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)

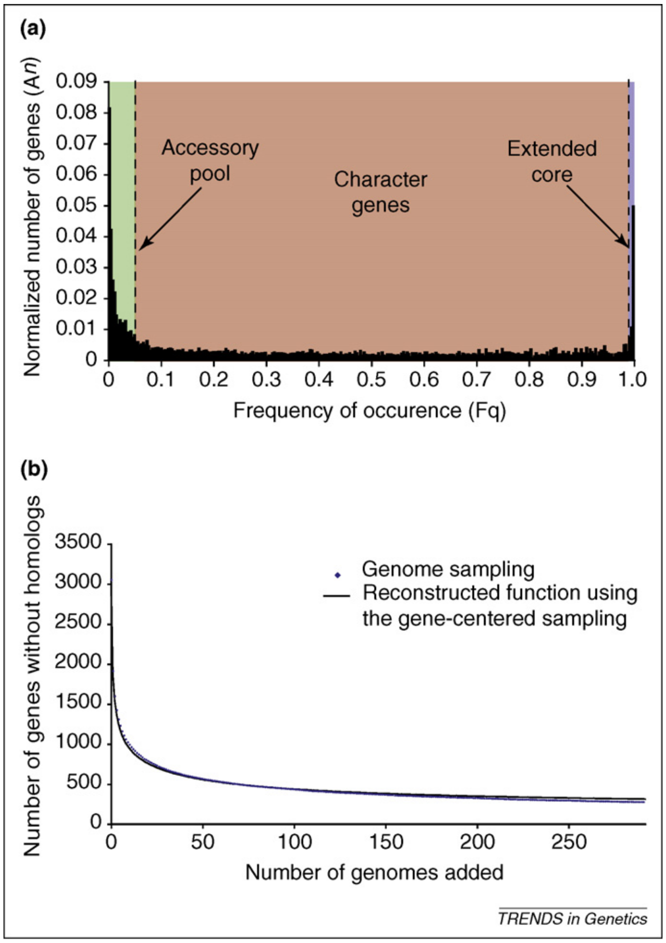

Figure 1. Frequency of occurrence of randomly selected genes in 293 bacterial genomes. **(a)** 15 000 genes were sampled to determine their frequencies of occurrence among genomes. Each bar corresponds to the normalized number of genes [*n* genes at Fq(*x*)/15000] having the indicated frequency (Fq) 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, $F(x) = \sum[A_n{}^{*}e^{(Kn^{*}x)}]$, $K = \mathrm{Ln}(1 - \mathrm{Fq})$, 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.

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.

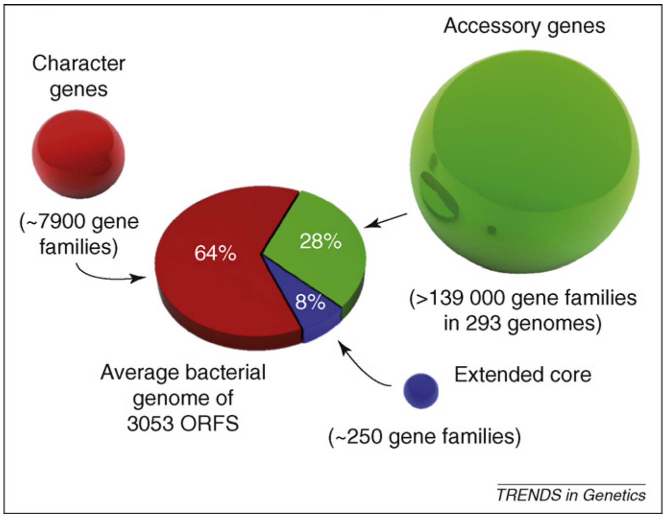

Figure 2. The bacterial pan-genome. Each gene found in the bacterial genome represents one of three pools: genes found in all but a few bacterial genomes comprise the extended core of essential genes (~250 gene families that encode proteins involved in translation, replication and energy homeostasis); 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.

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 of ABC transporters has diverse substrate specificities [6], which is caused by the substitutions in the periplasmic binding subunit [7,8]. Type I polyketide-synthases (PKSs) are large modular proteins that rely mainly on the combination of different protein sub-domains to achieve different functions [9]. Spreading in genomes through gene duplications and transfers, the seven characterized PKS domains assemble into multifunctional enzymes that synthesize many important secondary metabolites [9,10].

The creation of new protein folds might be reflected in the accessory pool. Many of these genes are ORFans (i.e. genes that do not have homologous sequences in other genomes). A closer analysis of the accessory genes found in *Escherichia coli* str. K12 substr. MG1655 reveals that, on average, these ORFs have a greater AT%, tend to be shorter and many are composed of insertion sequence (IS) elements and prophage sequences (see supplementary material online). We have found >139 000 rare gene families scattered throughout the bacterial genomes included in this study. The finding that the fitted exponential function approaches a plateau indicates an open pan-genome (i.e. the bacterial protein universe is of infinite size); a finding supported through extrapolation using a Kezdy-Swinbourne plot (Figure S3). This does not exclude the possibility that, with many more sampled genomes, the number of novel genes per additional genome might ultimately decline; however, our analyses and those presented in Ref. [11] do not provide any indication for such a decline and confirm earlier observations that many new protein families with few members remain to be discovered [12].

This set of accessory genes does not seem to be tightly bound to a particular organismal lineage. Their low level of conservation might indicate processes that can create new proteins [13]. These genes are frequently not subject to strong selective pressures [13,14] and they have high turnover rates in genomes [15]. Their likely association with bacteriophages and plasmids indicates that their evolution might transcend the organismal line of descent [16–18]. Regardless of their mode of insertion into bacterial genomes, the genes of the accessory pool seem to represent an ongoing gene creation process different from domain shuffling. Some of the genes in the accessory pool represent annotation artifacts resulting in ORFs that are not actually transcribed and translated. However, the number of falsely identified ORFs is usually estimated to be much smaller than the size of the accessory pool (1–4% [19] versus 28% of accessory genes per genome found in this study). In most instances, the process of gene creation might not lead to useful functions and the genes can be lost from the genomes. Occasionally, a new invention might arise from this cloud of genes and spread in and between populations as a result of the adaptive advantage provided, thereby moving the encoding gene to the pool of character genes.

Extending the pan-genome concept to higher taxonomic levels exacerbates the ambiguity in deciding if a gene should count as a new addition to the pan-genome or be considered as already present. This problem already exists for the pan-genome of a single species, especially in case of paralogs; however, for organisms belonging to different phyla, a protein with the same function might be so divergent that only the use of PSI blast or clustering might identify the homology [20–23]. Incorporating lineage-specific duplications and distinguishing them from ancient paralogs might be a useful extension to our gene and genome-centered classification schemes. A paralogous protein with an alignment score above the cut-off (a bit-score of 50), which is present in a target genome and which has lost the orthologous gene, would falsely cause the query gene to be considered to be present in the target genome. Under both approaches, the sampling of genes does not discriminate between orthologs and paralogs. However, because every gene is used as a query, paralogous genes present in the same genome are counted as separate families, resulting in ~8000 gene families in the character gene pool and ~250 gene families in the extended core. The choice of a simple blast hit cut-off to identify homologs might lead to falsely classifying a gene as different, just because it has diverged beyond recognition. This results in an underestimation of the number of genes in the extended core (two character gene families might be joined into a single family present in almost all bacteria), although within the bacterial domain divergence for core genes to score below a bitscore of 50 seems unlikely. Conversely, our simple approach to identify homologs probably overestimates the number of character genes: A small number of these character genes might have diverged below the chosen similarity cut-off and could be joined into a single family if conserved domains were used as classification. For example, many between-phyla comparisons of reaction center proteins from different photosynthetic bacteria score below the cut-off but, considering their similar fold and function, these should be considered as homologs. In other words, the surprisingly small number of families in the character gene pool would be even smaller.

