Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome"

Hervé Tettelin[a,b], Vega Masignani[b,c], Michael J. Cieslewicz[b,d,e], Claudio Donati[c], Duccio Medini[c], Naomi L. Ward[a,f], Samuel V. Angiuoli[a], Jonathan Crabtree[a], Amanda L. Jones[g], A. Scott Durkin[a], Robert T. DeBoy[a], Tanja M. Davidsen[a], Marirosa Mora[c], Maria Scarselli[c], Immaculada Margarit y Ros[c], Jeremy D. Peterson[a], Christopher R. Hauser[a], Jaideep P. Sundaram[a], William C. Nelson[a], Ramana Madupu[a], Lauren M. Brinkac[a], Robert J. Dodson[a], Mary J. Rosovitz[a], Steven A. Sullivan[a], Sean C. Daugherty[a], Daniel H. Haft[a], Jeremy Selengut[a], Michelle L. Gwinn[a], Liwei Zhou[a], Nikhat Zafar[a], Hoda Khouri[a], Diana Radune[a], George Dimitrov[a], Kisha Watkins[a], Kevin J. B. O'Connor[h], Shannon Smith[i], Teresa R. Utterback[i], Owen White[a], Craig E. Rubens[g], Guido Grandi[c], Lawrence C. Madoff[e,j], Dennis L. Kasper[e,j], John L. Telford[c], Michael R. Wessels[d,e], Rino Rappuoli[c,k,l], and Claire M. Fraser[a,b,k,m]

In 1987, it was proposed (1) that bacterial strains showing 70% DNA DNA reassociation and sharing characteristic phenotypic traits should be considered to be strains of the same species.

Thus far, the genome sequence of one or two strains for each species has provided unprecedented information; however, the question of how many genomes are necessary to fully describe a bacterial species has yet to be asked.

Comparative analysis of the six newly sequenced genomes and the two genomes already available in the databases suggests that a bacterial species can be described by its "pan-genome" (pan, from the Greek word , meaning whole), which includes a core genome containing genes present in all strains and a dispensable genome composed of genes absent from one or more strains and genes that are unique to each strain. Surprisingly, unique genes were still detected after eight genomes were sequenced, and mathematical extrapolation predicts that new genes will still be found after sequencing many more strains.

Methods

- Sequence
- Assemble
- Annotate
- Compare
- Modeling

Sequencing, Annotation, and Unfinished Genomes. Genome sequences were generated by the whole-genome shotgun sequencing approach (13, 14). Draft genomes were sequenced to 8×-sequence coverage, and the sequences were assembled by using the Celera Assembler (Celera Genomics, Rockville, MD) (15). Contigs were ordered and oriented according to their alignment to strain 2603V/R by using PROMER (16). Ordered matching contigs were pasted together into a pseudochromosome, and nonmatching contigs were tacked on the end in random order. In the pseudochromosome, contigs were separated by the sequence NNNNNCATTCCATTCATTAATTAATTAATGAATGAATGNNNNN, which (*i*) generates a stop codon in all six reading frames so that no gene is predicted across junctions and (*ii*) provides a start site in all frames, pointing toward contigs to predict incomplete genes at their extremities. ORFs were predicted and annotated by using an automated pipeline that combines GLIMMER gene prediction (17, 18), ORF and non-ORF feature identification, and assignment of functional role categories to genes (14). Assembly of strain 18RS21 resulted in a higher number of contigs than for the other unfinished genomes, leading to the prediction of >3,500 genes. Many small contigs did not harbor protein-coding genes, and several were fragments of rRNAs or coded for tRNAs or structural RNAs.

Shared and Strain-Specific Genes. Each strain pair was compared by means of the following: (*i*) a Smith and Waterman protein search on all of the predicted proteins by using the SSEARCH program (version 3.4) (19, 20); (*ii*) a DNA search of all of the predicted ORFs of a strain against the complete DNA sequence of the other strain, by using the FASTA program (version 3.4) (20); and (*iii*) a translated protein search of all of the predicted proteins of a strain against the complete DNA sequence of the other strain, by using the TFASTY program (version 3.4) (20). A gene was considered conserved if at least one of these three methods produced an alignment with a minimum of 50% sequence conservation over 50% of the protein/gene length.

