**Supporting Text**

**Comparative Genomics (Basis for Figs. 2 and 3).** To define conserved genes among the different strains, the variable degree of accuracy of sequence annotation was taken into account. In addition to the caveats inherent to the unfinished genomes, different criteria might have been applied to the annotation of strain NEM316 because it was published by a different group. Therefore, a simple ORF-versus-ORF sequence similarity search could give only a conservative estimate of the number of conserved genes. For this reason, we have used the three different algorithms indicated in the text to perform an all-versus-all comparison of the strains. Results are shown in Table 7. Because a conserved gene can appear in various numbers of paralogous copies, the number of genes shared by two strains can be asymmetric.

In Table 8, the number of strain-specific genes absent from all other strains is shown. All the paralogs within a genome were considered as independent genes, with the exception of the three identical islands III, VII, and VIII containing 49 genes in the NEM316 strain (see island numbers in ref. 1) that were counted only once.

**Shared Genes and Core-Genome Extrapolation (Fig. 2).** The number of core genes shared by all GBS isolates depends on how many strains are taken into account. The sequential inclusion of up to eight strains was simulated in all possible combinations. By measuring the number of conserved genes as a function of the number of sequenced strains, it is possible to extrapolate the size of the species core genome. For $n = 1, 2, \ldots, 8$: (*i*) a panel of $n - 1$ genomes is considered already known; (*ii*) the $n$th strain $A$ is chosen from the $8 - (n - 1)$ strains not present in the panel; and (*iii*) the genes in $A$ that are present in all of the $n - 1$ genomes in the panel are counted. The procedure is repeated for all independent permutations in the order of the genomes, and the results are reported in Fig. 2 as a function of $n$. The points corresponding to $n = 1$ are the number of genes present in the eight genomes. For $n = 2$, 56 points are reported, corresponding to the $8 \cdot 7 = 56$ independent pairs that can be chosen among eight strains. Because no distinction is made between paralogs, a gene conserved across two strains can be present in a different number of copies in the two genomes. As a consequence, the number of genes in genome $A$ shared with genomes $B$ and $C$ can be different from the corresponding number of genes in $B$ shared with genomes $A$ and $C$. In addition, the number of genes that are conserved in the $n$th genome depends only on the identity of the already known $n - 1$ genomes but not on their sequencing order. Taking these two facts into account, the number $N$ of independent measurements of the shared genes present in the $n$th genome is:

$$N = 8!/[(n - 1)! \cdot (8 - n)!].$$

This formula gives the number of points reported in Fig. 2 as a function of *n*.

To evaluate the size of the GBS species core genome, the average values of the shared genes were extrapolated by fitting the exponential decay function:

$$F_c(n) = \kappa_c \exp\left[-n/\tau_c\right] + \Omega \qquad \textbf{[1]}$$

to shared genes data, with a least-square Levenberg-Marquardt algorithm with $\kappa_c$, $\tau_c$, and $\Omega$ as free parameters. In Eq. **1**, $\kappa_c$ is the amplitude of the exponential decay, $\tau_c$ is the decay constant that measures the speed at which $F_c(n)$ converges to its asymptotic value, and $\Omega$ measures the size of the core genome for $n \to \infty$, where *n* is the continuous extrapolation of the number of strains taken into account. Results of the fitting procedure show that, with Pearson's correlation coefficient $r^2$ = 0.990, the size of the GBS core genome converges to a plateau value of $\Omega$ = 1,806 ± 16 genes (Table 3). Because of the variable number of paralogs present in different strains and the low level of sequence conservation required to consider a gene as conserved, this value gives an upper limit to the size of the core genome of GBS. Estimated values of the other free parameters are $\kappa_c$ = 610 ± 38 and $\tau_c$ = 2.16 ± 0.28.

Alternatively, the parameters of Eq. **1** can be estimated by directly fitting Eq. **1** to the data, instead of to the averages; results for the parameters are indicated in Table 4.