Concluding remarks

Since the completion of the first genomic sequence, we have come to appreciate the many forces acting on genome evolution [24]. The view of stable genomes functioning only as slowly changing repositories of genetic information gave way to a dynamic view whereby genomes function like collecting bins, continuously gaining and losing genes along the way. This constant rain of genetic material on genomes from a cloud of frequently transferred genes enhances the chance of survival of species by introducing variability in the population. We have identified three categories of gene that compose each genome: the extended core, the character genes and an accessory pool of genes. Proteins in these categories are evolving under different constraints and rules. Genes in the extended core are under high selective pressure and only minute changes at the sequence level are allowed. Although many instances of gene transfers have been documented, they mainly spread in populations through vertical inheritance. Gene duplication and domain shuffling are the preferred mode of evolution of the character genes. This set of genes enables organisms to quickly adapt to changing conditions and to exploit new niches. Of the three sets of genes, the character genes are the most likely to be transferred between organisms. The last category of genes consists of genes with low levels of conservation, which are scattered at low frequencies throughout the bacterial domain. This accessory pool of genes might represent in part genes that had previous functions in genomes (now pseudogenes) but that are now stripped of selective pressure. These fast evolving genes, perhaps residing in phage genomes most of the time, explore sequence spaces and, occasionally, a new useful protein fold might arise from this pool and spread through populations.

How is protein space explored in biological evolution? *A priori*, two extreme points of view are possible: protein evolution is predominantly the result of selection and rearrangements of already existing proteins or protein evolution is an ongoing process in which new proteins evolve as the exploration of the protein landscape continues. Our results provide evidence for both processes operating in the bacterial world.

Acknowledgements

This work was supported by NASA Applied Information Systems Research (NNG04GP90G), NASA Exobiology (NAG5-12367) and NSF Microbial Genetics (MCB-0237197). We acknowledge the Bioinformatics services at the Biotech Center of the University of Connecticut for their computational support. We also thank the three anonymous reviewers for their constructive comments.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2008.12.004.

References

1 Snel, B. *et al.* (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25

2 Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719

3 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

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

5 Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* 1, 127–136

6 Davidson, A.L. and Chen, J. (2004) ATP-binding cassette transporters in bacteria. *Annu. Rev. Biochem.* 73, 241–268

7 Nanavati, D.M. *et al.* (2005) Substrate specificities and expression patterns reflect the evolutionary divergence of maltose ABC transporters in *Thermotoga maritima*. *J. Bacteriol.* 187, 2002–2009

8 Fukami-Kobayashi, K. *et al.* (2003) Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol. Biol. Evol.* 20, 267–277

9 Weissman, K.J. (2004) Polyketide biosynthesis: understanding and exploiting modularity. *Philos. Transact. A Math Phys. Eng. Sci.* 362, 2671–2690

10 Jenke-Kodama, H. *et al.* (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLOS Comput. Biol.* 2, e132

11 Falkowski, P.G. *et al.* (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320, 1034–1039

12 Yooseph, S. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16

13 Daubin, V. and Ochman, H. (2004) Start-up entities in the origin of new genes. *Curr. Opin. Genet. Dev.* 14, 616–619

14 Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687

15 Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44, 383–397

16 Hendrix, R.W. *et al.* (2000) The origins and ongoing evolution of viruses. *Trends Microbiol.* 8, 504–508

17 Daubin, V. *et al.* (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol.* 4, R57

18 Hsiao, W.W. *et al.* (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet.* 1, e62

19 Aggarwal, G. and Ramaswamy, R. (2002) *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* 27, 7–14

20 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28

21 Enright, A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584

22 Zhang, Z. *et al.* (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26, 3986–3990

23 Harlow, T.J. *et al.* (2004) A hybrid clustering approach to recognition of protein families in 114 microbial genomes. *BMC Bioinformatics* 5, 45

24 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512

0168-9525/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2008.12.004 Available online 23 January 2009

Supplementary data

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

METHODS

Completely sequenced bacterial genomes, including multiple chromosomes and plasmids were retrieved from the NCBI databank (ftp://ftp.ncbi.nih.gov/). The list of genomes used in this study is provided in Table 1S of the supplemental material.

For the genome-centered approach, samplings were generated by keeping track of the number of genes with and without homologs from randomly selected seed genomes as the number sampled genome (also selected at random) using BLAST searches increases (Figure 1S). A total of 1011 sampling runs were performed over 293 bacterial genome using successive blast searches (BLASTP) and using a bit score cutoff of 50 (corresponding to an E-value of ~10^{-4}) to count homologous matches. For every sampling run, each individual genome was selected only once. The genome sampling was terminated after 1011 sampling runs when it was determined that additional sampling runs will not change the sampling curves and their decomposition. Using averages for each sampling point, we tested the quality of the fit of different non-linear functions for the shared (core) and unique genes (ORFans) using the "fit command" in GNUPLOT 4.0

(http://www.gnuplot.info/). The best fits were determined by comparing the final sum of squares of residuals and the asymptotic standard errors of the function tested. The core sampling was best described by a three exponential decay functions ($F(x) = 2434.45*e^{(-0.707072*x)} + 410.543*e^{(-0.112567*x)} + 92.1156*e^{(-0.0100862*x)} + 116.735$). For the rare genes, the data was following a five exponential decay functions ($F(x) = 939.38*e^{(-2.08366*x)} + 731.11*e^{(-0.435631*x)} + 455.203*e^{(-0.0984529*x)} + 328.632*e^{(-0.0318822*x)} + 385.519*e^{(-0.00614865*x)} + 213.655$).

The gene-oriented approach consisted of randomly selecting single genes from the original dataset of 293 genomes, and later expanded to 573 genomes or sampling of 508 genomes that excluding highly reduced genomes (Table 2S). Each selected gene was categorized based on the number of genomes with detected homologous sequences (bit score of 50). The number of genes per frequency was normalized to the total number of sampled genes (A*n* = number of genes at Fq(x)/15000). From these frequencies, sampling functions $F(x) = \sum[A_n*e^{(Kn*x)}]$ were reconstructed for the core and rare genes using the number of genes A_n found for each frequency and the decay constant K [Ln(present in *n* other genomes/(total number of genomes - 1))] of each category. The extended core function can be calculated by using the frequency of occurrence of the genes in genomes while the rare gene function uses the frequency of absence of genes. For example, a gene that is present in 90% of the genomes (Fq of presence = 0.9), is absent in 10% of the genome (Fq of absence = 0.1). The average number of ORFs present in the 293 sampled genomes (3053 ORFs) was used to reconstruct the sampling functions (Figure 2S). Because the accuracy of an estimate is greater when more data points are utilized, we used the reconstructed core function to estimate the extended core and the reconstructed rare gene function to estimate the size of the accessory pool. A decay function fitted to the reconstructed sampling curves was used to estimate the size of the three different group of gene and to extrapolate the tendency of the sampling curve as additional genomes are sampled. The resulting decay function best describing the reconstructed data for the extended core genes was found to be (F(x) = $1524.99*e^{(-2.85629*x)} + 740.743*e^{(-0.669252*x)} + 385.758*e^{(-0.188697*x)} + 155.119*e^{(-0.0527433*x)} + 85.7964*e^{(-0.00844936*x)} + 160.59$). The function best describing the reconstructed data for the rare genes was found to be (F(x) = $400.851*e^{(-4.08971*x)} + 415.025*e^{(-1.653*x)} + 380.265*e^{(-0.8651*x)} + 330.216*e^{(-0.4652*x)} + 261.534*e^{(-0.24675*x)} + 208.711*e^{(-0.1324*x)} + 193.966*e^{(-0.0681*x)} + 216.167*e^{(-0.03383*x)} + 190.82*e^{(-0.01444*x)} + 202.914*e^{(-0.004469*x)} + 253.058$). The integral of each component (area under curve) of the decay function (A_n/e^{Kn} for the extended core function or $A_n/1-e^{Kn}$ for the rare gene function) was used to calculate to the number of genes present in each of the components and their expected frequency of occurrence in genomes (e^{Kn} for extended core or $1-e^{Kn}$ for the rare gene function). The extended core and accessory pool are represented by genes present in at least 99% in the core function and absent in 95% or less of the sampled genomes in the rare gene function, respectively.

Table 1S : List of the genomes used in this study.