Core-Genome and Pan-Genome Extrapolation. The number of genes shared by all GBS isolates and the number of strain-specific genes depend on how many strains are taken into account. The sequential inclusion of up to eight strains was simulated in all possible combinations. The number (N) of independent measurements of the shared (see Fig. 2) and strain-specific genes (see Fig. 3) present in the nth genome is $N = 8!/[(n - 1)!\cdot(8 - n)!]$. The size of the species core genome and the number of strain-specific genes for a large number of sequenced strains were extrapolated by fitting the exponential decaying functions $F_c = \kappa_c \exp[-n/\tau_c] + \Omega$ and $F_s = \kappa_s \exp[-n/\tau_s] + tg(\theta)$, respectively, to the amount of conserved genes (see Fig. 2) and of strain-specific genes (see Fig. 3), where n is the number of sequenced strains and κ_c, κ_s, τ_c, τ_s, Ω, and $tg(\theta)$ are free parameters. $tg(\theta)$ represents the extrapolated rate of growth of the pan-genome size, $P(n)$, as a greater number of independent GBS strain sequences become available, i.e.,

$$\lim_{n\to\infty}[P(n)] \approx tg(\theta)\cdot n.$$

The *Inset* of Fig. 3 displays the measured size of the pan-genome as a function of n [in this case, $N = 8!/(8 - n)!$; points are obtained for each value of n] together with a plot of the calculated $P(n)$ (see *Supporting Text*, which is published as supporting information on the PNAS web site).

Synteny. Paralog clusters in each genome were generated by using the Jaccard algorithm (21), with ≥80% identity, and a Jaccard coefficient ≥0.6. Members of paralog clusters were then organized into ortholog clusters by allowing any member of a paralog cluster to contribute to the reciprocal best matches used to construct the ortholog clusters. Syntenic blocks are defined as a set of five or more consecutive pairs of genes from the same ortholog cluster. Because they do not participate in clusters, all contigs that do not contain protein-coding genes from the five draft genomes were searched against all genomes by using the NUCMER program (16). Syntenic blocks and NUCMER results were drawn (Fig. 1) by using SYBIL (http://sybil.sourceforge.net/).

In Fig. 1, genomic islands of diversity >5 kb are predicted as follows: (*i*) strains are inspected from the top panel and down and from left to right on each panel; (*ii*) regions of at least 1 kb not shared with another strain are identified; (*iii*) regions are merged into single islands if they are within 5 kb of each other; and (*iv*) resulting islands >5 kb are considered. It should be noted that some islands are composed of more than one contig. Genomic islands discussed in the text are the following: the α-galactosidase region in strain H36B, island 7.4; the prophage region in strain H36B, island 7.5; the DNA restriction/modification system in strain 515, part of island 4.5; the Tn*916* regions in strains 2603V/R, 515, CJB111, and COH1, islands 1.8 and the left side of 5.3; and serine-rich protein and glycosyltransferases flanked by cell-wall-anchored proteins and sortases in strain COH1, unnumbered region between islands 6.5 and 1.15. Fig. 1 reveals many non-protein-coding regions in strain 18RS21 that display NUCMER matches elsewhere in the 18RS21 genome. Most of these regions correspond to fragments of rRNAs, tRNAs, or structural RNAs, all of which exhibit an expected atypical nucleotide composition.

χ^2 Analysis. Regions of atypical nucleotide composition were identified by the χ^2 analysis; the distribution of all 64 trinucleotides (3mers) was computed for the complete genome in all six reading frames, followed by the 3mer distribution in 5,000-bp windows. Windows overlapped by 500 bp. For each window, the χ^2 statistic on the difference between its 3mer content and that of the whole genome was computed. Peaks in Fig. 1 indicate regions of atypical nucleotide composition.