**Strain-Specific Genes and Pan-Genome Extrapolation (Fig. 3).** As for shared genes, the number of strain-specific genes or genes not present in any other GBS isolate depends on the number of strains taken into account. Following a procedure similar to the one used to evaluate the size of the core genome, for each GBS strain and for *n* = 1, 2, …, 8 the number of strain-specific genes, i.e., genes that are not present in any of the previously known *n* – 1 sequences, that are first seen by sequencing the *n*th strain was counted. Again, all possible combinations of the sequencing order were considered, and the results are reported in Fig. 3 as a function of *n*. The number of specific genes pertaining to the *n*th genome depends only on the identity of the previously sequenced *n* – 1 genomes, but not on their sequencing order, and, therefore, the number of independent combinations is again:

$$N = 8!/[(n-1)!\cdot(8-n)!].$$

The average number of specific genes per strain closely follows an exponential decay that was fitted with the function

$$F_s(n) = \kappa_s \exp\left[-n/\tau_s\right] + tg(\theta), \qquad \textbf{[2]}$$

where $tg(\theta)$ measures the number of specific genes for $n \to \infty$, and $\kappa_s$, $\tau_s$, and *n* have the same meaning as in Eq. **1**. Least-squares fit of Eq. **2** to strain-specific genes data shows, with extremely high values of the Pearson correlation coefficient $r^2$ = 0.995, that the average number of specific genes converges to the finite plateau value of $tg(\theta) = 33 \pm 3.5$ genes. The maximum likelihood value, the standard error, and the upper and lower limits for 95% and 99% confidence levels for each parameter are reported in Table 5. In particular, from these data we can estimate that the probability that $tg(\theta) = 0$ is $6 \times 10^{-4}$.

Alternatively, the parameters of Eq. **2** can be estimated by directly fitting Eq. **2** to the data, instead of to the averages (Table 6). In this case, the probability that $tg(\theta) = 0$ is smaller than $2 \times 10^{-16}$.

Although the precise value of the parameters can change depending on the fitting procedure, in both cases the fit excludes that the data are statistically compatible with $F_s(n)$ decaying to 0 for large values of $n$.

Because of the low level of sequence conservation required to consider a gene as conserved, $tg(\theta)$ gives a conservative lower bound estimate of the extrapolated number of strain-specific genes that would be found, on average, upon sequencing of additional independent GBS strains.

The value of $tg(\theta)$ is crucial in the extrapolation of the pan-genome size, i.e., the total number of distinct genes pertaining to the GBS species. From Eq. **2**, the pan-genome size *P*(*n*) can be calculated as a function of the number of independent strains *n*:

$$P(n) = D + \sum_{j=2}^{n} \left\{ \kappa_s \exp\left[ -j / \tau_s \right] + tg(\theta) \right\}, \qquad \textbf{[3]}$$

where *D* = 2,199 is the average number of genes per sequenced genome. With simple algebra:

$$P(n) = D + tg(\theta)[n-1] + \kappa_s \exp\left[ -2/\tau_s \right] \cdot \sum_{j=0}^{n-2} \exp\left[ -j/\tau_s \right]$$

$$= D + tg(\theta)[n-1] + \kappa_s \exp\left[ -2/\tau_s \right] \cdot \frac{1 - \exp\left[ -(n-1)/\tau_s \right]}{1 - \exp -1/\tau_s}. \qquad \textbf{[4]}$$

From Eq. **4**, $\lim_{n \to \infty}[P(n)] \approx tg(\theta) \cdot n$, where θ represents the extrapolated angle of growth of the pan-genome size *P*(*n*) as more independent GBS strain sequences become available. Eq. **4** defines the pan-genome size growth without introducing new free parameters, because parameter values are

obtained from the fit of Eq. **3** to specific-genes data. Fig. 3 *Inset* displays the measured size of the pan-genome as a function of $n$ [in this case the order of genome addition is relevant, so $N = 8!/(8 - n)!$ points are obtained for each value of $n$] together with a plot of $P(n)$ calculated from Eq. **4**.

1. Maione, D., Margarit, I., Rinaudo, C. D., Masignani, V., Mora, M., Scarselli, M., Tettelin, H., Brettoni, C., Iacobini, E. T., Rosini, R., *et al.* (2005) *Science* **309,** 148–150.