus*) than do distributed genes, but the degree of variation among the core genomes is much less than in the smaller fraction of the genome sampled by distributed genes. For instance, *Staph. aureus* strains JH1 and JH9 differ at 10^{-5} of the bases, or about 20 bp, in their core genes, but they differ at 0.0553 of the distributed genes in the *Staph. aureus* pan-genome, i.e. in the presence or absence of 125 genes. Similarly, *E. coli* K-12 strains MG1655 and W3110 differ at only 1.23×10^{-6} of the base pairs in their 2610 core genes, or at about 4 bp, but they differ in the presence of 0.0195 of the 10 489

Table 4. NG results for *B. cereus*, *E. coli*, *Salm. enterica*, *Staph. aureus*, *S. pneumoniae* and *S. pyogenes*

Groups are enclosed in parentheses and complexes of groups are enclosed in square brackets. Strains with the same MLST ST are underlined.

Species	Core gene groupings	Distributed gene groupings	Grouping similarity
<i>B. cereus</i>	[(03BB102, AH820, E33L), (AH187, ATCC10987)], (ATCC14579, B4264, G9842), (NVH391_98)	[(03BB102, AH820, E33L), (AH187, ATCC10987), (ATCC14579, B4264, G9842)], (NVH391_98)	1.0
<i>E. coli</i>	[(K12_MG1655, K12_W3110, K12_DH10B), (ATCC8739, HS), (O157:H7_EDL933, <u>O157:H7_Sakai</u> , <u>O157:H7_EC4115</u>), (E24377A, IAI1, SE11, 55989)], (CFT073, <u>UTI89</u> , 536, <u>APECO1</u> , E2348/69, <u>S88</u> , ED1a), (SMS_3_5, IAI39), (UMN026)	[(K12_MG1655, K12_W3110, ATCC8739, HS, SMS_3_5, K12_DH10B, IAI39, E2348_69), (CFT073, <u>UTI89</u> , 536, <u>S88</u> , <u>APECO1</u> , ED1a), (O157:H7_EDL933, <u>O157:H7_Sakai</u> , <u>O157:H7_EC4115</u>), (E24377A, SE11 IAI1, UMN026, 55989)]	0.8442
<i>Salm. enterica</i>	(Arizonae_62:z4, z23:-), [(Choleraesuis_SC_B67, Paratyphi_C_RKS4594), (Heidelberg_SL476, Typhimurium_LT2, Newport_SL254, Paratyphi_B_SPB7, Agona_SL483, Schwarzengrund_CVM19633), (Paratyphi_A_ATCC9150, Paratyphi_A_AKU_12601), (Typhi_CT18, Typhi_Ty2), (Dublin_CT_02021853, Enteritidis_P125109, Gallinarum_287_91)]	(Arizonae_62:z4,z23:-), [(Choleraesuis_SC_B67, Paratyphi_C_RKS4594), (Heidelberg_SL476, Newport_SL254, Agona_SL483, Schwarzengrund_CVM19633), (Typhi_CT18, Typhi_Ty2, Paratyphi_A_AKU_12601), (Typhimurium_LT2, Enteritidis_P125109, Dublin_CT_02021853, Gallinarum_287_91)], (Paratyphi_A_ATCC9150), (Paratyphi_B_SPB7)	0.9083
<i>Staph. aureus</i>	(RF122), [(COL, Newman), (JH1, JH9), (MSSA476, MW2), (<u>Mu3</u> , <u>Mu50</u> , <u>N315</u>), (<u>USA300_FPR3757</u> , <u>USA300_TCH1516</u> , <u>NCTC8325</u>)], (MRSA252)	(RF122), [(COL, Newman, <u>USA300_FPR3757</u> , <u>USA300_TCH1516</u> , <u>NCTC8325</u>), (JH1, JH9), (MRSA252, <u>N315</u> , <u>Mu3</u> , <u>Mu50</u>), (MSSA476, MW2)]	0.8352
<i>S. pneumoniae</i>	[(70585, P1031), (ATCC700669, JJA, CGSP14), (D39, R6, G54, TIGR4)], (Hungary19A_6), (Taiwan19F_14)	[(70585, D39, CGSP14, R6, TIGR4, Taiwan19F_14, P1031, G54) (ATCC700669, JJA)], (Hungary19A_6)	0.6545
<i>S. pyogenes</i>	[(MGAS315, SSI-1), (MGAS6180, MGAS2096, MGAS9429)], (Manfredo, MGAS8232, MGAS10394), (MGAS10270), (MGAS10750), (<u>MGAS5005</u> , <u>M1_GAS</u>)	[(MGAS315, SSI-1), (Manfredo, MGAS8232, MGAS10394), (MGAS6180, MGAS10270, MGAS10750), (<u>MGAS9429</u> , <u>MGAS2096</u>), (<u>MGAS5005</u> , <u>M1_GAS</u>)]	0.9242