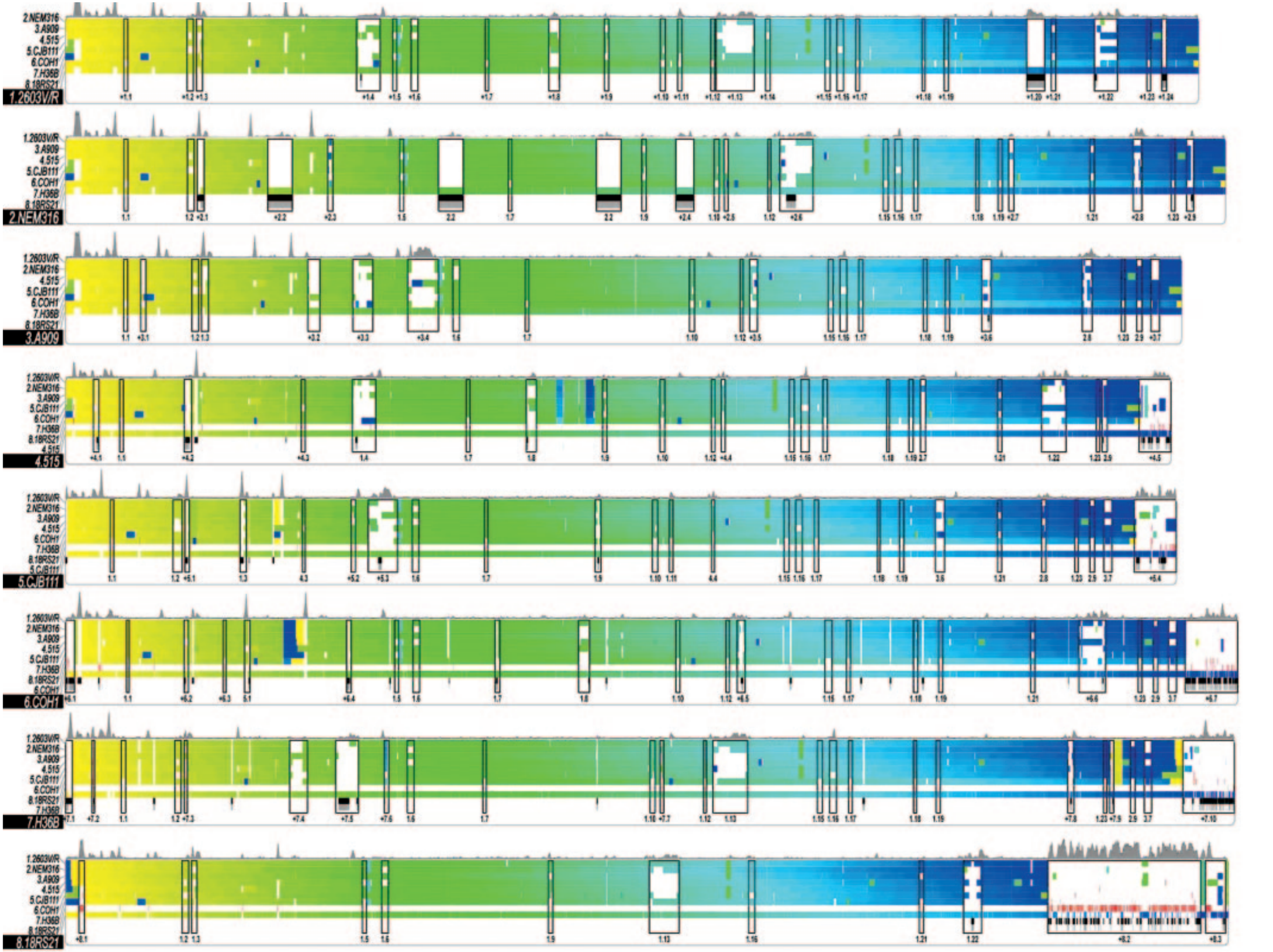

2.NEM316
3.A909
4.515
5.CJB111
6.COH1
7.H36B
8.1BRS21
1.2603V/R
1.2603V/R
3.A909
4.515
5.CJB111
6.COH1
7.H36B
8.1BRS21
2.NEM316
1.2603V/R
2.NEM316
4.515
5.CJB111
6.COH1
7.H36B
8.1BRS21
3.A909
1.2603V/R
2.NEM316
3.A909
5.CJB111
6.COH1
7.H36B
8.1BRS21
4.515
1.2603V/R
2.NEM316
3.A909
4.515
6.COH1
7.H36B
8.1BRS21
5.CJB111
1.2603V/R
2.NEM316
3.A909
4.515
5.CJB111
7.H36B
8.1BRS21
6.COH1
1.2603V/R
2.NEM316
3.A909
4.515
5.CJB111
6.COH1
8.1BRS21
7.H36B
1.2603V/R
2.NEM316
3.A909
4.515
5.CJB111
6.COH1
7.H36B
8.1BRS21
0 Mb
0.25 Mb
0.5 Mb
0.75 Mb
1 Mb
1.25 Mb
1.5 Mb
1.75 Mb
2 Mb
2.25 Mb