Acaryochloris marina MBIC11017^{23}	Escherichia coli K12^{123}	Pseudomonas syringae pv B728a^{123}
Acidiphilium cryptum JF-5^{23}	Escherichia coli O157H7^{123}	Pseudomonas syringae tomato DC3000^{123}
Acidobacteria bacterium Ellin345^{23}	Escherichia coli O157H7 EDL933^{123}	Psychrobacter arcticum 273-4^{123}
Acidothermus cellulolyticus 11B^{23}	Escherichia coli UTI89^{23}	Psychrobacter cryohalolentis K5^{23}
Acidovorax avenae citrulli AAC00-1^{23}	Escherichia coli W3110^{123}	Psychrobacter PRwf-1^{23}
Acidovorax JS42^{23}	Fervidobacterium nodosum Rt17-B1^{23}	Psychromonas ingrahamii 37^{23}
Acinetobacter baumannii ATCC 17978^{23}	Flavobacterium johnsoniae UW101^{23}	Ralstonia eutropha H16^{23}
Acinetobacter sp ADP11^{23}	Flavobacterium psychrophilum JIP02 86^{23}	Ralstonia eutropha JMP134^{123}
Actinobacillus pleuropneumoniae L20^{23}	Francisella tularensis FSC 198^{23}	Ralstonia metallidurans CH34^{23}
Actinobacillus succinogenes 130Z^{23}	Francisella tularensis holarctica^{123}	Ralstonia solanacearum^{123}
Aeromonas hydrophila ATCC 7966^{23}	Francisella tularensis holarctica FTA^{23}	Renibacterium salmoninarum ATCC 33209^{23}
Aeromonas salmonicida A449^{23}	Francisella tularensis holarctica OSU18^{23}	Rhizobium etli CFN 42^{123}
Agrobacterium tumefaciens C58^{123}	Francisella tularensis novicida U112^{23}	Rhizobium leguminosarum bv viciae 3841^{23}
Alcanivorax borkumensis SK2^{23}	Francisella tularensis tularensis^{123}	Rhodobacter sphaeroides 2 4 1^{123}
Alkalilimnicola ehrlichei MLHE-1^{23}	Francisella tularensis WY96-3418^{23}	Rhodobacter sphaeroides ATCC 17025^{23}
Alkaliphilus metalliredigens QYMF^{23}	Frankia alni ACN14a^{23}	Rhodobacter sphaeroides ATCC 17029^{23}
Alkaliphilus oremlandii OhILAs^{23}	Frankia CcI3^{123}	Rhodococcus RHA1^{23}
Anabaena variabilis ATCC 29413^{123}	Frankia EAN1pec^{23}	Rhodoferax ferrireducens T118^{123}
Anaeromyxobacter dehalogenans 2CP-C^{123}	Fusobacterium nucleatum^{123}	Rhodopseudomonas palustris BisA53^{23}
Anaeromyxobacter Fw109-5^{23}	Geobacillus kaustophilus HTA426^{123}	Rhodopseudomonas palustris BisB18^{123}
Anaplasma marginale St Maries^{13}	Geobacillus thermodenitrificans NG80-2^{23}	Rhodopseudomonas palustris BisB5^{23}
Anaplasma phagocytophilum HZ^{13}	Geobacter metallireducens GS-15^{123}	Rhodopseudomonas palustris CGA009^{123}
Aquifex aeolicus^{123}	Geobacter sulfurreducens^{123}	Rhodopseudomonas palustris HaA2^{123}
Arcobacter butzleri RM4018^{23}	Geobacter uraniumreducens Rf4^{23}	Rhodospirillum rubrum ATCC 11170^{123}
Arthrobacter aurescens TC1^{23}	Gloeobacter violaceus^{123}	Rickettsia akari Hartford^{3}
Arthrobacter FB24^{23}	Gluconacetobacter diazotrophicus PAl 5^{23}	Rickettsia bellii OSU 85-389^{3}
Aster yellows witches-broom phytoplasma AYWB^{13}	Gluconobacter oxydans 621H^{123}	Rickettsia bellii RML369-C^{3}
Azoarcus BH72^{23}	Gramella forsetii KT0803^{23}	Rickettsia canadensis McKiel^{3}
Azoarcus sp EbN1^{123}	Granulobacter bethesdensis CGDNIH1^{23}	Rickettsia conorii^{13}
Azorhizobium caulinodans ORS 571^{23}	Haemophilus ducreyi 35000HP^{123}	Rickettsia felis URRWXCal2^{13}
Bacillus amyloliquefaciens FZB42^{23}	Haemophilus influenzae^{123}	Rickettsia massiliae MTU5^{3}
Bacillus anthracis Ames^{123}	Haemophilus influenzae 86 028NP^{123}	Rickettsia prowazekii^{13}

Bacillus anthracis Ames 0581^{123}	Haemophilus influenzae PittEE^{23}	Rickettsia rickettsii Sheila Smith^{3}
Bacillus anthracis str Sterne^{123}	Haemophilus influenzae PittGG^{23}	Rickettsia typhi wilmington^{13}
Bacillus cereus ATCC 10987^{123}	Haemophilus somnus 129PT^{23}	Roseiflexus castenholzii DSM 13941^{23}
Bacillus cereus ATCC14579^{123}	Hahella chejuensis KCTC 2396^{123}	Roseiflexus RS-1^{23}
Bacillus cereus cytotoxis NVH 391-98^{23}	Halorhodospira halophila SL1^{23}	Roseobacter denitrificans OCh 114^{23}
Bacillus cereus ZK^{123}	Helicobacter acinonychis Sheeba^{23}	Rubrobacter xylanophilus DSM 9941^{23}
Bacillus clausii KSM-K16^{123}	Helicobacter hepaticus^{123}	Saccharophagus degradans 2-40^{123}
Bacillus halodurans^{123}	Helicobacter pylori 26695^{123}	Saccharopolyspora erythraea NRRL 2338^{23}
Bacillus licheniformis ATCC 14580^{123}	Helicobacter pylori HPAG1^{23}	Salinibacter ruber DSM 13855^{123}
Bacillus licheniformis DSM 13^{123}	Helicobacter pylori J99^{123}	Salinispora arenicola CNS-205^{23}
Bacillus pumilus SAFR-032^{23}	Herminiimonas arsenicoxydans^{23}	Salinispora tropica CNB-440^{23}
Bacillus subtilis^{123}	Herpetosiphon aurantiacus ATCC 23779^{23}	Salmonella enterica arizonae serovar 62 z4 z23 ^{23}
Bacillus thuringiensis Al Hakam^{23}	Hyphomonas neptunium ATCC 15444^{23}	Salmonella enterica Choleraesuis^{123}
Bacillus thuringiensis konkukian^{123}	Idiomarina loihiensis L2TR^{123}	Salmonella enterica Paratypi ATCC 9150^{123}
Bacillus weihenstephanensis KBAB4^{23}	Jannaschia CCS1^{123}	Salmonella enterica serovar Paratyphi B SPB7^{23}
Bacteroides fragilis NCTC 9434^{123}	Janthinobacterium Marseille^{23}	Salmonella typhi^{123}
Bacteroides fragilis YCH46^{123}	Kineococcus radiotolerans SRS30216^{23}	Salmonella typhi Ty2^{123}
Bacteroides thetaiotaomicron VPI-5482^{123}	Klebsiella pneumoniae MGH 78578^{23}	Salmonella typhimurium LT2^{123}
Bacteroides vulgatus ATCC 8482^{23}	Lactobacillus acidophilus NCFM^{123}	Serratia proteamaculans 568^{23}
Bartonella bacilliformis KC583^{23}	Lactobacillus brevis ATCC 367^{23}	Shewanella amazonensis SB2B^{23}
Bartonella henselae Houston-1^{123}	Lactobacillus casei ATCC 334^{23}	Shewanella ANA-3^{23}
Bartonella quintana Toulouse^{123}	Lactobacillus delbrueckii bulgaricus^{23}	Shewanella baltica OS155^{23}
Bartonella tribocorum CIP 105476^{23}	Lactobacillus delbrueckii bulgaricus ATCC BAA-365^{23}	Shewanella baltica OS185^{23}
Baumannia cicadellinicola Homalodisca coagulata^{3}	Lactobacillus gasseri ATCC 33323^{23}	Shewanella baltica OS195^{23}
Bdellovibrio bacteriovorus^{123}	Lactobacillus helveticus DPC 4571^{23}	Shewanella denitrificans OS217^{23}
Bifidobacterium adolescentis ATCC 15703^{23}	Lactobacillus johnsonii NCC 533^{123}	Shewanella frigidimarina NCIMB 400^{23}
Bifidobacterium longum^{123}	Lactobacillus plantarum^{123}	Shewanella loihica PV-4^{23}
Bordetella bronchiseptica^{123}	Lactobacillus reuteri F275^{23}	Shewanella MR-4^{23}
Bordetella parapertussis^{123}	Lactobacillus sakei 23K^{123}	Shewanella MR-7^{23}
Bordetella pertussis^{123}	Lactobacillus salivarius UCC118^{123}	Shewanella oneidensis^{123}
Bordetella petrii^{23}	Lactococcus lactis^{123}	Shewanella pealeana ATCC 700345^{23}
Borrelia afzelii PKo^{23}	Lactococcus lactis cremoris MG1363^{23}	Shewanella putrefaciens CN-32^{23}
Borrelia burgdorferi^{123}	Lactococcus lactis cremoris SK11^{23}	Shewanella sediminis HAW-EB3^{23}
Borrelia garinii PBi^{123}	Lawsonia intracellularis PHE MN1-00^{23}	Shewanella W3-18-1^{23}
Bradyrhizobium BTAi1^{23}	Legionella pneumophila Corby^{23}	Shigella boydii Sb227^{123}
Bradyrhizobium japonicum^{123}	Legionella pneumophila Lens^{123}	Shigella dysenteriae^{123}

