Update

Genome Analysis

# Estimating the size of the bacterial pan-genome

**Pascal Lapierre[1] and J. Peter Gogarten[2]**

[1] University of Connecticut Biotechnology Center, 91 North Eagleville Road, Storrs, CT 06269-3149, USA
[2] Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125, USA

**The 'pan-genome' denotes the set of all genes present in the genomes of a group of organisms. Here, we extend the pan-genome concept to higher taxonomic units. Using 573 sequenced genomes, we estimate the size of the bacterial pan-genome based on the frequency of occurrences of genes among sampled genomes. Using gene- and genome-centered approaches, we characterize three distinct pools of gene families that comprise the bacterial pan-genome, each evolving under different evolutionary constraints. Our findings indicate that the pan-genome of the bacterial domain is of infinite size (the Bacteria as a whole have an open pan-genome) and that ~250 genes per genome belong to the extended bacterial core genome.**

## Genome plasticity and evolution

The availability of several hundred completely sequenced genomes has changed our views of genome evolution and uncovered extensive gene sharing between organisms. The view of stable genomes that function as unchanging information repositories has given way to a more dynamic view in which genomes frequently lose genes and incorporate foreign genetic materials [1,2]. The term 'pan-genome' or 'supragenome' denotes the set of all genes present in the genomes of members of a group of organisms, usually a species [3,4]. The pan-genome includes genes present in only one organism (known as ORFans), in the genomes of a few members of the group or in genes that are present in all genomes of the group (known as the core genome). Previously, Tettelin *et al.* [3] have shown that each individual strain of Group-B *Streptococci* (GBS) contains 13–61 unique genes and that, extrapolated to infinity, one would expect to find ~30 new genes for every additional GBS genome sequenced. Here, we apply this pan-genome concept to the bacterial branch of the tree of life, evaluating the dynamics of genome and gene family evolution and characterizing two modes of evolution: reuse with variation and *de novo* creation.

## From gene frequency to pan-genome

The approach developed by Tettelin *et al.* [3] to define the pan-genome consisted of tracking the number of unique genes among genomes in successive blast searches. This genome-oriented method is useful when a limited number of genomes are analyzed but computationally difficult when the number of genomes sampled is too large (total number of different sequential paths for $n$ genomes sampled is equal to $n!$). Because this method enables estimation of the frequency of occurrence of genes in genomes, the reverse also hold true. By using the frequency of occurrence of genes among genomes (i.e. in how many genomes do sampled genes have a homolog?), one can extrapolate back the sampling curve of the actual pan-genome of the group of organisms studied. This gene-oriented method has the advantage of being computationally less intensive and simultaneously providing a direct assessment of the gene frequencies among genomes, regardless of their genome of origin. Both approaches were initially compared using 293 completely sequenced genomes that were available at the time when this analysis was first conducted. The gene-oriented approach was later expanded to 573 bacterial genomes (for a list of all genomes sampled, see Table S1 in the supplementary material online) and yields very similar results (Table S2). We did not include archaeal genomes in our analyses because archaeal and bacterial homologs often are too divergent to establish homology through simple blast searches.

A total of 15 000 open reading frames (ORFs) were randomly selected from any of the 293 genomes (each ORF could only be selected once) and basic local alignment search tool (BLAST) searches were used to determine for each gene the number of genomes in which homologous sequences could be found (we required a bitscore >50 to classify a gene as present in the target genome and as a member of the same gene family). The total of 15 000 genes is sufficient to accurately reconstruct the sampling curve from the genome-centered approach. The resulting data were used to build a histogram in which each point represents the normalized number of genes ($An$) at the different frequencies (Fq) of occurrence in genomes (Figure 1a). The frequency distribution shows clustering of genes at both extremities of the histogram and most frequency categories contain approximately the same number of genes in the central part. The reconstruction of the sampling curve by adding up each individual component of the histogram, $f(x) = \sum[A_n{}^{*}e^{(Kn^{*}x)}]$, agrees with the data generated using the genome-centered approach, showing the equivalence of the two approaches (Figure 1b). From this histogram, three groups of ORFs are distinguished: (i) the extended core made up of ORFs on the right hand side of the diagram that occur in all or nearly all genomes; (ii) the accessory pool represented by ORFs on the left hand side of the diagram, comprising genes present in only one or a few genomes; and (iii) the remainder of the diagram comprised of proteins that are encoded in only a subset of the genomes. Here, we term these character genes because they define or can be used to define the character of groups of genomes. A decay function fitted to the reconstructed sampling curves was used to estimate