distributed genes in the *E. coli* pan-genome, i.e. in 205 distributed genes. Thus, microarrays imprinted with all of the genes from a species' distributed genome (Tettelin *et al.*, 2005; Ehrlich *et al.*, 2005; Hogg *et al.*, 2007; Hiller *et al.*, 2007) can be used to obtain accurate typing information that can be used in clinical diagnostics and for epidemiological studies. This type of approach was used when the data from a seven strain *Staph. aureus* pan-genome was used to build a microarray to estimate a phylogenetic tree of 161 *Staph. aureus* strains (Lindsay *et al.*, 2006), and a similar study has been conducted in *E. coli* (Willenbrock *et al.*, 2007). The distributed genes of both species, some 10–15 % of which are associated with extra-chromosomal elements, suggest considerable recombination, thus we doubt that there is sufficient phylogenetic signal to permit estimating a valid phylogenetic tree of either species. Of course, phylogenetic software will estimate phylogenetic trees whether or not those trees are valid. Nevertheless, both studies indicate that isolates can be distinguished on the basis of the presence or absence of genes.

It might be argued that such a high level of strain typing resolution is not only unnecessary but also counter productive. According to that view, perfect resolution (each strain is distinct) is equally as uninformative as

zero resolution. That would be true only if we could not estimate the relationships among those 'perfectly resolved' isolates. The NG clustering method provides a convenient and reasonable way to estimate those relationships and allows high-level typing resolution to be useful.

NG provides an alternative to phylogenetic analysis that, like MLST and eBURST analysis, depends not on identity by descent but only upon identity by state (genome similarity) for estimating relationships among strains. Because NG samples a much higher fraction of the genome than MLST, it provides much better resolution than MLST. Comparing the NG method based on core genes alleles to MLST, it is hardly surprising that the complete sequences of thousands of genes resolve strains that are irresolvable by the partial sequences of seven genes. It might be argued that the resolution achieved from the sequences of core genes is almost too high. For instance, the *E. coli* K12 strains MG1655 and W3110 differ at only 1.23×10^{-6} of the bases in their core genes; i.e. at only about 4 bases among those genes. Indeed, in private conversations, some have argued that two strains that are that similar should be considered, for epidemiological purposes, to be members of the same clone. However, our distributed gene-based

NG analyses show that these strains are significantly different on a gene possession basis, therefore, it is likely that if the core NG method finds a difference, it is real and important clinically and epidemiologically. Thus, with the high resolution that comes with whole genome analyses, it is essential to have a clustering method available that shows which strains are most closely related to each other and how close those relationships are.

Our study of six bacterial species shows that core gene sequences provide much higher resolution than MLST, in all cases distinguishing among strains that have the same MLST ST. NG analyses are entirely consistent with MLST clustering by eBURST, yet they provide additional discriminatory power. Despite the rapid development in DNA sequencing technologies, whole-genome sequencing remains far too costly to be used in routine epidemiological studies or in many academic studies of microbial population structure. For that reason, strain typing and estimating relatedness based on core gene sequences is not yet practical. However, it is both practical and cost effective to use comparative genome hybridization to microarrays of distributed gene probes both to type strains and to estimate relationships among strains by the NG method. Our results show that for all six species, strain typing by the presence or absence of distributed genes affords even higher resolution than typing by sequencing the core genes.

NG by distributed genes is completely consistent with MLST grouping by eBURST, and is also highly consistent with NG by core gene sequences (Table 4). If a pair of strains that differ at only 4 bp can be distinguished by the presence or absence of 205 distributed genes, it seems likely that distributed gene microarrays will be able to distinguish virtually all isolates. At this time, the only advantages of MLST are lower cost and the large number of strains in some MLST databases. When distributed gene arrays are widely available, cost is likely to decrease dramatically, and strain typing on the basis of pan-genome distributed genes will offer a high-resolution alternative to MLST and other DNA-based typing methods; as the databases of distributed gene types strains grow, their value will surpass that of MLST for epidemiological purposes. It is our hope, indeed our expectation, that as large numbers of strains are analysed by distributed gene arrays, correlations will be found between the presence/absence of certain genes and clinically important phenotypes such as virulence and antibiotic resistance etc. Determination of those correlations is expected to permit development of specialized microarrays that can assess many phenotypes simultaneously and inexpensively.

