J Mol Evol (2006) xx:1–14
DOI: 10.1007/s00239-005-0250-9

JOURNAL OF MOLECULAR EVOLUTION

Computational Prediction of Genomic Functional Cores Specific to Different Microbes

Alessandra Carbone

Génomique Analytique, Université Pierre et Marie Curie-Paris 6, INSERM U511, 91, Bd de I'Hôpital, 75013 Paris, France

Received: 20 October 2005 / Accepted: 10 July 2006 [*Reviewing Editor:* Martin Kreitman]

Abstract. Computational and experimental attempts tried to characterize a universial core of genes representing the minimal set of functional needs for an organism. Based on the increasing number of available complete genomes, comparative genomics has concluded that the universal core contains <50 genes. In contrast, experiments suggest a much larger set of essential genes (certainly more than several hundreds, even under the most restrictive hypotheses) that is dependent on the biological complexity and environmental specificity of the organism. Highly biased genes, which are generally also the most expressed in translationally biased organisms, tend to be over represented in the class of genes deemed to be essential for any given bacterial species. This association is far from perfect; nevertheless, it allows us to propose a new computational method to detect, to a certain extent, ubiquitous genes, nonorthologous genes, environment-specific genes, genes involved in the stress response, and genes with no identified function but highly likely to be essential for the cell. Most of these groups of genes cannot be identified with previously attempted computational and experimental approaches. The large variety of life-styles and the unusually detectable functional signals characterizing translationally biased organisms suggest using them as reference organisms to infer essentiality in other microbial species. The case of small parasitic genomes is discussed. Data issued by the analysis are compared with previous computational and experimental studies. Results are discussed both on methodological and biological grounds.

Correspondence to: A. Carbone; *email:* alessandra.carbone@lip6.fr

Key words: Codon bias — Codon Adaptation Index — Evolution — Genomic core — Minimal gene sets

Introduction

The notion of a minimal gene set, or genomic core, was introduced with the early development of comparative genomics, beginning with comparison (Mushegian & Koonin 1996) of the two sequenced genomes, *Mycoplasma genitalium* (Fraser et al. 1995) and *Haemophilus influenzae* (Fleischmann et al. 1995), aimed to identify a small set of genes common to all genomes that describe the minimal set of functional needs for an organism. Computational attempts to define the universal core based on the increasing number of available complete genomes and experimental attempts to determine the number of essential genes (for which disruption implies lethality) are listed in Tables 1 and 2. The increasing number of essential genes provided by experiments and the decreasing number of shared genes suggested by computational analysis indicate that the notion of a minimal gene set should account for several factors, such as the complexity of the metabolic machinery and the specificity of the living environment. Computations appear to underestimate the minimal gene set by taking into account only those genes that have remained similar enough during the course of evolution to be recognized as true orthologues. In contrast, for a substantial number of essential functions,

Table 1. Detection of essential genes by comparative genomics

No. of organisms	No. of homologous genes	References
2	256	Mushegian & Koonin, 1996
4 (euryarchaea)	543	Makarova et al. 2003
4 (euryarchaea)	521	Nesbø et al. 2001
34	80	Harris et al. 2003
45	23 (strong requirements)	Brown et al. 2001
100	60	Koonin 2003
147	35	Charlebois & Doolittle 2004

different organisms use genes that are not orthologues and in some cases, not even homologues (Koonin 2003). Experimental approaches have other limitations: (1) transposon mutagenesis might overestimate the minimal set by misclassifying nonessential genes that slow down growth without arresting it (for instance, ribosomal protein L24 has been detected as essential for *Escherichia coli* [Gerdes et al. 2003; Hashimoto et al. 2005] but it has been found not absolutely essential in [Nishi et al. 1985]), and it can miss essential genes that tolerate transposon insertions (Akerley et al. 2002; Gerdes et al. 2002); (2) the use of antisense RNA is limited to genes for which an adequate expression of inhibitory RNA can be obtained in the organism under study; and (3) systematic inactivation has the drawback that each gene is inactivated singly. All of these methods leave open the possibility that a different gene set is essential under different growth conditions.

We aimed to detect a set of essential genes under a variety of growth conditions and to relax the notion of essentiality so that disruption of essential genes implies a devastating life, possibly, but not necessarily leading to lethality. We searched for several *functional genomic cores*, one for each organism, rather than for a single minimal gene set fitting several organisms. The method we introduce here suggests that core genes are the most biased genes in translationally biased genomes. In fact, for these genomes, highly biased genes are also the most expressed (Grantham et al. 1980; Sharp & Li 1987) and tend to be over represented in the class of genes deemed to be essential for any given bacterial species. Using a tool designed for an automatic large-scale analysis of codon bias in genomes based on no previous biological knowledge of the organism (Carbone et al. 2003), we identified core genes as having a high Self Consistent Codon Index (SCCI) (this notion, in translationally biased organisms, corresponds to the Codon Adaptation Index [CAI] Sharp & Li 1987; Carbone et al. 2003). It has been demonstrated that evolutionary signals detected from codon bias are highest for translationally biased organisms and can be used to identify important metabolic pathways (Carbone & Madden 2005). Based on this fact and on the possibility of detecting translational bias for an organism by computational means through two statistical criteria applied to genome sequences (Carbone et al. 2004), we considered 27 translationally biased microbial organisms (25 bacteria and 2 archaea); derived the corresponding functional genomic cores; and analyzed and compared functional cores. We demonstrated that genes occurring in functional cores are informational genes (involved in transcription, translation, recombination, repair, replication, secretion, signaling, and cell envelope), operational genes (involved in energy production and conversion, amino acid and nucleotide metabolism, and carbohydrate transport and metabolism), and operational genes specific to physiological and environmental factors, including stress-response genes. Some functionally uncharacterized genes were also detected. Genes participating in basic cellular activity are generally found throughout all species (i.e., ubiquitous) and have been detected in previous computational and experimental studies. They might be non-orthologous and shared by different organisms (e.g., the ribonuclease HI family gene *rnhA* occurs in the functional core of *Synechocystis*, and the ribonuclease HII family gene *rnhB* occurs in the functional cores of *Methanosarcina acetivorans* and *Streptococcus agalactiae*); although these genes can be detected experimentally, they cannot be found by comparative genomics. Genes sustaining life in specific environmental conditions have been detected through experiments conducted on a few species and on a limited pool of living environments (Table 2), but comparative genomics misses them all. Among the most interesting findings, it is worth noticing that translationally biased genome of *Mycoplasma tuberculosis* displays a high SCCI value for genes involved in chorismate biosynthesis and pyrodoxal 5′ phosphate biosynthesis, suggesting essentiality of these two pathways (Carbone & Madden 2005), which are unimportant for most bacteria.