Fig. 1. Whole genome alignment of GBS strains. The eight genomes are compared to each other by using COG (41) and NUCMER analyses (see *Materials and Methods*). Each genome (shaded strain name) is colored with a gradient that ranges from yellow (nucleotide 1) to blue (end). Differences in color between a reference sequence (the last colored line in each genome) and the other genomes indicate conserved protein-coding regions that have been rearranged. Uncolored segments denote coding regions in which no conserved genes were detected. NUCMER matches for contigs that do not contain protein-coding genes are displayed by red blocks (matches within the reference strain are displayed on the line directly above it). Genomic islands of diversity are boxed and numbered "*x*.*y*," where *x* is the panel or strain number where the island first appeared and *y* is the island location in that genome from left to right. A indicates an island that was not identified in a previous genome. Islands that overlap by at least 50% (based on the number of shared genes) with previously identified islands receive the same number as the initial island. The gene content of the 69 islands identified is listed in Table 2, which is published as supporting information on the PNAS web site. Strain-specific regions, free of COG or NUCMER matches, are displayed in black at the bottom of each panel. Portions of these regions that harbor protein-coding genes are indicated in gray below the black blocks. The curves on top of each panel represent the nucleotide composition (^{2} analysis) (see *Materials and Methods*) of the reference strain of the panel, and peaks indicate regions of atypical composition.