Bradyrhizobium ORS278^{23}	Legionella pneumophila Paris^{123}	Shigella flexneri 2a^{123}
Brucella abortus 9-941^{123}	Legionella pneumophila Philadelphia 1^{123}	Shigella flexneri 2a 2457T^{123}
Brucella canis ATCC 23365^{23}	Leifsonia xyli xyli CTCB0^{123}	Shigella flexneri 5 8401^{23}
Brucella melitensis^{123}	Leptospira borgpetersenii serovar Hardjo-bovis JB197^{23}	Shigella sonnei Ss046^{123}
Brucella melitensis biovar Abortus^{123}	Leptospira borgpetersenii serovar Hardjo-bovis L550^{23}	Silicibacter pomeroyi DSS-3^{123}
Brucella ovis^{23}	Leptospira interrogans serovar Copenhageni^{123}	Silicibacter TM1040^{23}
Brucella suis 1330^{123}	Leptospira interrogans serovar Lai^{123}	Sinorhizobium medicae WSM419^{23}
Brucella suis ATCC 23445^{23}	Leuconostoc mesenteroides ATCC 8293^{23}	Sinorhizobium meliloti^{123}
Buchnera aphidicola^{13}	Listeria innocua^{123}	Sodalis glossinidius morsitans^{123}
Buchnera aphidicola Cc Cinara cedri^{3}	Listeria monocytogenes^{123}	Solibacter usitatus Ellin6076^{23}
Buchnera aphidicola Sg^{13}	Listeria monocytogenes 4b F2365^{123}	Sorangium cellulosum So ce 56 ^{23}
Buchnera sp^{13}	Listeria welshimeri serovar 6b SLCC5334^{23}	Sphingomonas wittichii RW1^{23}
Burkholderia 383^{123}	Magnetococcus MC-1^{23}	Sphingopyxis alaskensis RB2256^{23}
Burkholderia cenocepacia AU 1054^{23}	Magnetospirillum magneticum AMB-1^{123}	Staphylococcus aureus aureus MRSA252^{123}
Burkholderia cenocepacia HI2424^{23}	Mannheimia succiniciproducens MBEL55E^{123}	Staphylococcus aureus aureus MSSA476^{123}
Burkholderia cepacia AMMD^{23}	Maricaulis maris MCS10^{23}	Staphylococcus aureus COL^{123}
Burkholderia mallei ATCC 23344^{123}	Marinobacter aquaeolei VT8^{23}	Staphylococcus aureus JH1^{23}
Burkholderia mallei NCTC 10229^{23}	Marinomonas MWYL1^{23}	Staphylococcus aureus JH9^{23}
Burkholderia mallei NCTC 10247^{23}	Mesoplasma florum L1^{13}	Staphylococcus aureus Mu3^{23}
Burkholderia mallei SAVP1^{23}	Mesorhizobium BNC1^{23}	Staphylococcus aureus Mu50^{123}
Burkholderia multivorans ATCC 17616^{23}	Mesorhizobium loti^{123}	Staphylococcus aureus MW2^{123}
Burkholderia pseudomallei 1106a^{23}	Methylibium petroleiphilum PM1^{23}	Staphylococcus aureus N315^{123}
Burkholderia pseudomallei 1710b^{123}	Methylobacillus flagellatus KT^{23}	Staphylococcus aureus NCTC 8325^{123}
Burkholderia pseudomallei 668^{23}	Methylobacterium extorquens PA1^{23}	Staphylococcus aureus Newman^{23}
Burkholderia pseudomallei K96243^{123}	Methylococcus capsulatus Bath^{123}	Staphylococcus aureus RF122^{123}
Burkholderia thailandensis E264^{123}	Moorella thermoacetica ATCC 39073^{123}	Staphylococcus aureus USA300^{123}
Burkholderia vietnamiensis G4^{23}	Mycobacterium avium 104^{23}	Staphylococcus aureus USA300 TCH1516^{23}
Burkholderia xenovorans LB400^{23}	Mycobacterium avium paratuberculosis^{123}	Staphylococcus epidermidis ATCC 12228^{123}
Caldicellulosiruptor saccharolyticus DSM 8903^{23}	Mycobacterium bovis^{123}	Staphylococcus epidermidis RP62A^{123}
Campylobacter concisus 13826^{23}	Mycobacterium bovis BCG Pasteur 1173P2^{23}	Staphylococcus haemolyticus^{123}
Campylobacter curvus 525 92^{23}	Mycobacterium gilvum PYR-GCK^{23}	Staphylococcus saprophyticus^{123}
Campylobacter fetus 82-40^{23}	Mycobacterium JLS^{23}	Streptococcus agalactiae 2603^{123}
Campylobacter hominis ATCC BAA-381^{23}	Mycobacterium KMS^{23}	Streptococcus agalactiae A909^{123}
Campylobacter jejuni^{123}	Mycobacterium leprae^{123}	Streptococcus agalactiae NEM316^{123}

Campylobacter jejuni 81116^{23}	Mycobacterium MCS^{23}	Streptococcus gordonii Challis substr CH1^{23}
Campylobacter jejuni 81-176^{23}	Mycobacterium smegmatis MC2 155^{23}	Streptococcus mutans^{123}
Campylobacter jejuni doylei 269 97^{23}	Mycobacterium tuberculosis CDC1551^{123}	Streptococcus pneumoniae D39^{23}
Campylobacter jejuni RM1221^{123}	Mycobacterium tuberculosis F11^{23}	Streptococcus pneumoniae R6^{123}
Candidatus Blochmannia floridanus^{13}	Mycobacterium tuberculosis H37Ra^{23}	Streptococcus pneumoniae TIGR4^{123}
Candidatus Blochmannia pennsylvanicus BPEN^{13}	Mycobacterium tuberculosis H37Rv^{123}	Streptococcus pyogenes M1 GAS^{123}
Candidatus Carsonella ruddii PV^{3}	Mycobacterium ulcerans Agy99^{23}	Streptococcus pyogenes Manfredo^{23}
Candidatus Pelagibacter ubique HTCC1062^{13}	Mycobacterium vanbaalenii PYR-1^{23}	Streptococcus pyogenes MGAS10270^{23}
Candidatus Ruthia magnifica Cm Calyptogena magnifica ^{3}	Mycoplasma agalactiae PG2^{3}	Streptococcus pyogenes MGAS10394^{123}
Candidatus Sulcia muelleri GWSS^{3}	Mycoplasma capricolum ATCC 27343^{13}	Streptococcus pyogenes MGAS10750^{23}
Candidatus Vesicomyosocius okutanii HA^{3}	Mycoplasma gallisepticum^{13}	Streptococcus pyogenes MGAS2096^{23}
Carboxydothermus hydrogenoformans Z-2901^{123}	Mycoplasma genitalium^{13}	Streptococcus pyogenes MGAS315^{123}
Caulobacter crescentus^{123}	Mycoplasma hyopneumoniae 232^{13}	Streptococcus pyogenes MGAS5005^{123}
Chlamydia muridarum^{13}	Mycoplasma hyopneumoniae 7448^{13}	Streptococcus pyogenes MGAS6180^{123}
Chlamydia trachomatis^{13}	Mycoplasma hyopneumoniae J^{13}	Streptococcus pyogenes MGAS8232^{123}
Chlamydia trachomatis A HAR-13^{13}	Mycoplasma mobile 163K^{13}	Streptococcus pyogenes MGAS9429^{23}
Chlamydophila abortus S26 3^{13}	Mycoplasma mycoides^{13}	Streptococcus pyogenes SSI-1^{123}
Chlamydophila caviae^{13}	Mycoplasma penetrans^{13}	Streptococcus sanguinis SK36^{23}
Chlamydophila felis Fe C-56^{13}	Mycoplasma pneumoniae^{13}	Streptococcus suis 05ZYH33^{23}
Chlamydophila pneumoniae AR39^{13}	Mycoplasma pulmonis^{13}	Streptococcus suis 98HAH33^{23}
Chlamydophila pneumoniae CWL029^{13}	Mycoplasma synoviae 53^{13}	Streptococcus thermophilus CNRZ1066^{123}
Chlamydophila pneumoniae J138^{13}	Myxococcus xanthus DK 1622^{23}	Streptococcus thermophilus LMD-9^{23}
Chlamydophila pneumoniae TW 183^{13}	Neisseria gonorrhoeae FA 1090^{123}	Streptococcus thermophilus LMG 18311^{123}
Chlorobium chlorochromatii CaD3^{123}	Neisseria meningitidis 053442^{23}	Streptomyces avermitilis^{123}
Chlorobium phaeobacteroides DSM 266^{23}	Neisseria meningitidis FAM18^{23}	Streptomyces coelicolor^{123}
Chlorobium tepidum TLS^{123}	Neisseria meningitidis MC58^{123}	Sulfurovum NBC37-1^{23}
Chloroflexus aurantiacus J 10 fl^{23}	Neisseria meningitidis Z2491^{123}	Symbiobacterium thermophilum IAM14863^{123}
Chromobacterium violaceum^{123}	Neorickettsia sennetsu Miyayama^{13}	Synechococcus CC9311^{23}
Chromohalobacter salexigens DSM 3043^{23}	Nitratiruptor SB155-2^{23}	Synechococcus CC9605^{123}
Citrobacter koseri ATCC BAA-895^{23}	Nitrobacter hamburgensis X14^{23}	Synechococcus CC9902^{123}
Clavibacter michiganensis NCPPB 382^{23}	Nitrobacter winogradskyi Nb-255^{123}	Synechococcus elongatus PCC 6301^{123}
Clostridium acetobutylicum^{123}	Nitrosococcus oceani ATCC 19707^{123}	Synechococcus elongatus PCC 7942^{123}
Clostridium beijerinckii NCIMB 8052^{23}	Nitrosomonas europaea^{123}	Synechococcus RCC307^{23}
Clostridium botulinum A^{23}	Nitrosomonas eutropha C71^{23}	Synechococcus sp WH8102^{123}