Issues pertaining to functional cores being artifactually small because of the ubiquity requirement and problems related to gene divergence beyond detectability do not influence our method. To uniformly compare functional genomic cores across microbial species, for each organism we considered 200 core genes, including divergent proteins, even if the actual number of core genes was much larger, as discussed later.

Functional signals detected from statistical analysis are more easily identifiable for translationally biased than for other organisms (Carbone & Madden 2005), and the large spread of their life-styles and living environments (Carbone et al. 2004) makes them obvious reference models for this analysis. The 27

Table 2. Detection of essential genes by experiments

Organisms	Technique	Essential genes	Genes considered	References
B. subtilis	Chromosomal mutagenesis	≈300	≈4000	Itaya 1995
B. subtilis	Systematic inactivation	271	≈4100	Kobayashi et al. 2003
M. genitalium	Transposon mutagenesis	265	482	Hutchison et al. 1999
M. genitalium	Transposon mutagenesis	382	482	Glass et al. 2006
H. influenzae	Transposon mutagenesis	670	1272	Akerley et al. 2002
E. coli	Transposon mutagenesis	620	3746	Gerdes et al. 2003
E. coli	Transposon mutagenesis	234	2994	Hashimoto et al. 2005
H. pylori	Transposon mutagenesis	326	1491	Salama et al. 2004
S. aureus	Antisense RNA	150	482	Ji et al. 2001; Forsyth et al. 2002
S. pneumoniae	High throughput gene disruption	113	347	Thanassi et al. 2002
S. cerevisiae	Systematic inactivation	406	2026	Winzeler et al. 1999
S. cerevisiae	Systematic inactivation	1105	5916	Giaever et al. 2002
C. elegans	RNA interference	1722	19427	Kamath et al. 2003

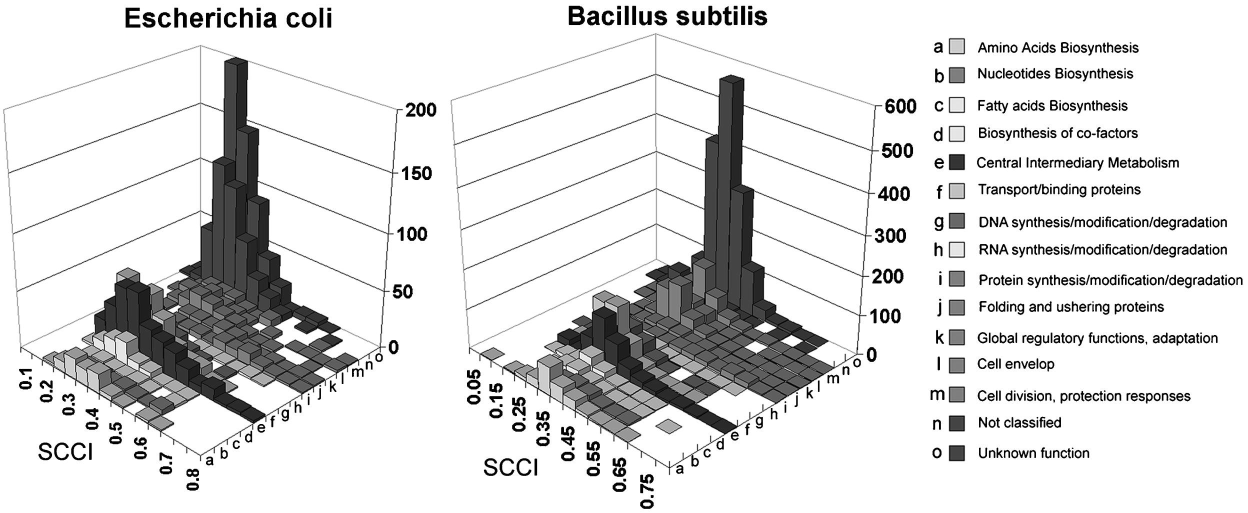

Fig. 1. *SCCI* values and number of coding sequences are plotted for each functional category of *E. coli* and *B. subtilis*. Histograms follow a long tail distribution within all functional classes with the exception of the protein synthesis category (row "i"), which displays two pronounced peaks.

organisms we considered belong to a large variety of phylogenetic taxa, γ and δ proteobacteria, firmicutes, actinobacteria, thermococcales, and methanosarcinales. They do not display strong GC or AT content (see Figure 1 in [Carbone et al. 2004]) and they are characterized by different optimal growth temperatures (Carbone et al. 2004). The well-structured organization of core genes suggests how to infer functional cores for organisms that are not translationally biased and where biological signals are much weaker: such organisms are indeed expected to have the same set of ubiquitous core genes, a pool of nonorthologous genes, and a remaining set of core genes for surviving in a specific living environment. Nonorthologous core genes related to physiology and environmental genes are expected to be shared with translationally biased organisms satisfying similar living conditions. Also, recently transferred genes, which might be possibly vital to a recipient's current activities but that have not yet "ameliorated" within its genome so as to be detected by statistical analysis, are expected to appear as functional core genes in some other translationally biased genome, and this will identify them with high probability as potential core genes.

It is plausible to think of genes that are fundamental to the survival of an organism and whose products are needed only in small amounts. These genes will likely not be detected by our methodology (their SCCI value is expected to be low), but we tried to show, by comparing core sets with genomes of small sizes, that such missed genes are not many. Indeed, we found that approximately half of genes in parasitic genomes of very small sizes, such as *Buchnera aphidicola*, are core genes in some phylogenetically close translationally biased genome like *E. coli*, and that more than two-thirds of *B. aphidicola* genes with homologs in *M. genitalium* are core genes in *E. coli*. Because the ongoing further reduction of *B. aphidicola*, the number of essential genes missed by our method appears to be moderate. This observation is also positively supported by the analysis of firmicutes and γ-proteobacteria persistent genes (that is genes that are preserved in the genomes of the two species) which are asserted to be highly biased in (Fang et al. 2005) and by experimental evidence reported in (Winzeler et al. 1999) where for >99% of the 406 essential genes in *S. cerevisiae*, transcripts were detected, and the average number of these transcripts was 70% higher than for all unessential genes. Even if, in general, there is limited correlation between mRNA and protein expression levels, the amount of experimental noise is smaller and the correlation higher for highly expressed genes (Gygi et al. 1999), i.e., those relevant to our analysis.