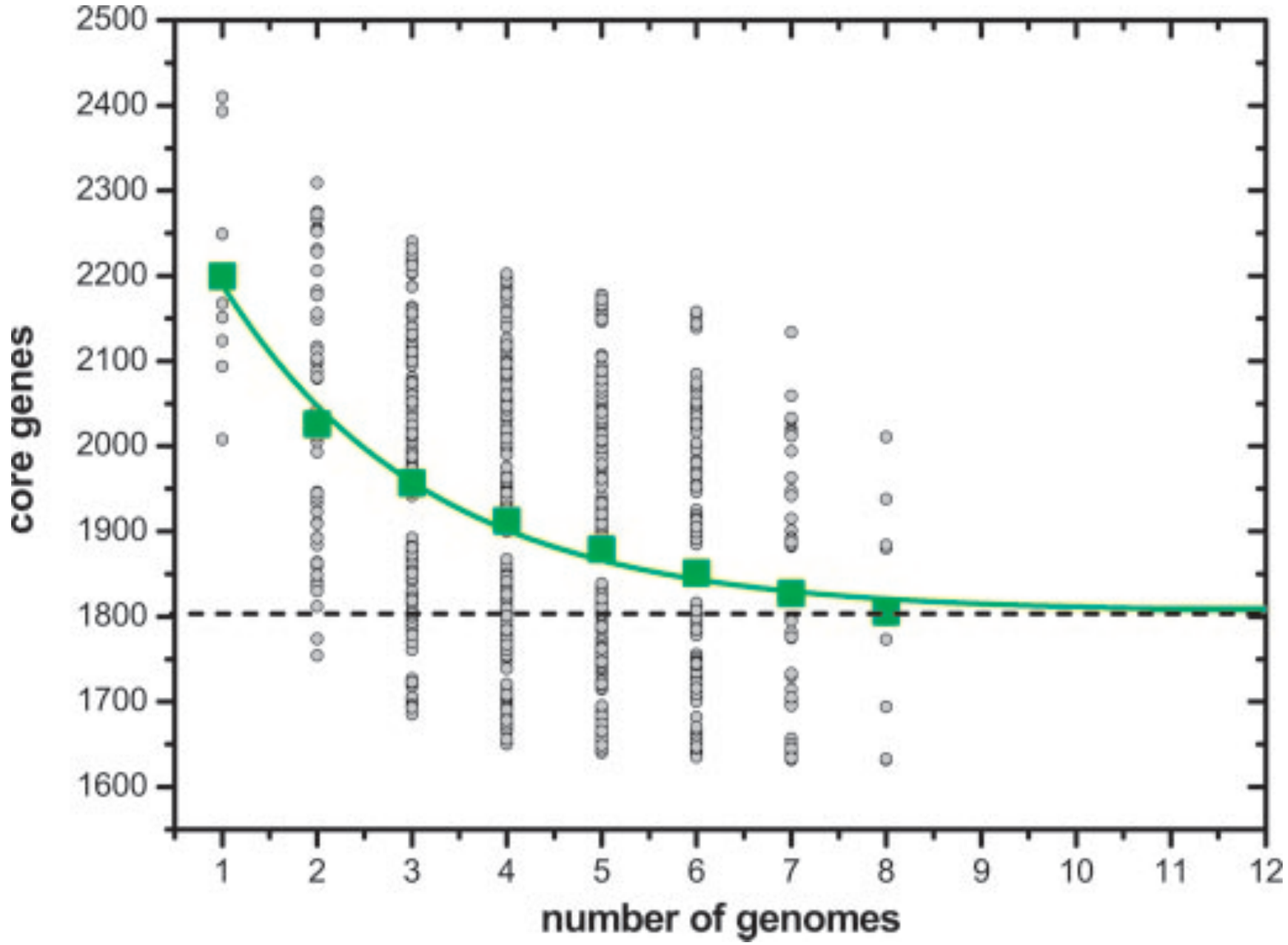


Fig. 2. GBS core genome. The number of shared genes is plotted as a function of the number n of strains sequentially added (see *Materials and Methods*). For each n, circles are the $8!/[(n-1)!\cdot(8-n)!]$ values obtained for the different strain combinations. Squares are the averages of such values. The continuous curve represents the least-squares fit of the function $F_c = \kappa_c \exp[-n/\tau_c] + \Omega$ (see Eq. **1** in *Supporting Text*) to data. The best fit was obtained with correlation $r^2 = 0.990$ for $\kappa_c = 610 \pm 38$, $\tau_c = 2.16 \pm 0.28$, and $\Omega = 1{,}806 \pm 16$. The extrapolated GBS core genome size Ω is shown as a dashed line.