*Corresponding author:* Lapierre, P. (pascal.lapierre@uconn.edu)

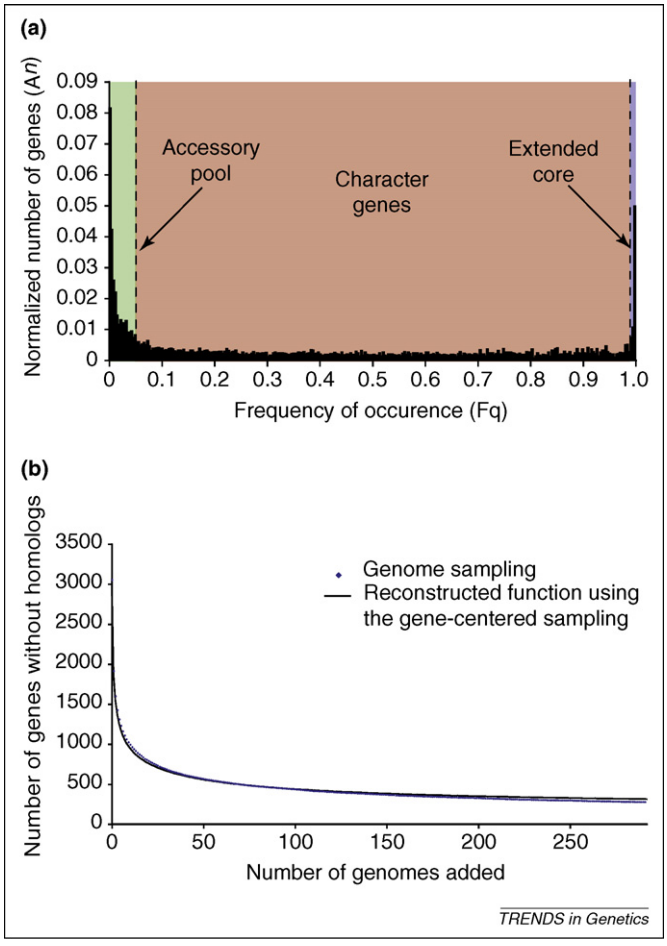


**Figure 1**. Frequency of occurrence of randomly selected genes in 293 bacterial genomes. **(a)** 15 000 genes were sampled to determine their frequencies of occurrence among genomes. Each bar corresponds to the normalized number of genes [*n* genes at Fq(*x*)/15000] having the indicated frequency (Fq) of occurrence (present in *n* other genomes/Total number of genomes −1). Genes without any homologs (Fq = 0) represent ORFans, whereas genes present in 292 other genomes (Fq = 1) represent strict core genes. Parts of the histogram that mainly contribute to the extended core, the character genes and the accessory pool are colored in blue, red and green, respectively (see Figure 2 for a definition of each of these categories). From the decay components (K) of the sampling functions for extended core and rare genes (see supplementary material online), the boundaries between the three pool of genes were determined by genes present in at least 99% of the genomes for the core set of genes and genes absent in at least 95% of the genomes for the accessory pool. **(b)** The frequency sampling can be used to reconstruct the sampling curve expected from the genome-centered approach. The sampling function reconstructed from the frequency histogram, $F(x) = \sum[A_n^* e^{(Kn^*x)}]$, $K = Ln(1 - Fq)$, agree with the data obtained with the sampling using the genome-centered approach. The slight difference between the genome-centered and the reconstructed sampling curves is caused by the probability of the sampling of individual gene. In the gene-centered approach, each gene, regardless of its genome of origin, has the same probability to be sampled, causing over representation of genes from larger genomes compared to the genome-centered approach. Because large genomes tend to harbor more duplicates and ORFans, it will cause the sampling curve to decay faster and to reach stability at slightly higher values.