Our interest in this study was not in detecting a genomic core to deduce the composition of ancestral genomes (Mushegian & Koonin 1996; Koonin 2003), facilitate reconstruction of phylogenetic trees (Makarova et al. 2003; Nesbø et al. 2001; Daubin et al. 2002; Lerat et al. 2003) or to address questions on species comparison based on codon bias ([Carbone et al. 2004; Carbone & Madden 2005], see also [Kreil & Ouzounis 2001; Lynn et al. 2002; Tekaia et al. 2002; Sharp et al. 2005] for a similar analysis pursued on codon usage and amino-acid use). We wished to test the evolutionary hypothesis that most essential genes in microbial organisms have high SCCI and we wished to extract biological information on gene functional classification that hopefully will be useful to the working biologist. Depending on our knowledge of the organism in question, the number of uncharacterized genes we detect as essential might be very large, and functional cores might furnish new biological insights. The gathered data can provide new information to attain genome minimization conditioned by specific environmental conditions and metabolic activities (Venter et al. 2003; Zimmer 2003; Smith et al. 2003). The systematic study of microbial differences from functional genomic cores might play a crucial role in the identification of specific molecules targeting cohabiting microbial species, as well as the identification of good growth conditions for *in vitro* culture.

Materials and Methods

Organisms and Genomes

Genomes and gene annotation were retrieved from the Genomes Directory of GenBank via file transfer protocol. All coding sequences (CDSs) were considered, including those annotated as hypothetical and those predicted by computational methods only.

Gene Classification

Genes were classified according to the initial version of Clusters of Orthologous Groups (COG) classification (NCBI www.ncbi.nlm.nih.gov/COG). To the COG classes we added two more, one to collect phage-related proteins and the other for virulence. Genes classified in the COG database as having a general predicted function were listed separately and were classified, whenever possible, in COG classes (see Supplementary Table 3) by hand. Whenever this manual classification was too uncertain, we listed the gene as unknown.

Homologous Genes

Lists of homologous genes shared by pairs of bacteria were taken from Genplot (www.ncbi.nlm.nih.gov/). Detection of homologous genes is nonsymmetric, and we consider the list of genes of the smallest genome against the larger one.

Essential Genes

Lists of experimentally identified essential genes for several species is available at tubic.tju.edu.cn/deg/. The Profiling of *E. coli* Chromosome (PEC) database provides a list of essential and nonessential genes in *E. coli* that were identified after genome minimization; see www.shigen.nig.ac.jp/ecoli/pec/index.jsp for data and classification criteria.

Genes Involved in Stress Response

We used the list of 296 *E. coli* genes related to repair, recombination, and physiological adaptations to different stresses compiled by Rocha et al. (2002) and available at www.abi.snv.jussieu.fr/people/erocha/stress/index.html.

Translational Bias

Translational selection refers to the benefit of an increased translational output for a fixed investment in the translational machinery (ribosomes, tRNA, elongation factors, etc.) if only a subset of codons (and their corresponding tRNAs) are used preferentially. Because the benefit of using a particular codon depends on how often it is translated, the strength of translational selection, and hence the degree of codon bias, is expected to vary with the expression level of a gene within an organism. Mutational bias (e.g., an excess or deficit of GC content compared to AT content) might obscure translational selection, which can appear in strong or weak forms (see later).

Calculation of the Self Consistent Codon Index

Sharp (Sharp & Li 1987) formulated the hypothesis that for translationally biased genomes G, there is a *reference set* S of coding sequences, constituting approximately the 1% of the genes in G that are representative of codon adaptation in G. This bias can be described by listing a set of codon weights calculated on genes in S: Given an amino acid j, its synonymous codons might have different frequencies in S; if x_{ij} is the number of times that the codon i for the amino acid j occurs in S, then one associates to i a weight w_{ij} relative to its sibling of maximal frequency y_j in S,

$$w_{ij} = x_{ij}\Big/y_j. \tag{1}$$

Such weights describe codon preferences in G, and they were successfully used by Sharp to correlate gene expression levels to translational codon bias in fast-growing organisms. To do this, one computes the Codon Adaptation Index (CAI) (Sharp & Li 1987) for all genes, CAI $(g) = \left(\prod_{k=1}^{L} w_k\right)^{1/L}$, where g is a gene, w_k is the weight of the k-th codon in g, L is the number of codons in g; and where the reference set S is manually defined as the set of genes

Table 3. List of functional classes represented in the functional core of at least nine organisms