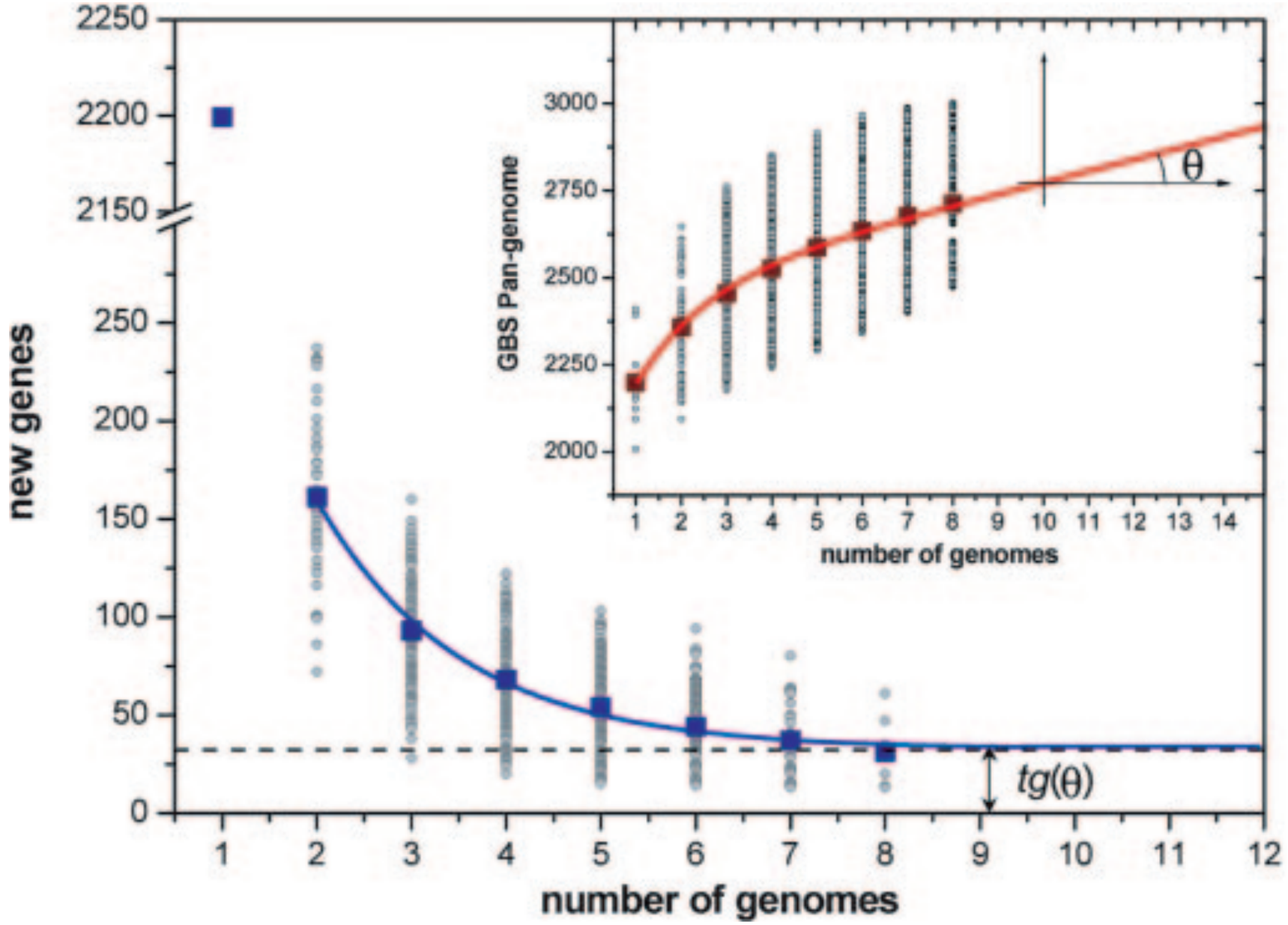


Fig. 3. GBS pan-genome. The number of specific genes is plotted as a function of the number n of strains sequentially added (see *Materials and Methods*). For each n, circles are the $8!/[(n-1)!\cdot(8-n)!]$ values obtained for the different strain combinations; squares are the averages of such values. The blue curve is the least-squares fit of the function $F_s(n) = \kappa_s \exp[-n/\tau_s] + tg(\theta)$ (see Eq. **2** in *Supporting Text*) to the data. The best fit was obtained with correlation $r^2 = 0.995$ for $\kappa_s = 476 \pm 62$, $\tau_s = 1.51 \pm 0.15$, and $tg(\theta) = 33 \pm 3.5$. The extrapolated average number $tg(\theta)$ of strain-specific genes is shown as a dashed line. (*Inset*) Size of the GBS pan-genome as a function of n. The red curve is the calculated pan-genome size