Clostridium botulinum A ATCC 19397^{23}	Nitrosospira multiformis ATCC 25196^{123}	Synechococcus WH 7803^{23}
Clostridium botulinum A Hall^{23}	Nocardia farcinica IFM10152^{123}	Synechocystis PCC6803^{123}
Clostridium botulinum F Langeland^{23}	Nocardioides JS614^{23}	Syntrophobacter fumaroxidans MPOB^{23}
Clostridium difficile 630^{23}	Nostoc sp^{123}	Syntrophomonas wolfei Goettingen^{23}
Clostridium kluyveri DSM 555^{23}	Novosphingobium aromaticivorans DSM 12444^{123}	Syntrophus aciditrophicus SB^{23}
Clostridium novyi NT^{23}	Oceanobacillus iheyensis^{123}	Thermoanaerobacter tengcongensis^{123}
Clostridium perfringens^{123}	Ochrobactrum anthropi ATCC 49188^{23}	Thermobifida fusca YX^{123}
Clostridium perfringens ATCC 13124^{23}	Oenococcus oeni PSU-1^{23}	Thermosipho melanesiensis BI429^{23}
Clostridium perfringens SM101^{23}	Onion yellows phytoplasma^{13}	Thermosynechococcus elongatus^{123}
Clostridium phytofermentans ISDg^{23}	Orientia tsutsugamushi Boryong^{3}	Thermotoga lettingae TMO^{23}
Clostridium tetani E88^{123}	Parabacteroides distasonis ATCC 8503^{23}	Thermotoga maritima^{123}
Clostridium thermocellum ATCC 27405^{23}	Parachlamydia sp UWE25^{13}	Thermotoga petrophila RKU-1^{23}
Colwellia psychrerythraea 34H^{123}	Paracoccus denitrificans PD1222^{23}	Thermus thermophilus HB27^{123}
Corynebacterium diphtheriae^{123}	Parvibaculum lavamentivorans DS-1^{23}	Thermus thermophilus HB8^{123}
Corynebacterium efficiens YS-314^{123}	Pasteurella multocida^{123}	Thiobacillus denitrificans ATCC 25259^{123}
Corynebacterium glutamicum ATCC 13032 Bielefeld^{123}	Pediococcus pentosaceus ATCC 25745^{23}	Thiomicrospira crunogena XCL-2^{123}
Corynebacterium glutamicum ATCC 13032 Kitasato^{123}	Pelobacter carbinolicus^{123}	Thiomicrospira denitrificans ATCC 33889^{123}
Corynebacterium glutamicum R^{23}	Pelobacter propionicus DSM 2379^{23}	Treponema denticola ATCC 35405^{123}
Corynebacterium jeikeium K411^{123}	Pelodictyon luteolum DSM 273^{123}	Treponema pallidum^{123}
Coxiella burnetii^{123}	Pelotomaculum thermopropionicum SI^{23}	Trichodesmium erythraeum IMS101^{23}
Coxiella burnetii Dugway 7E9-12^{23}	Petrotoga mobilis SJ95^{23}	Tropheryma whipplei TW08 27^{13}
Coxiella burnetii RSA 331^{23}	Photobacterium profundum SS9^{123}	Tropheryma whipplei Twist^{13}
Cyanobacteria bacterium Yellowstone A-Prime^{123}	Photorhabdus luminescens^{123}	Ureaplasma urealyticum^{13}
Cyanobacteria bacterium Yellowstone B-Prime^{123}	Pirellula sp^{123}	Verminephrobacter eiseniae EF01-2^{23}
Cytophaga hutchinsonii ATCC 33406^{23}	Polaromonas JS666^{23}	Vibrio cholerae^{123}
Dechloromonas aromatica RCB^{123}	Polaromonas naphthalenivorans CJ2^{23}	Vibrio cholerae O395^{23}
Dehalococcoides BAV1^{23}	Polynucleobacter QLW-P1DMWA-1^{23}	Vibrio fischeri ES114^{123}
Dehalococcoides CBDB1^{123}	Porphyromonas gingivalis W83^{123}	Vibrio harveyi ATCC BAA-1116^{23}
Dehalococcoides ethenogenes 195^{123}	Prochlorococcus marinus AS9601^{23}	Vibrio parahaemolyticus^{123}
Deinococcus geothermalis DSM 11300^{23}	Prochlorococcus marinus CCMP1375^{123}	Vibrio vulnificus CMCP6^{123}
Deinococcus radiodurans^{123}	Prochlorococcus marinus MED4^{123}	Vibrio vulnificus YJ016^{123}
Delftia acidovorans SPH-1^{23}	Prochlorococcus marinus MIT 9211^{23}	Wigglesworthia brevipalpis^{13}
Desulfitobacterium hafniense	Prochlorococcus marinus MIT	Wolbachia endosymbiont of