Functional classes	Aci	Bha	Bsu	Bth	Bba	Cdi	Efa	Eca	Eco	Hin	Lpl	Lla	Mac	Pmu	Plu	Pab	Sty	Sat	Son	Sfl	Sag	Smu	Spn	Spy	Syn	Vch	Ype
INFORMATION STORAGE AND PROCESSING																											
J Translation and associated functions																											
Ribosomal proteins (including subunits)*	49	65	48	34	11	49	49	41	45	46	50	53	39	51	47	49	47	46	48	44	52	46	51	52	22	53	51
Elongation factors: *tufA, TufB, efp, typA, fus*	5	4	4	4	1	4	4	5	5	5	4	4	2	5	5	3	5	5	6	5	5	4	6	5	3	7	3
Initiation factors: *infA, infB, infC, eif, fam, aif fam*	2	2	1	1	3	1		1	1	2	1	1	2	2	2	3	1	1	1	1	2	2		2		2	2
Aminoacyl transferRNA synthetases*	1		2	13	5	5	5	7	9	6	6	8	6	6	9	5	7	7	7	11	6	7	11	9		10	2
Polyribonucleotide and tRNA nucleotidyltransferase: *pnp, cca, rrp42*	1	1			1		1	1	1	1		1	2	1	1		1	1	1	1			1			1	1
Ribosome recycling/releasing/binding factors: *frr, rrf, rbfA*	1		1	1			1		2	1	1			1	2		1	1	1	1	2	1	1	1		1	1
K Transcription																											
Cold shock proteins: *csp fam*	2	1	3	5		1	1	2	3		3	2		2	3		3	3	2	3	1			1		3	7
RNA polymerase; *rpo fam*	3	1	4	1	1	3	3	4	5	3	5	5	2	4	4	4	3	3	3	5	6	4	4	5	5	4	4
Transcription antiterminator: *nusG*					1	1	1		1	1	1	1		1	1		1	1		1		1					
Transcription terminator: *nusA, rho*					1		1					1		1			2	2	1	1			1	1		2	1
L DNA replication, recombination and repair																											
Bacterial nucleoid DNA-binding protein and histones:	1	1	1					1	3	1		1		1	2	2	1	2		3			1	1		2	2
hupA, hupB, hbs, hlpA, hpyA1–2, han1																								1		2	2
RNA helicase: *deaD, rheA/rhe1*					1	1	1	1	1		1	1	1		1		1	1		1	1			1	1	1	1
Single-strand binding protein: *ssb*					1				1	1					1		1	1		1		1		1			1
Recombination protein: *RecA*					1				1			1			1		1			1	1			1			1
CELLULAR PROCESSING AND SIGNALING																											
D Cell division and chromosome partitioning																											
Cell division proteins: *ftsZ, DivIVA*							1	1				1	1		1	1	1	1			2	2	1				1
O Posttranslational modification, protein Turnover chaperons																											
Chaperone proteins: *dnaK, dnaJ, GroEL, GroES, grpE, hscB, mopA, mopB, htpG, htpX, hsp90*	3	3	3	2	5	4	3	3	3	3	3	3	3	3	6	1	3	3	3	3	1	2	3	2	4	5	3
Peptidyl-prolyl cis-trans isomerase: *ppiA, ppiB, slyD, fkpA, fklB, cyp*	2	1	1	1	2	1		3	3	3		1	2	2	3		3	3	3	3				1	1	2	3
Thioredoxin: *trxA, trxB, trxM, trxH*	1	3	1	1			1			1	3	2		1	1				1		1	1	2	1	2	1	
Alkyl hydroperoxide reductase protein: *ahpC, ahpF*			1	2			1	1	1						1		1	2	1	1	1	1		1		1	1
Trigger factor: *tig, ropA*	1	1	1	1			1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1		1	1
Clp protease: *ClpX, ClpP, HslU*			1	1		2	1	1	1		1	1										2	1		1		
Ribose-phosphate pyrophosphokinase: *PrsA*		1	1	1		1			1					1			1	1	1							1	
Cell division: *ftsH*							1				1	1			1		1		1			1		1	2		1
M Cell envelop biogenesis, outer membrane																											
Channel forming, conductance: *tsx, mscL*	1			1		1			1					1			1	1		1	1	1		1			
Lipoproteins: *pal, lpp*					1			2	1	1				1	2		3	2	2	2						2	2
Outer membrane proteins: *omp fam, plp, nmpC, fadL*								4	5	4				3	3		8	7	9	5						6	7

(Continued)

Table 3. Continued

Functional classes	Aci	Bha	Bsu	Bth	Bba	Cdi	Efa	Eca	Eco	Hin	Lpl	Lla	Mac	Pmu	Plu	Pab	Sty	Sat	Son	Sfl	Sag	Smu	Spn	Spy	Syn	Vch	Ype
N Cell mobility and secretion																											
Secretory proteins: *sec fam, gspD, gspG, yajC, yidC*	1				2	1	1	2	3	2				3	3		3	4	3	4	2				1		2
Flagellin proteins: *hag, fli fam, fla fam*		1	2	1	1			1							1				1							1	3
Membrane GTP-binding proteins: *typA, lepA*	1			1	1	1		1			1	1			1		1	1		1							
P organic ion transport and metabolism																											
Superoxide dismutase: *sodA, sodB*	1	1	1	1		1	1	1	2	1		1		1	1		1	1	1	2	1	1	1	1	1	1	1
Phosphate binding proteins: *pst fam, phoU, phnA, phnD*		2			1			3	2		1	1					3	3		2		1		1	2	1	1
Metal-ion binding proteins: *fhuD, fecB, yfeA, afuA, nifH, fbpA, cyaY, copP*		4		2						2			1	4	1						2		2	1			
METABOLISM																											
C energy production and conversion																											
Hydratase: *citB, acnB, fumA, fumC*	2	1	1			1		1	1	1				1	1		1	1	1	1							
Dehydrogenases*	7	8	5	10	7	7	6	5	6	5	5	6	14	6	6	6	6	5	11	5	2	3	5	5	3	6	6
Membrane-bound ATP-synthase: *atp fam*	4	3	2	3	2	5	2	4	3	4	3	3	3	5	4	3	3	3	4	3	3	2	2	3	3	4	3
Succinyl-CoA synthetase: *suc fam*	2	2		1				2	3						2	1	2	2	4	3						2	2
Ferredoxin: *fer, fdx, fdxA, petF*	2	2	1	1									5		1	1						1			2		1
Reductases: *nqr fam, drgA, frdA, frdB, hdrD, gor, nirB, napA, mer, ycgG, ftrC*								2	3	4		1	2	3	1			1	3	3	1				2	6	2
Cytochrome oxidases; *cyo fam, qoxB, ctaD, cco fam*	2		1			1									1		1	2	2	1						1	
Cytochrome: *cccA, qcrA, qcrB, cyd fam, napC, napB, scyA, petB*		1	1			2			1	3				3		1		1	6	1					5		
Inorganic pyrophosphatase; *ppa, ppaC*	1	1			1	1	1	1	1	1				1	1	1	1	1	1	1	1	1	1	1		1	1
Phosphate acetyltransferase: *pta*			1	1			1	1	1	1			1	1	1		1	1		1			1	1		1	1
Formate acetyltransferase: *pfl fam*								1	1	1		1		1	1		1	1	1	1	1	1	1	1		1	1
Acetate Kinase: *ackA*				1		1	1	1	1	1			1	1	1		1	1		1	1			1		1	1
Flavodoxin: *fldA, yfjE, flaW*									1	1		1		1					1		1	1	1	1		1	1
G carbohydrate transport and metabolism																											
Enolase: *eno*	1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1	1	1	1	1	1	1	1	1	1
1-6-Bisphosphate aldolase: *fda, fbaA, fbp, lacD, glpX, cbbA*	1	1	1	1		1	1	1	1	1	1	1		1	2		1	1	1	1	1	1	2	2	2	2	2
Glyceraldehyde-3-phosphate dehydrogenase: *gap fam, gapdh, plr*	1	1	1	2	1	1	1	1	1	1	1	2		1	1	1	1	1	1	1	2	2	1	2	1	1	1
Inositol-1-monophosphatase: *suhB*	1								1	1				1	1		1	1	1	1							
Transketolase: *tkt*	1		1	1				2	1	1				2	1		1	1	1	1	1	1	1	1	1	2	1
Transporter (ABC and others): *malE, malX, glpT*		4						1		3	2	4	1	3						2	1		7	3		3	2
PTS system: *pts fam, mtlA, crr, man fam, ptn fam, ptcB, lacE*		2	2	2			2	2	5	1	8	6					5	6		4	9	13	12	9		3	2
Triosephosphate isomerase: *tpi, tpiA*		1	1	1			1	1	1	1	1	1		1	1		1	1		1	1		1	1		1	1
Phosphoglycerate kinase/mutase: *pgk, pgm, gpmA, pmg9*		1	2	1	1	1	2	2	2	2	2	2		2	1		3	3		3	3	3	2	3	1	2	3
Pyruvate kinase: *pyk fam, ppdK*			1	1	1		1	1	1	1	1	1	1	1	1	1	1	1		1	1		1	1		1	1
6-Phosphofructokinase: *pfk*				1			1		1	1	1	1					1	1		1	1		1	1		1	
Glucose-6- phosphate isomerase: *pgi, gnp*				1			1		1		2	1					1	1		1	1	1	1	1			1
Transaldolase: *tal*				1				1	1	1				1	1		1	1		1					1	1	1