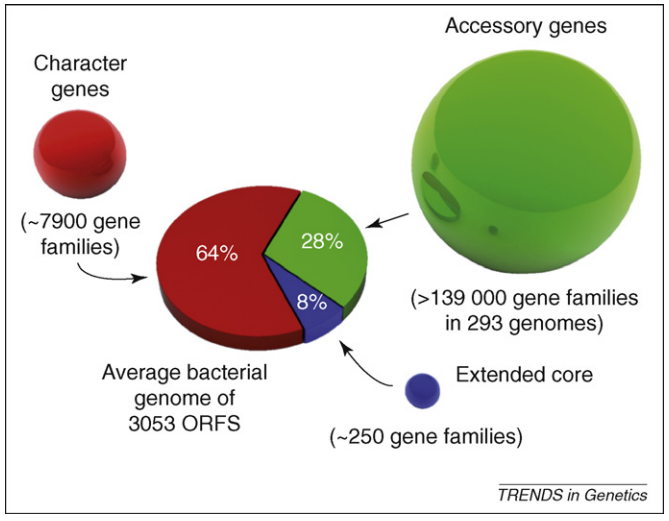


**Figure 2**. The bacterial pan-genome. Each gene found in the bacterial genome represents one of three pools: genes found in all but a few bacterial genomes comprise the extended core of essential genes (~250 gene families that encode proteins involved in translation, replication and energy homeostasis); the character genes (~7900 gene families) represent genes essential for colonization and survival in particular environmental niches (e.g. symbiosis and photosynthesis); and finally, the accessory genes, a pool of apparently infinite size, contains genes that can be used to distinguish strains and serotypes; the function of most genes in this category is unknown. At the genomic level, a typical bacterial genome is composed of ~8% of core genes, 64% of character genes and 28% of accessory genes. Although the character genes contain only 7900 gene families, they are the most abundant at the genomic level. Expanding the gene-centered approach to 573 bacterial genomes or sampling of 508 genomes, excluding highly reduced genomes, yields similar results (Table S2), except that the total number of families in the accessory pool is increased as expected for an open pan-genome.

the size of the three different groups of genes and to extrapolate the sampling curves to higher numbers of sampled genomes as additional genomes are sampled (see methods in the supplementary material online for more details).

## The extended core, character genes and accessory pool

The existence of a core set of genes present in all bacteria is testament to the conservative nature of evolution. Within several billions of years of bacterial evolution, no successful replacement of the core genes evolved in any of the lineages leading to the studied genomes. The core set of genes is under high selective pressure for a function that prevents drastic changes. The gene frequency approach presented here enables relaxing the core definition to include genes that are missing in only a small fraction of the genomes. This extended core of shared genes, which represent genes present in at least 99% of the sampled genomes (as determined by the fastest decay component of the sampling function), constitutes ~8% of the genes present in a typical bacterial genome (Figure 2). As pointed out by Koonin *et al.* [5], this set of core genes does not correspond to the minimal set of genes necessary for an organism to survive and thrive in nature. It is rather a backbone of essential components on which the rest of the genome is built.

Interestingly, although the character genes were found to be the main component of every bacterial genome (~64% of the total genes on average), this set of genes only contains ~7900 gene families. The rather small number of protein families found in the character pool is offset by the flexibility of these genes in their ability to adapt to new functions. Although similar on the sequence level, the character gene families demonstrate high diversity of substrate specificity. Instead of using a random process of creating new genes *de novo* to adjust to a situation, the limited number of character gene families indicates that the preferred mode of adaptation in bacteria consists of exploring new solutions from existing sequences via gene duplications, mutations and a mix and match assembly of modular proteins [6–9]. For example, the large gene family