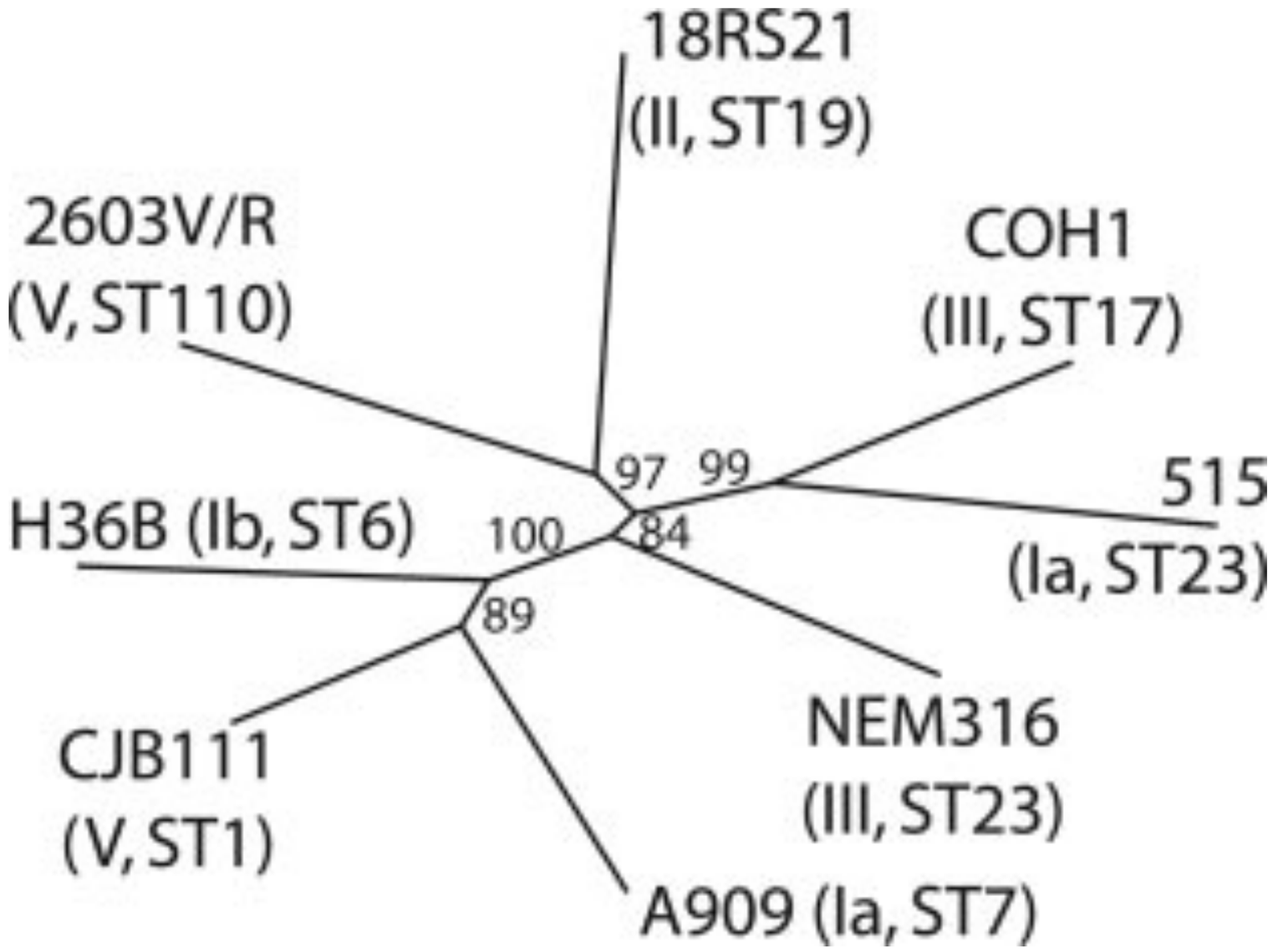
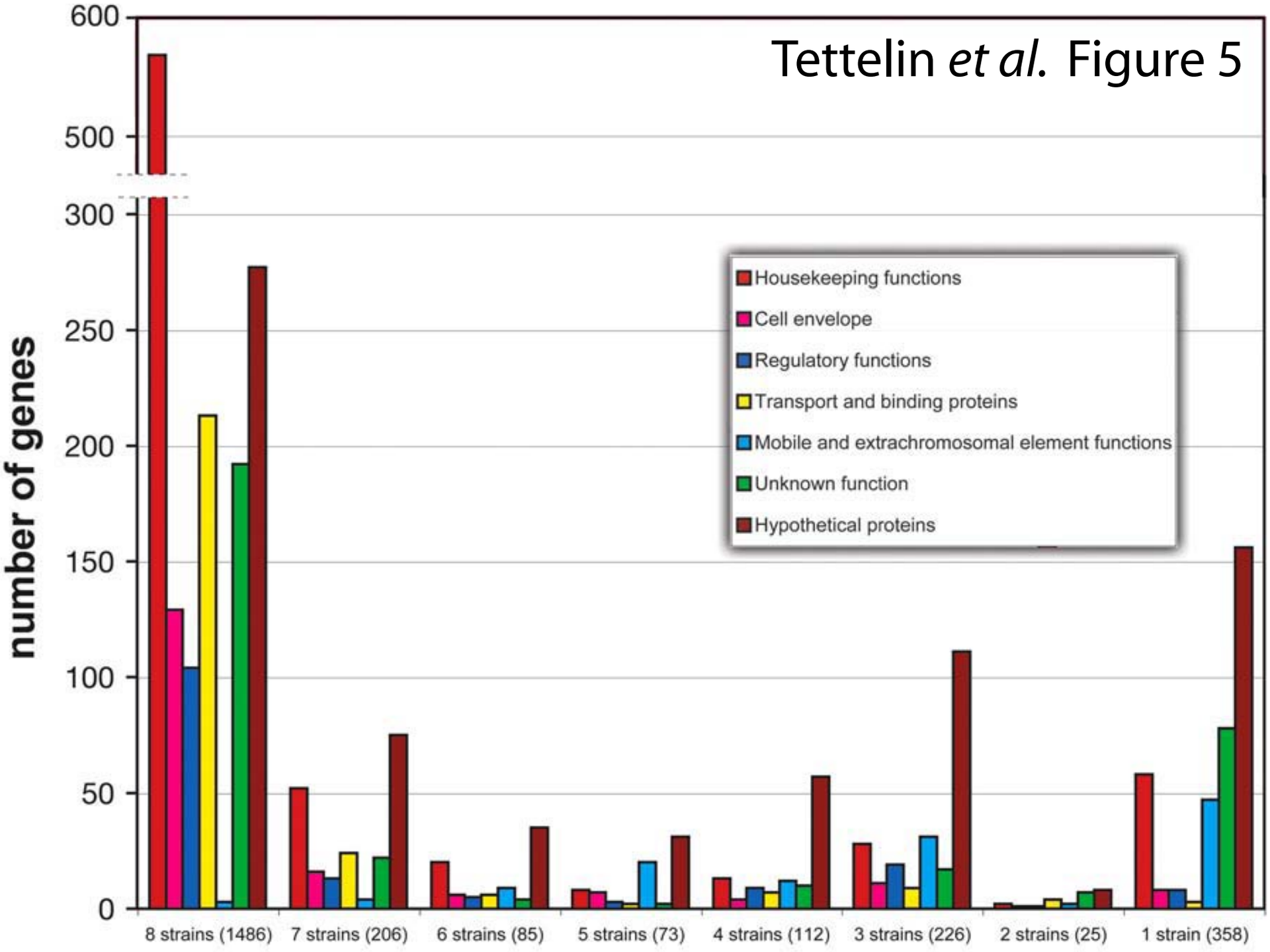