(Continued)

Table 3. Continued

Functional classes	Aci	Bha	Bsu	Bth	Bba	Cdi	Efa	Eca	Eco	Hin	Lpl	Lla	Mac	Pmu	Plu	Pab	Sty	Sat	Son	Sfl	Sag	Smu	Spn	Spy	Syn	Vch	Ype
6-Phosphogluconate dehydrogenase: *gnd*							1	1		1	1	1		1	1								1			1	1
Isomerases: *uxaC, araA, xylA, manC, rpiA, araD, fucA, lacA, lacB, gmhA*									1		1	1				1			1	1	1	1	2	2	1		
E amino-acid transport and metabolism																											
Transporters*	1	2			3			1	1	2	1	2	2	1		2	1	2		1	5	4	2			3	1
Glutamine synthetase: *glnA*	1	1	1	1	1	1		1	1		1	1	2		1		1	1	1	1	1	1	1		1		1
Serine hydroxymethyltransferase: *glyA*	1		1	1		1		1	1	1				1	1	1	1	1	1	1		1					1
Aminotransferases: *aspC, dapD, rocD, speE, argD, argF-1, nifS, otcA, arcB, aspB, glmS*	2	1		2	1		1	1		2	1	2	2	1	1	4	1	3		1	1	1	1	1		1	2
Ketol-acid reductoisomerase: *ilvC*	1	1	1					1	1	1		1	1		1		1	1				1					1
Dehydrogenases: *quiA, ald, dhlE, gdh, asd, serA, leuB, tdh, aroE*				3		2		1		2			3		2	3			1			3	2				
Synthases: *cysK, thrC, asnB, asnH, gltD, argG, leuA, speE, cpa, cysK*	1	1						2				1	4		1	1						2	1		2	1	1
Lyases: *aspA, nana*								1	1	2				1	1		1	1		1			1			1	
F nucleotide transport and metabolism																											
Nucleoside diphosphate kinase: *ndk, ndkR*	1	1				1		1	1				1		1				1	1					1		1
Other kinases: *adk, pyrH, cmk*						1		1	1	1	1	1	1	1	1		2	2	2	1	I	1	1	1	1	1	1
Nucleoside di(tri)phosphate reductase: *ndkH, nrdA, nrdB, nrdF, nrdD nrdH*							1	2		1	1		1	2	2						2	1	1			1	
GMP reductase: *guaB, guaC*			1	1		1	1	1	1	1		1					1	1	1	1	1	1	1	1		1	1
GMP synthase: *guaA*			1	1		1	1	1	1	1		1				2	1	1		2	1		2	1			
CTP synthetase: *ctrA, pyrG*	1		1	1			1					1	1						1		1	1				1	
Adenylosuccinate synthetase: *purA*	1			1		1	1	1	1	1				1	1		1	1		1	1						1
Transferases: *hpt, pyrE, pyrl, apt, gpt, upp*			1	2				1	1		1	2		1		1				1	2		3	1			
Purine-nucleoside and phosphorylase; *deoD, punA*				2		1		1	1	1		1		1	1		1	1	1	1	1			2		1	1
Deoxyribose- phosphate aldolase: *deoC*				1		1			1			1	1		1		1	1		1	1	1		1			
H coenzyme metabolism																											
S-adenosylmethionine synthetase: *metK, metX*	1		1	1	1			1	1	1	1				1		1	1	1	1					1	1	1
I lipid metabolism																											
Acyl carrier protein: *acp fam, ydiD*	1	1	1				1	2	1	1		1		1	1		1	1	2	1	1		1	1	1	1	1
AcetylCoA carboxylase: *acc fam*	3	2	1			1		2	1	1				2	2		2	2		2						1	2
β-ketoacyl-acyl carrier protein synthase: *fabB, fabF*	1								1	1		1		1			1	1		1		1					1

[a] Proteins occurring in the functional core of 27 organisms are grouped in functional classes corresponding to COG classification. Each value in the table corresponds to the number of proteins with a given function (among those listed in the corresponding row) occurring in the functional core of a given organism. Missing values correspond to 0. For functional classes marked by*, protein names are given in supplementary tables 1 and 2. A precise listing of protein names for families (*fam*) is also given in supplementary tables 1 and 2. *Acinetobacter sp ADP1* (Aci), *Bacillus halodurans* (Bha), *Bacillus subtilis* (Bsu), *Bacillus thuringiensis konkukian* (Bth), *Bdellovibrio bacteriovorus* (Bba), *Corynebacterium diphtheria* (Cdi), *Enterococcus faecalis V583* (Efa), *Erwinia carotovora atroseptica SCRI1043* (Eca), *Escherichia coli K12* (Eco), *Haemophylus influenzae* (Hin), *Lactobacillus plantarum* (Lpl), *Lactococcus lactis* (Lla), *Methanosarcina acetivorans* (Mac), *Pasteurella multocida* (Pmu), *Photorhabdus luminescens* (Plu), *Pyrococcus abyssi* (Pab), *Salmonella typhi* (Sty), *Salmonella typhimurium LT2* (Sat), *Shewanella oneidensis* (Son), *Shigella flexneri* 2a (Sfl), *Streptococcus agalactiae* 2603 (Sag), *Streptococcus mutans* (Smu), *Streptococcus pneumoniae TIGR4* (Spn), *Streptococcus pyogenes* (Spy), *Synechocystis PCC6803* (Syn), *Vibro cholerae* (Vch), *Yersinia* pestis CO92 (Ype).

coding for proteins known to be highly expressed as ribosomal and glycolytic proteins are for fast growers. Then one ranks the genes by *CAI* values: Genes ranking the highest are the most biased, and those ranking the lowest are the least affected by selective bias. For fast growers, genes with high *CAI* value turned out to be the ones that are highly expressed (Sharp & Li 1987).