Y51[123]	9215[23]	Brugia malayi TRS[13]
Desulfotalea psychrophila LSv54[123]	Prochlorococcus marinus MIT 9301[23]	Wolbachia endosymbiont of Drosophila melanogaster[13]
Desulfotomaculum reducens MI-1[23]	Prochlorococcus marinus MIT 9303[23]	Wolinella succinogenes[123]
Desulfovibrio desulfuricans G20[123]	Prochlorococcus marinus MIT 9312[123]	Xanthobacter autotrophicus Py2[23]
Desulfovibrio vulgaris DP4[23]	Prochlorococcus marinus MIT 9515[23]	Xanthomonas campestris[123]
Desulfovibrio vulgaris Hildenborough[123]	Prochlorococcus marinus MIT9313[123]	Xanthomonas campestris 8004[123]
Dichelobacter nodosus VCS1703A[23]	Prochlorococcus marinus NATL1A[23]	Xanthomonas campestris vesicatoria 85-10[123]
Dinoroseobacter shibae DFL 12[23]	Prochlorococcus marinus NATL2A[123]	Xanthomonas citri[123]
Ehrlichia canis Jake[13]	Propionibacterium acnes KPA171202[123]	Xanthomonas oryzae KACC10331[123]
Ehrlichia chaffeensis Arkansas[13]	Prosthecochloris vibrioformis DSM 265[23]	Xanthomonas oryzae MAFF 311018[23]
Ehrlichia ruminantium Gardel[13]	Pseudoalteromonas atlantica T6c[23]	Xylella fastidiosa[123]
Ehrlichia ruminantium str. Welgevonden[13]	Pseudoalteromonas haloplanktis TAC125[123]	Xylella fastidiosa Temecula1[123]
Ehrlichia ruminantium Welgevonden[13]	Pseudomonas aeruginosa[123]	Yersinia enterocolitica 8081[23]
Enterobacter 638[23]	Pseudomonas aeruginosa PA7[23]	Yersinia pestis Angola[23]
Enterobacter sakazakii ATCC BAA-894[23]	Pseudomonas aeruginosa UCBPP-PA14[23]	Yersinia pestis Antiqua[23]
Enterococcus faecalis V583[123]	Pseudomonas entomophila L48[23]	Yersinia pestis biovar Mediaevails[123]
Erwinia carotovora atroseptica SCRI1043[123]	Pseudomonas fluorescens Pf-5[123]	Yersinia pestis CO92[123]
Erythrobacter litoralis HTCC2594[123]	Pseudomonas fluorescens PfO-1[123]	Yersinia pestis KIM[123]
Escherichia coli 536[23]	Pseudomonas mendocina ymp[23]	Yersinia pestis Nepal516[23]
Escherichia coli APEC O1[23]	Pseudomonas putida F1[23]	Yersinia pestis Pestoides F[23]
Escherichia coli CFT073[123]	Pseudomonas putida KT2440[123]	Yersinia pseudotuberculosis IP 31758[23]
Escherichia coli E24377A[23]	Pseudomonas stutzeri A1501[23]	Yersinia pseudotuberculosis IP32953[123]
Escherichia coli HS[23]	Pseudomonas syringae phaseolicola 1448A[123]	Zymomonas mobilis ZM4[123]

[1] 293 genomes sampling dataset

[2] 508 genomes sampling dataset

[3] 573 genomes sampling dataset

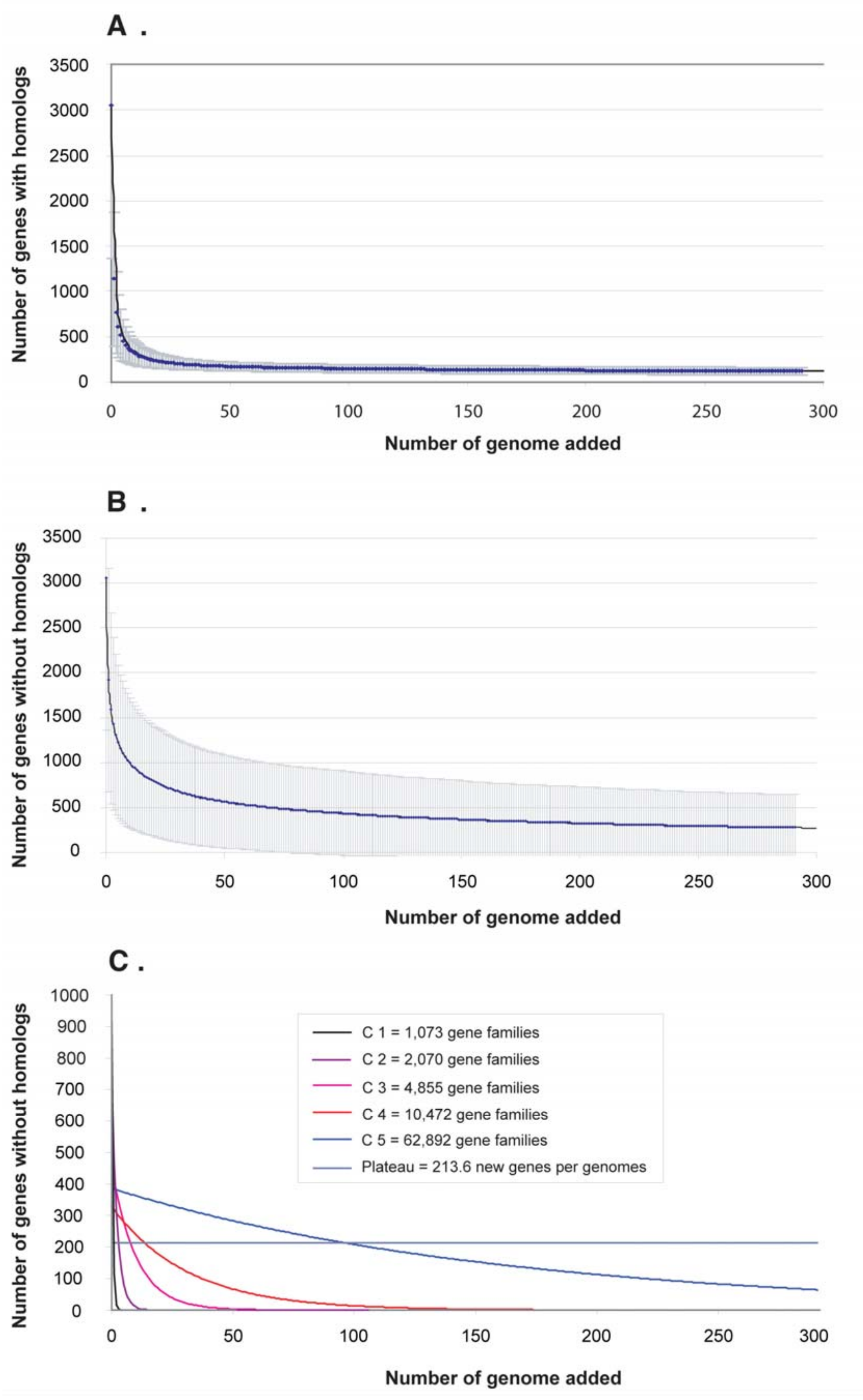
A .
Number of genes with homologs
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
Number of genome added
B .
Number of genes without homologs
3500
3000
2500
2000
1500
1000
500
0
0
50
100
150
200
250
300
Number of genome added
C .
Number of genes without homologs
1000
900
800
700
600
500
400
300
200
100
0
C 1 = 1,073 gene families
C 2 = 2,070 gene families
C 3 = 4,855 gene families
C 4 = 10,472 gene families
C 5 = 62,892 gene families
Plateau = 213.6 new genes per genomes
0
50
100
150
200
250
300
Number of genome added

Figure 1S. Genome-centered approach to assess the Bacterial Core- and *Pan-genomes*. (A) Average number of shared genes and (B) unique genes in the starting genome as the number of sampled genomes increases up to 293 genomes. The sampling at point zero corresponds to the average genome size (3053 ORFs) of the all sampled genomes. As more genomes are sampled, the numbers of genes from the starting genome that are shared by all genomes (in A) or that are unique in the starting genome (in B) is getting smaller. (C) A representation of the five decay components of the sampling function of unique genes. The area under the curve ($A_n/1–e^{Kn}$) allows the estimation the total number of genes families present in each component (Insert C1 through C5). The existence of a plateau in (B) reveals an open pan-genome.