Fig. 4. Dendrogram of the eight GBS genomes. Shared gene information was used to cluster proteins into groups by using the single-linkage method of the program CLUSTER (http://rana.lbl.gov). Groups were then converted into profiles of presence or absence of each gene (0 or 1) in the eight GBS strains and used as input to PAUP* 4.0b10 (Sinauer, Sunderland, MA) for dendrogram drawing and bootstrapping. Numbers at the nodes indicate bootstrap values. Serotypes and MLST types of each strain are within parentheses.

Tettelin *et al.* Figure 5



Distribution of shared and unique genes sorted by functional categories. Groups of shared genes and strain-specific genes were sorted by broad functional categories as follows: housekeeping functions, cell envelope proteins, regulatory functions, transport and binding proteins, mobile and extrachromosomal elements, proteins of unknown function, and hypothetical proteins. The total number of genes in each of these categories was counted and displayed for groups shared by all eight strains, groups shared by any combination of seven to two strains, and strain-specific genes.

Implications for Bacterial Taxonomy. Methods commonly used to define bacterial species (DNA·DNA reassociation, 16S rRNA typing, MLST, etc.) rely mostly on features associated with the core genome (40). Our work confirms that the essence of the species is linked to the core genome. However, the majority of the genetic traits linked to virulence, capsular serotype, adaptation, and antibiotic resistance pertain to the dispensable genome. Therefore, sequencing of multiple strains is necessary to understand the virulence of pathogenic bacteria and to provide a more consistent definition of the species itself. We identified species with an open pan-genome, such as GBS and GAS, and species with a closed pan-genome, such as *B. anthracis*. Nevertheless, a different interpretation of the same data may lead to the conclusion that the present definition of bacterial species is inconsistent because, in reality, only species with an open pan-genome are species, whereas *B. anthracis* is not a true genetic species on its own, but only a clone of *Bacillus cereus*, with very distinctive phenotypic traits provided by the acquisition of the virulence plasmid coding for the anthrax toxin.

Concluding Comment. Our data clearly show that the strategy to sequence one or two genomes per species, which has been used during the first decade of the genomic era, is not sufficient and that multiple strains need to be sequenced to understand the basics of bacterial species. The methods presently used to evaluate the species diversity, such as complete genome hybridization and MLST, can explain only the presence, absence, and variability of the genetic loci that are already known and do not provide information on the genes that are not present in the reference genome. Our work provides a clear demonstration that, by these approaches, we fail to include in the analysis the entire dispensable genome, the size of which can be vastly larger than the core genome. Our work on the protein-based vaccine against GBS has shown that this is not just a theoretical disadvantage but has very important practical consequences because a universal vaccine is possible only by including dispensable genes (8).