In (Carbone et al., 2003), we extended Sharp's hypothesis, saying that for a genome *G*, there is a reference set *S* of coding sequences, constituting approximately 1% of the genes in *G* (the size of *S* is suggested by Sharp's original work), which is representative of *dominating* codon bias in *G*. We consider SCCI to be defined as SCCI $(g)=\left(\prod_{k=1}^{L} w_k\right)^{1/L}$, where the reference set *S* is the most biased set of genes in the organism with respect to this formula, that is, *S* is the (self-consistent) set of genes that take maximum value in the formula when *S* is chosen as a reference set. We showed that SCCI correlates to the dominating bias in a genome, such as GC content, preference for codons with G or C at the third nucleotide position; or a leading strand richer in GT than a lagging strand, and that for translationally biased organisms, it computes codon adaptation (that is, SCCI coincides with CAI). This is demonstrated by observing that *S* can be automatically computed by a pure statistical analysis of all genes in a genome, which is not based on biological knowledge of the organism. For fast growers, *S* was found to consist of highly expressed genes (notice that in Sharp analysis, *S* was manually defined). The lack of reliance on biological knowledge allows computation of weights for organisms of unknown life-style. The name SCCI is for the first time employed in this article; in Carbone et al. (2003) and Carbone et al. (2004), we call it CAI, but because it is not only of codon adaptation that we speak except in the case of translationally biased genomes, it seems less confusing to give to the notion a new and more appropriate name. Also, CAI has been employed with a manual and explicit choice of *S*, whereas the formula CAI parameterized with *S* (i.e., SCCI) turns out to be a *universal measure* to study codon bias.

Codon weights, reference set *S* and SCCl values were calculated with the program CAIJava (Carbone et al. 2003), which uses parsers of GenBank flat files from the Biojava (www.biojava.org) programming package. A description of the algorithm and a validation of the approach reported by Carbone et al. (2004). The program CAIJava is available at www.ihes.fr/~carbone/data.htm.

Detection of Translational Bias

In (Carbone et al. 2004), two numeric criteria were introduced to detect translational bias in a genome. The *ribosomal criterion* defines the z-score (SCCI(r) $-\mu$) $/\sigma$ for each gene of a ribosomal protein *r*, where mean μ and standard deviation σ are calculated for the SCCI distribution over all CDS; this allows to define the average $\overline{Z}_{Rib}$ of z-scores for ribosomal proteins and say that an organism characterized by translational bias is expected to have high $\overline{Z}_{Rib}$, i.e., >1. The *strength criterion* computes codon weights, as in (1) based on all genes in the genome *G* (w_k (*G*)) and on the genes in the set *S* (w_k) and expects the difference between w_k (G) and w_k to be large for translationally biased genomes (i.e., $\sum_{k=1}^{64} (w_k(G) - w_k)/2 > 8$; this sum is an indicator of the number of amino acids having different preferred codons in the entire genome and in the set of most biased genes; the threshold, was empirically calculated on known translationally biasedorganisms (Carbone et al. 2004)). The combination of the two criteria allows to determine which genomes are strongly translationally biased, that is, those satisfying both criteria, from those that are weakly so, that is those that satisfy the ribosomal criterion only. Notice that our numerical criteria provide quantitative values ranging within a continuous interval, and that based on these values, one can identify strong, weak, and absent forms of bias as well as finer classifications. The 27 genomes considered here satisfy both ribosomal and strength criteria and have been characterized as strongly biased. The list of the 27 functional genomic cores is given in Supplementary Table 1.

Table 4. Comparison of small genomes with *E. coli* genome

	W. brevipalips	*Buchnera aph.*	*M. genitalium*
No. of CDSs	617	504	484
E. coli homologous genes	606	498	266
E. coli homologous core genes	235	231	144

SCCI Distribution Tail

This is defined as the set of genes *g*, with SCCI (g) $\geq \mu + \sigma$ where μ and σ are mean and SD of the distribution. This means a SCCI >0.42 for *E. coli* and >0.44 for *B. subtilis* in Fig. 1. For all other organisms, thresholds are listed in Supplementary Table 4. This definition can be easily applied to all bacterial genomes, and it ensures that (1) genes in the tail largely deviate from the average behavior of the genome and (2) that a significant number of genes belongs to the tail. This number was >200 for all 27 genomes considered, and it allowed comparison of species in a meaningful manner (Table 3). Asking for SCCI (g) $> \mu + n\sigma$ with $n \geq 2$ would provide too little information on relevant genes: The genome of *H. pylori*, for instance, has only 97 genes with SCCI(g) $\geq \mu + 2\sigma$.

Results

We propose a new computational approach to detect a set of potentially essential genes for a microbial genome, called *functional genomic core*, that is not based on comparative genomics but on the analysis of codon bias on single genomes. The method ranks genes with respect to the SCCI, and highly biased genes are proposed to form the functional core. Before presenting a comparative study of core genes across species, it is instructive to observe some properties of distributions within single genomes and across functional classes and to discuss the potential size of a functional core for a single species. In Fig. 1, function-specific histograms show that *E. coli* and *B. subtilis* have comparable landscapes of SCCI distributions; that the distribution of genes within functional classes does not depend on SCCI values; and, most important, that all functional classes display long tail representatives (the same is also true for eukaryotic species such as *S. cerevisiae*, not shown). This suggests verifying whether genes that form histogram long tails are shared across species. (We do not expect shared genes to be orthologous but rather to intervene in the same functional activity). Genes within long tails are significantly deviating from average to be considered important for the organisms life. Based on this intuition, we suggest them to constitute functional cores. As illustrated by Fig. 1 for *E. coli* and *B. subtilis* the distribution of SCCI values demonstrates no clear discontinuity to discriminate "core" from "non-core" genes in a principled manner.