ACKNOWLEDGEMENTS

This work was supported by the Bellingham Research Institute, Allegheny General Hospital, Allegheny-Singer Research Institute, grants from the National Institute on Deafness and Other Communication Disorders DC 05659 (G.D.E.), DC 02148 (G.D.E.), DC 04173 (G.D.E.), and the Health Resources and

Services Administration of the Department of Health and Human Services (G.D.E.). We thank Joshua Earl for assistance with MLST.

REFERENCES

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**, 403–410.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.
- Didelot, X. & Falush, D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266.
- Ehrlich, G. D., Hu, F. Z., Shen, K., Stoodley, P. & Post, J. C. (2005). Bacterial plurality as a general mechanism driving persistence in chronic infections. *Clin Orthop Relat Res* 20–24.
- Feil, E. J. & Spratt, B. G. (2001). Recombination and the population structures of bacterial pathogens. *Annu Rev Microbiol* **55**, 561–590.
- Feil, E. J., Smith, J. M., Enright, M. C. & Spratt, B. G. (2000). Estimating recombinational parameters in *Streptococcus pneumoniae* from multilocus sequence typing data. *Genetics* **154**, 1439–1450.
- Feil, E. J., Holmes, E. C., Bessen, D. E., Chan, M. S., Day, N. P., Enright, M. C., Goldstein, R., Hood, D. W., Kalia, A. & other authors (2001). Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc Natl Acad Sci U S A* **98**, 182–187.
- Feil, E. J., Cooper, J. E., Grundmann, H., Robinson, D. A., Enright, M. C., Berendt, T., Peacock, S. J., Smith, J. M., Murphy, M. & other authors (2003). How clonal is *Staphylococcus aureus*? *J Bacteriol* **185**, 3307–3316.
- Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol* **186**, 1518–1530.
- Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. (1996). Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J Bacteriol* **178**, 3934–3938.
- Guttman, D. S. & Dykhuizen, D. E. (1994). Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383.
- Hall, B. G. & Barlow, M. (2006). Phylogenetic analysis as a tool in molecular epidemiology of infectious diseases. *Ann Epidemiol* **16**, 157–169.
- Hiller, N. L., Janto, B., Hogg, J. S., Boissy, R., Yu, S., Powell, E., Keefe, R., Ehrlich, N. E., Shen, K. & other authors (2007). Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J Bacteriol* **189**, 8186–8195.
- Hogg, J. S., Hu, F. Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J. C. & Ehrlich, G. D. (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.
- Hughes, A. L. & Friedman, R. (2005). Nucleotide substitution and recombination at orthologous loci in *Staphylococcus aureus*. *J Bacteriol* **187**, 2698–2704.
- Lapierre, P. & Gogarten, J. P. (2009). Estimating the size of the bacterial pan-genome. *Trends Genet* **25**, 107–110.

- Lefebure, T. & Stanhope, M. J. (2007).** Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol* **8**, R71.
- Lindsay, J. A., Moore, C. E., Day, N. P., Peacock, S. J., Witney, A. A., Stabler, R. A., Husain, S. E., Butcher, P. D. & Hinds, J. (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.
- Meats, E., Feil, E. J., Stringer, S., Cody, A. J., Goldstein, R., Kroll, J. S., Popovic, T. & Spratt, B. G. (2003).** Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. *J Clin Microbiol* **41**, 1623–1636.
- Olive, D. M. & Bean, P. (1999).** Principles and applications of methods for DNA-based typing of microbial organisms. *J Clin Microbiol* **37**, 1661–1669.
- O'Rourke, M. & Spratt, B. G. (1994).** Further evidence for the non-clonal population structure of *Neisseria gonorrhoeae*: extensive genetic diversity within isolates of the same electrophoretic type. *Microbiology* **140**, 1285–1290.
- Perez-Losada, M., Browne, E. B., Madsen, A., Wirth, T., Viscidi, R. P. & Crandall, K. A. (2006).** Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. *Infect Genet Evol* **6**, 97–112.
- Snel, B., Huynen, M. A. & Dutilh, B. E. (2005).** Genome trees and the nature of genome evolution. *Annu Rev Microbiol* **59**, 191–209.
- Sorensen, T. I., Nielsen, G. G., Andersen, P. K. & Teasdale, T. W. (1988).** Genetic and environmental influences on premature death in adult adoptees. *N Engl J Med* **318**, 727–732.
- Tettelin, H., Massignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L. & other authors (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, P. F., Wassenaar, T. M. & Ussery, D. W. (2007).** Characterization of probiotic *Escherichia coli* isolates with a novel pan-genome microarray. *Genome Biol* **8**, R267.
- Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L. H., Karch, H., Reeves, P. R., Maiden, M. C. & other authors (2006).** Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60**, 1136–1151.

Edited by: D. W. Ussery