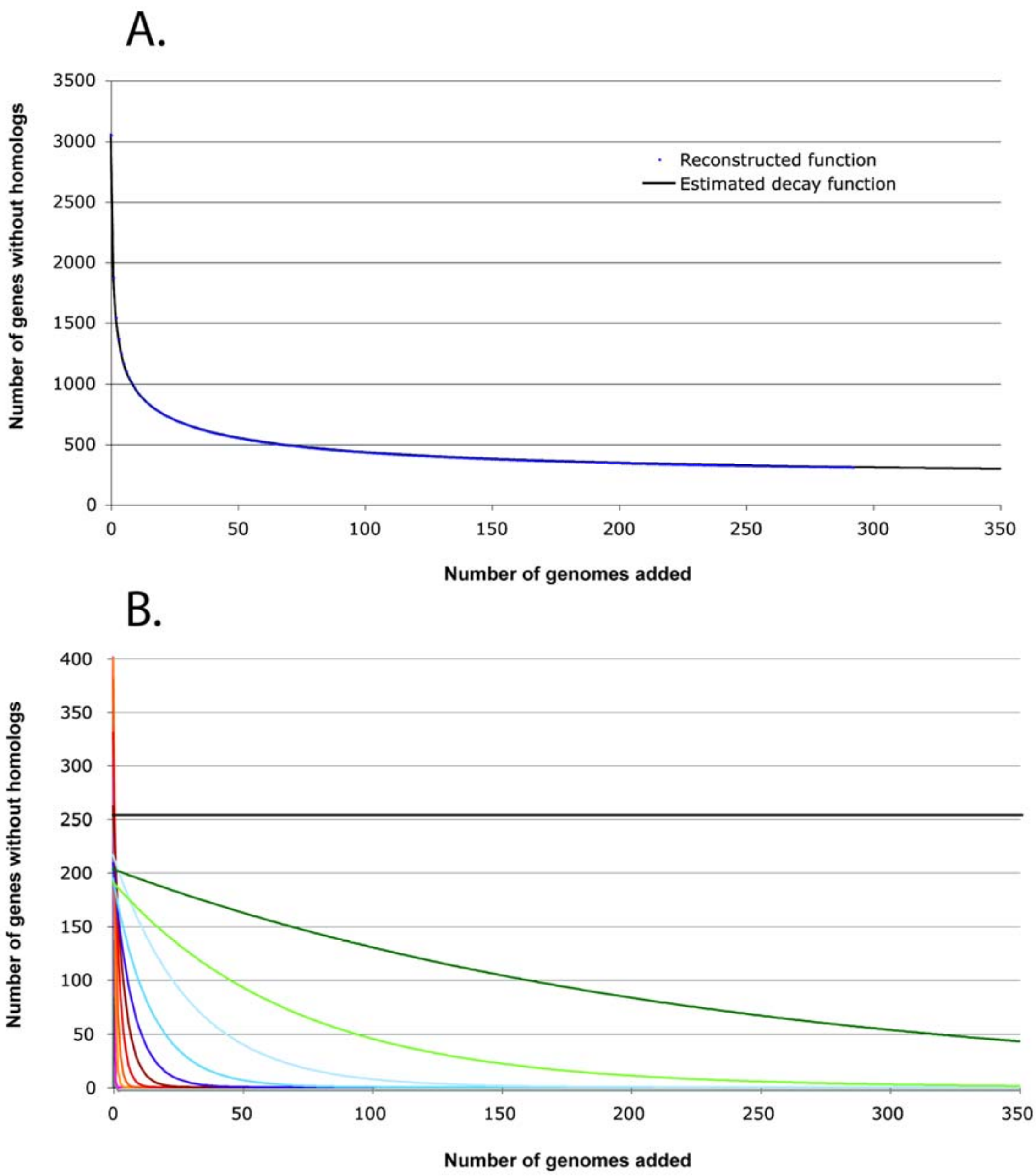

Figure 2S. Gene-centered approach to assess the Bacterial Core- and *Pan-genomes*. (A) The reconstructed a sampling function using the frequency of occurrence of rare genes among the 293 sampled genomes ($F(x) = \sum[A_n*e^{(Kn*x)}]$). The decay functions best describing the reconstructed data was ($F(x) = 400.851*e^{(-4.08971*x)} + 415.025*e^{(-1.653*x)} + 380.265*e^{(-0.8651*x)} + 330.216*e^{(-0.4652*x)} + 261.534*e^{(-0.24675*x)} + 208.711*e^{(-0.1324*x)} +$

$193.966*e^{(-0.0681*x)} + 216.167*e^{(-0.03383*x)} + 190.82*e^{(-0.01444*x)} + 202.914*e^{(-0.004469*x)} +$ 253.058). (B) Representation of the ten individual components with plateau of the decomposed decay function. The integral of each component (area under curve) of the decay function ($A_n/1-e^{Kn}$) was used to calculate to the number of genes present in the accessory pool and their expected frequency of occurrence in genomes ($1-e^{Kn}$).

Table 2S. The estimated contributions of extended core -, character -, and accessory genes to the average bacterial genome using the genome and the gene centered approach. Sampling of 293 bacterial genomes was performed using both approaches. The Gene centered approach was later expanded to 573 completely sequenced bacterial genomes and to a set of 508 genomes that excluded parasitic species and other highly reduced genomes.

	Genome Centered Approach	Gene Centered Approach			Average Gene Centered
	293 Genomes	293 Genomes	573 Genomes	508 Genomes	
Extended Core	6.8%	8%	5.6%	9.5%	**7.7%**
Character Genes	62.7%	63.7%	64.8%	61%	**63.2%**
Accessory Genes	30.4%	28.2%	29.6%	29.5%	**29.1%**

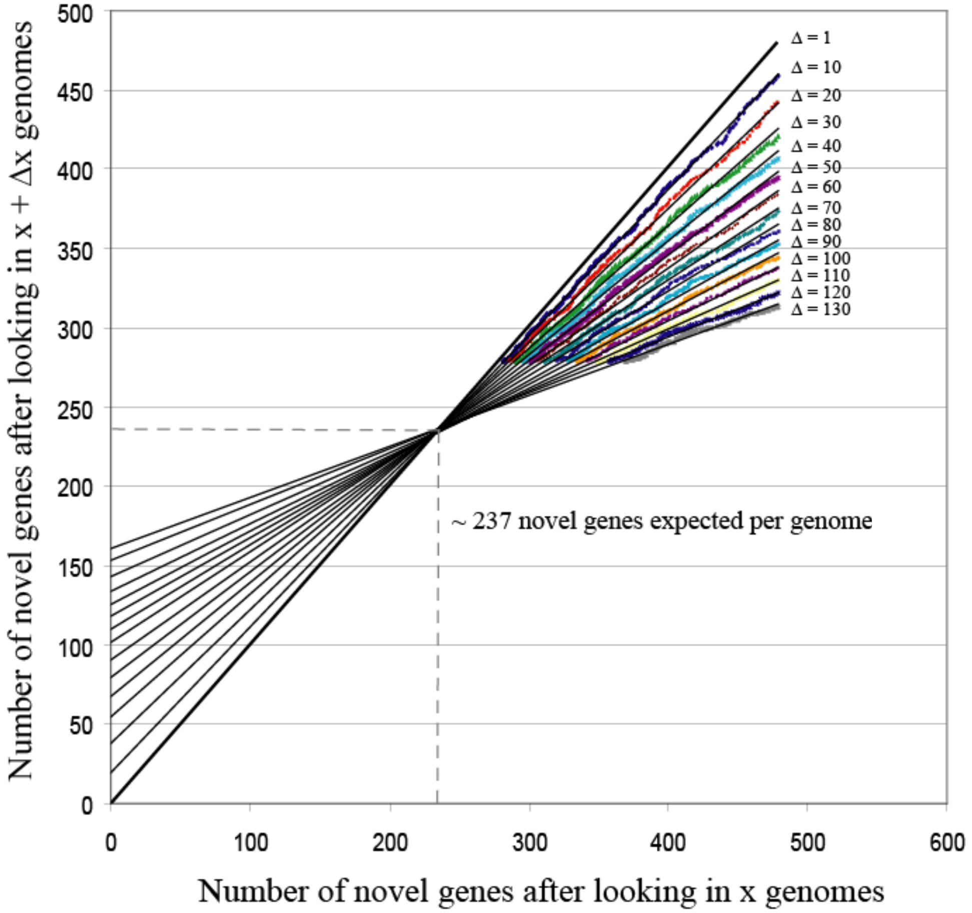

Figure 3S. Kezdy-Swinbourne plot for the number of unique genes. A Kézdy-Swinbourne (KS) plot estimates the value that a decay function (f(x)) approaches as x goes to infinity (Hiromi, K. 1979. *Kinetics of Fast Enzyme Reactions*. Halsted Press (Wiley), New York)). In a KS-plot the value of the function at point x +Δx is plotted against the value at point x. As x goes to infinity both f(x) and f(x+Δx) approach the same limit independent of the choice of Δx. The principle of the KS plot can be explained as follows: Assume the simple decay function $f(x) = K + A \bullet e^{-k\bullet x}$ (eq. 1), then $f(x + \Delta x) = K + A \bullet e^{-k\bullet(x+\Delta x)}$ (eq. 2). Through elimination of the constant A (solving eq. 1 for A and inserting into eq. 2): $f(x+\Delta x) = e^{-k \bullet \Delta x} \bullet f(x) + K'$ I.e., for a simple decay

function the plot of $f(x+\Delta x)$ against $f(x)$ results in a straight line with slope $e^{-k \cdot \Delta x}$. For $x \to \infty$ both $f(x)$ and $f(x+\Delta x)$ approach the same constant: $f(x) \to K$, $f(x+\Delta x) \to K$. Therefore, if one plots the value of a decay function at $x + \Delta x$ against the value at x, the plots result in straight lines. The lines obtained for different Δx all intersect the $x = y$ line at the same point K. Here we apply this method to the number of genes without detectable homologs after comparing the starting genome to *x* other genomes using genome-oriented method. The Kézdy-Swinbourne Plot is rather insensitive to deviations from a simple single component decay function. We only plot values obtained for more than 80 genomes sampled (i.e. after the faster components have already decayed).