To show that core genes are shared across translationally biased organisms and are proportional in number across functional classes, we considered the most biased 200 core genes as functional cores representatives and compared them. A complete list of functional classes and proteins appearing in functional cores is given in Supplementary Table 2, and a concise overview is described in Table 3 (displaying approximately 82 functional subclasses characterized by genes shared by > 9 microbes).

Genes Shared by Most Functional Cores

All functional genomic cores contain virtually complete translation, transcription, and replication machineries. All groups of chaperones are present. The recombination and repair system is rudimentary as is the cell division and chromosome partitioning system. Among proteins involved in cellular processes and signaling, posttranslational modification, secretion, inorganic ion transport and, metabolism (with superoxide dismutase, metal-ion binding proteins, and phosphate-binding proteins) are well represented. Much less represented are signal transduction mechanisms. Very well represented, but only in a few organisms, are outer-membrane proteins and lipoproteins. Within metabolic activities, production and conversion of energy and carbohydrate metabolism are well represented. All enzymes involved in the Embden-Meyerhof pathway and in the conversion of pyruvate into coenzyme A and acetate occur in functional cores. Inorganic pyrophosphatase, phosphate acetyltransferase, membrane-bound ATPase, and sugar phosphotransferase system (PTS) proteins occur in almost all functional cores. Amino acids and nucleotide transport and metabolism are also very well represented, whereas coenzyme metabolism, lipid metabolism, secondary metabolism biosynthesis, transport, and catabolism are poorly represented.

Genes Essential to Specific Living Conditions

A functional genomic core also collects genes that play a role in the life of an organism performing specific metabolic activities and living under specific environmental conditions (possibly with a limited amount of nutrients and in the presence of adverse factors including competition (Koonin 2000)). Genome specificity was observable in all functional cores we detected (see Supplementary Table 3).

Photosynthesis. *Synechocystis* functional genomic core contains phycobilisome proteins (such as phycocyanin, allophycocyanin, degradation, and linker polypeptide), photosystem I and II proteins, fructose-1,6-bisphosphatealdolase and ferredoxin. The presence of proteins involved in *photosynthesis* within the most biased genes is a good indicator of the known photosynthetic activity and life-style of *Synechocystis* (Carbone and Madden 2005). These genes are expected to appear in the functional genomic core of all photosynthetic organisms and to represent a necessary condition for a photosynthetic organism to function properly.

Methane metabolism. Methanol-5 hydroxybenzimidazolylcobamideco methyltransferase, methyl coenzyme M reductase, and methylcobamide methyltransferase isozyme M are among the most biased genes in *M. acetivorans*, which uses methane metabolism in an essential way. Within genes involved in methylotrophic methanogenesis (Galagan et al. 2002), four corrinoid proteins are core. Proteins involved in acetoclastic methanogenesis (Galagan et al. 2002), that is, *Ack, Pta*, and (the three copies of) *cdhA*, are core genes; the remaining genes in the *cdh* family (that is *cdhB, cdhC, cdhD*, and *cdhE)* are also highly biased.

Ferredoxin. Among the most biased genes of the archaea *P. abyssi*, we found ferredoxin, ferredoxin oxidoreductase, and keto-valine-ferredoxin oxidoreductase *y*-chain. Ferredoxin appears to be the major metabolic electron carrier in *Pyrococci* (Cohen et al. 2003; Silva et al. 2000; Schut et al. 2001; Ward et al. 2000).

Sporulation. The three *Bacillus* species—*halodurans, subtilis*, and *thuringiensis*—present small acid-soluble spore proteins (*ssp* family) that are involved in sporulation within their functional genomic core. *B. subtilis* functional core contains also spore coat proteins (*cotD*, *cotG*, and *cotN*), a general stress protein (*gsiB*), and a transcriptional regulator (*abrB*), all of which are also involved in sporulation.

Metabolism of carbohydrates. Genes for transport and metabolism of cellobiose, sucrose, and *β*-glucoside are found within the functional genomic core of *S. mutans*, which is capable of metabolizing a wide variety of carbohydrates (Ajdić et al. 2002) but not on the functional core of the other *Streptococci* we considered. Also, S. *mutans* is able to convert several sugar-alcohols to glycolytic intermediates, and the genes for metabolism of mannitol were also present in its functional core. Note that all *Streptococci* species contained genes for glucose, fructose, mannose, and maltose and maltodextrin metabolism in their functional core, and all species except *S. agalactiae* contained galactose enzymes. This example illustrates well how sensitive the method is to detect functional differences even between closely related species.

Genes with Unknown Function

The number of genes with uncharacterized function belonging to functional cores varies from organism to organism and goes from as few as 5 in *S. flexneri 2a* to up to 107 in *B. bacteriovorus* in the top 200 core genes considered. As discussed later this is a particularly interesting set of genes because it likely provides important candidates in the search of genes with specific functional activity.

Comparison With Data Issued by Comparative Genomics

We compared our 27 functional genomic cores (Table 3) against the two minimal gene sets proposed in (Mushegian and Koonin (1996) and in (Charlebois and Doolittle 2004). The remarkable fitting of the data we found in both cases implies that most essential genes constituting minimal gene sets are highly biased in translationally biased organisms. The most represented functional classes of genes issued by comparing *M. genitalium* with *H. influenzae* (Mushegian and Koonin 1996) correspond to the most represented functional classes in functional genomic cores (Table 3). Some proteins were expected to be essential but instead were found to be missing by Mushegian and Koonin (1996), whereas occur in functional genomic cores: In transcription, we observed a variety of sigma factors, mainly *rpoD* and *rpoE* but also *rpoH* in *E. coli* and *Shighella* and *rpoF* in *Synechocystis*. Termination factors *rho* are also present in functional cores. All groups of chaperones are present, *hsp*90 (*htpG*) included; for energy metabolism, proteins occurring in the PTS were detected. In translation, consistent with Mushegian and Koonin (1996), no tRNA nucteotidyltransferase was found in functional cores, with the exception of the archaea *M. acetivorans*.