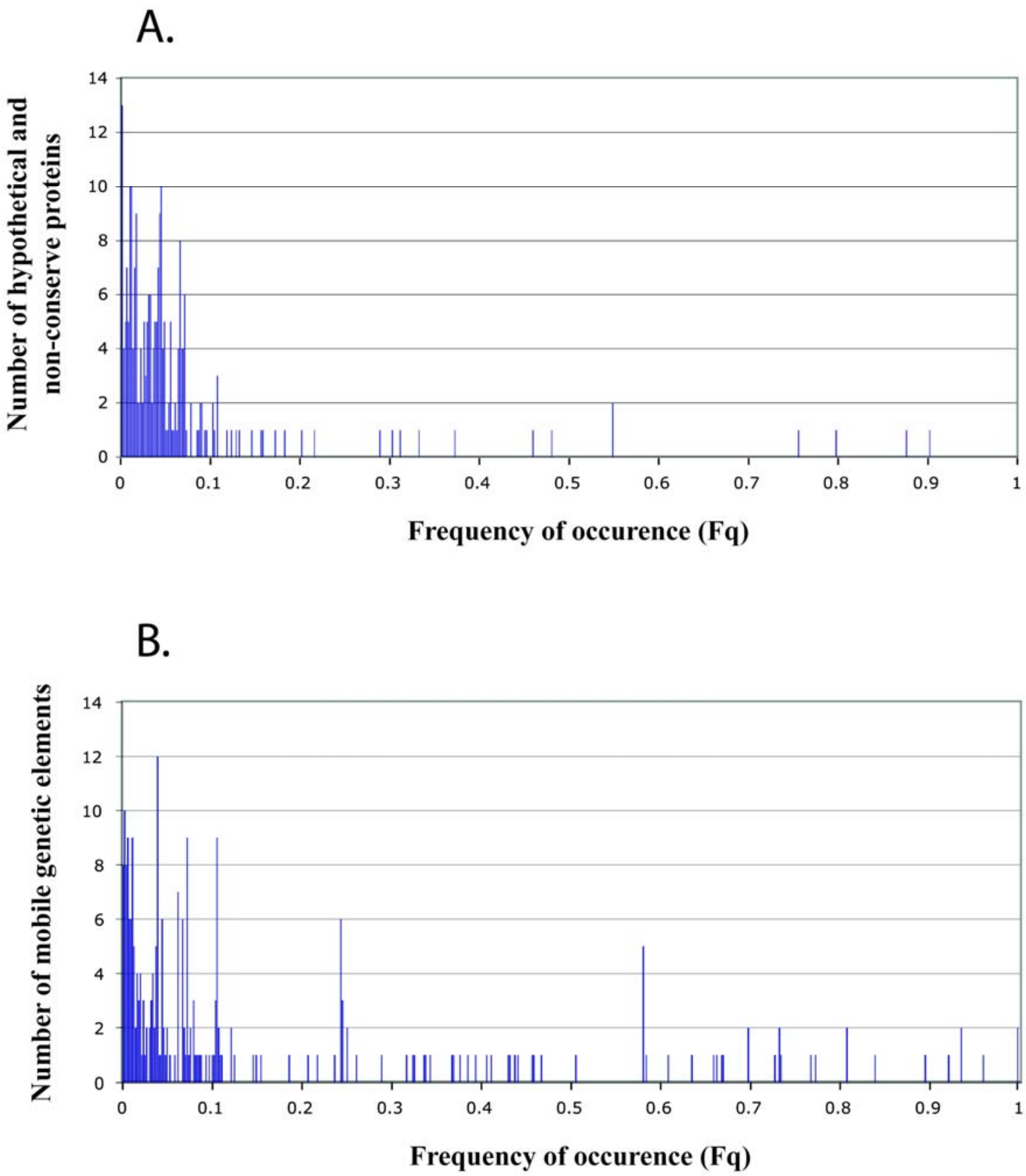

Figure 4S. Frequency of occurrence in genomes of hypothetical/non-conserve proteins and mobile genetic elements. Frequency distribution among genomes of genes present in *E. coli* K12 annotated as hypothetical or non-conserved proteins (A) and mobile genetic elements (B).

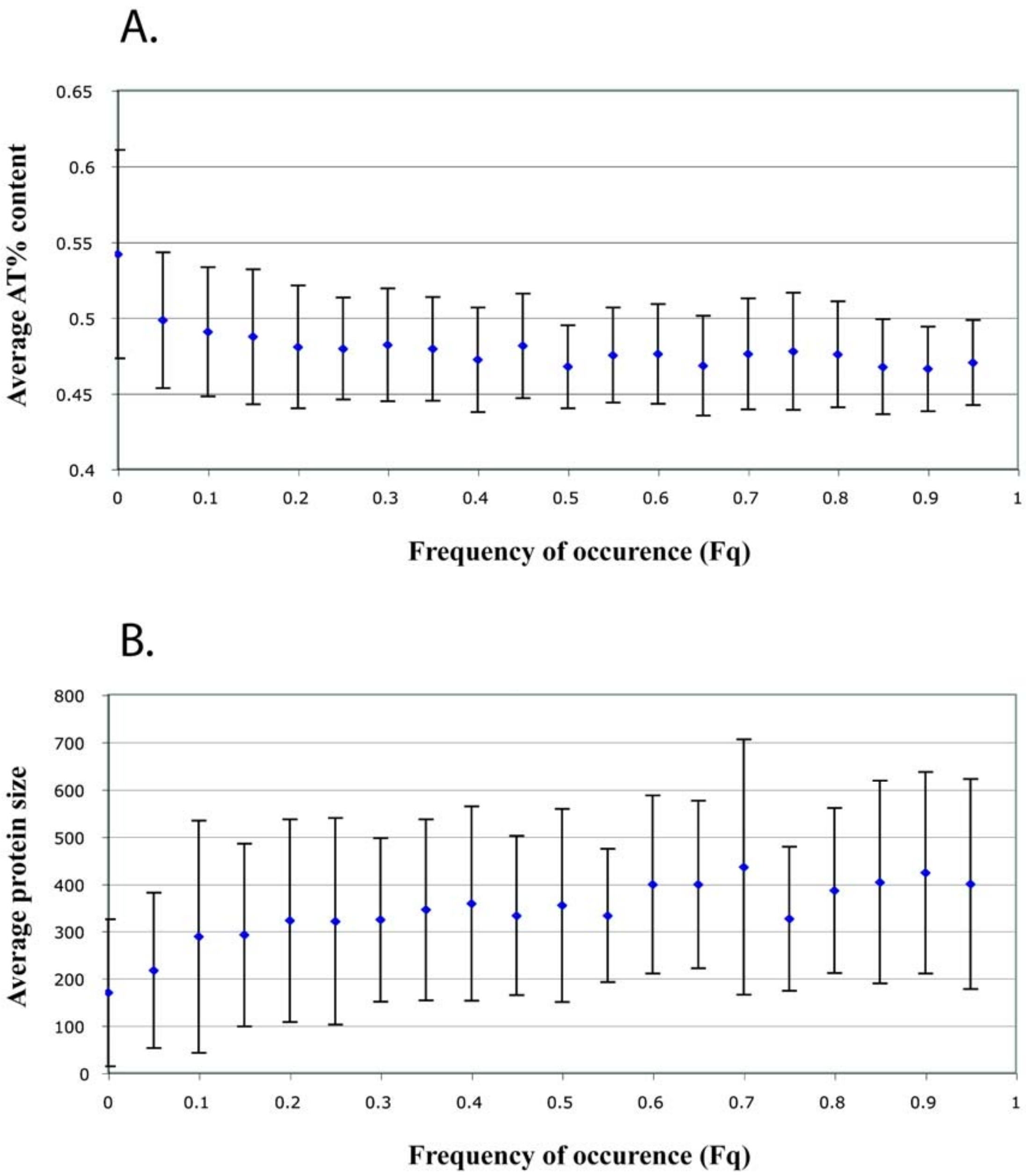

Figure 5S. Average percent AT content and protein size in relation to frequency of occurrence among genomes. Survey of the average AT% (A) and amino acids length (B) of all ORFs present in the genome of *E. coli* K12. Each value corresponds to the average number of proteins using a bins size frequency of 0.5. The first value at 0 frequency represent the average number of proteins found at a frequency [0, <0.5].