Most prevalent genes computed from cross-phylum analysis (that is, genes present in at least 80% of genomes from each of the 14 bacterial (12) and archaeal (2) phyla of 170 prokaryotes considered) in Charlebois & Doolittle (2004) appear in the functional genomic cores of our 27 organisms (as illustrated Table 3 and in Supplementary Table 2). Among these 71 prevalent genes, only a few do not appear in any of the 27 functional genomic cores: DNA polymerase III *(dnaX)*, DNA primase (*dnaG*), endonuclease III (*nth*), and topoisomerase IA (*topA*), implicated in replication, recombination, and repair; dimethyladenosine transferase (rRNA methylation, *ksgA)* and histidyl-tRNA synthetase *(hisS)* in translation; Xaa-Pro aminopeptidase *(pepP)* in amino-acid transport and metabolism; xanthosine triphosphate pyrophosphatase (*yggV*) in nucleotide transport and metabolism; and preprotein translocase subunit *(secY)* in intracellular tracking and secretion. Enlarging the number of core genes to 500 would allow recovery of *topA, hisS, secY, ksgA*, and *yggV* as core genes for a few (<4) organisms, with a *dnaG*-like primase appearing in *P. abyssi*. Functional genomic cores contain RNA polymerase subunits other than *rpoB* as well as many ribosomal proteins that have not been detected in (Charlebois and Doolittle 2004).

Comparison with Data Issued by Experiments Carried out in E. coli

Experiments based on transposon mutagenesis (Gerdes et al. 2003) have suggest that 625 genes of the 3746 analyzed are essential for robust aerobic growth of *E. coli* in rich media. Among the set of genes found to be essential under this growth condition and to be preserved in > 80% diverse bacterial genomes (Gerdes et al. 2003) (this accounts for 171 genes), we detected 40% of them within the first 200 most biased genes of *E. coli* and approximately 60% of them within a functional genomic core made of 572 genes. Note that genes such as enolase (*eno*), which is found universally present in genomes of all kingdoms, were not considered essential by Gerdes et al. (2003), whereas we know about its crucial role in glycolysis and gluconeogenesis and that its malfunctioning is likely to create difficult living conditions. This gene is the third most biased gene in *E. coli* and within the most biased ones for our 27 organisms. It belongs to all functional cores we studied. Also, genes expressed under anaerobic conditions, or genes involved in stress response and DNA maintenance and repair, were not susceptible to detection as essential by the experiments, but they were numerically well represented in our core set.

Large-scale chromosomal deletion (Hashimoto et al. 2005) allowed for a decrease of the *E. coli* genome by 30% inducing cells to grow albeit with an increased doubling time. The PEC database provides the list of essential and nonessential genes in *E. coli* that were identified after genome minimization: 234 coding sequences were essential; 1860 were nonessential; and 900 genes were classified with unknown behavior. Hundred and twenty nine essential genes, 278 nonessential ones, and 53 with unknown behavior are core genes. Among the 278 core genes classified as nonessential, 74 genes are involved in stress response. Sixty three core genes were deleted and do not appear in the minimized genome, and most of them are also stress response genes (32). In addition to those, we found some outer-membrane proteins that might be also induced in stress response, some transport proteins, and genes with unknown or putative function; no gene involved in stress response was detected as essential by the experimental methodology.

The high sensitivity of the comparison with different experimental methodologies is reflected in the fact that only 144 essential genes of the PEC pool are common to the 620 genes identified in Gerdes et al. (2003) and that the essential genes detected in Gerdes et al. (2003) are almost uniformly classified among the essential, nonessential, and unknown PEC categories (as observed in Fang et al. [2005]). This, together with the highly favorable living conditions in a laboratory setting, which are missing to a bacteria thriving in the wild or competing with other organisms (possibly mutants) for limited resources, fits with the finding that only 53.63% of the genes common to the two experimental pools are core. In fact, our computational method (see also Fang et al. [2005]) provides many genes that are not directly involved in growth, but rather in conditions of starvation or stress, and genes whose loss may lead to such a lower degree of fitness that their deletion will never be fixed in natural populations.

Comparison with Data Issued by Systematic Inactivation in B. subtilis

A systematic inactivation of *B. subtilis* genes, through several vectors constructed to perform insertional mutagenesis in the chromosome, was carried out (on a rich medium) and lead to the detection of 248 essential genes (Kobayashi et al. 2003). We compared the 519 *B. subtilis* functional core genes (detected with SCCI >0.44; see Supplementary Table 4) with the 248 essential genes and found 126 of them to be core. Most genes involved in the Embden-Meyerhof-Parnas pathway are core genes, and this is in agreement with their unexpected essentiality as argued in Kobayashi et al. (2003), where these enzymes are proposed as candidates for novel and unexpected functions in the cell. Core genes such as pgm (phosphoglycerate mutase) and eno (enolase), whose absence is known to induce slow growth in *B. subtilis* (Illades-Aguiar & Setlow 1994), were identified as essential.

Comparison with Small Parasitic Genomes

If the genetic patrimony of all organisms constitutes only a part of the set of essential genes, then comparison of small genomes with their closely related translationally biased organisms, although living in different environments, should identify these genes. We compared *E. coli* genome against two small γ-proteobacteria genomes, *Buchnera aphildicola str Bp* and *Wigglesworthia brevipalpis*, that have been proven (Gil et al. 2002; Akman et al. 2002) to be phylogenetically close to *E. coli.* Because of their minimal size and high similarity to *E. coli*, we expect them to contain a large number of core genes of *E. coli.* Results (summarized in Table 4) support the use of SCCI values as discriminators of essentiality in translationally biased genomes as well as the idea that the notion of minimality is intrinsically related to life–style. In fact, *Buchnera aph.* displays 498 of its 504 genes as homologous to *E. coli* genes (Gil et al. 2002; vanHam et al. 2003), and among those half (231) are genes with high SCCI value in *E. coli.* The genetic difference between the currently existing *Buchnera aph.* and *E. coli* shows the difference of living environments between them, one being a symbiont living in a cell and the other a free-living organism. For this, one should expect some core genes to be specific to *Buchnera aph.* Also, new findings show that this genome is still experiencing a reductive process toward a minimum set of genes necessary for its symbiotic lifestyle (Gil et al. 2002). This evidence supports our hypothesis that essential genes should be sought within highly biased genes.

The genome of *W. brevipalpis*, the primary endosymbiont of tsetse flies, displays 606 of 617 genes as homologous to *E. coli* genes, and, among these 235 genes have high SCCI in *E. coli.* It contains, in addition to genes for parasitic life, a subset of genes of free-living bacteria, such as the enteric *E. coli* and *S.typhimurium* (Akman et al. 2002). Although *Buchnera aph.* and *W. brevipalpis* share apparent functional and evolutionary similarities with regard to their symbiotic association with their insect hosts, their genetic blueprints are quite different: *W. brevipalpis* shares only 381 of its CDSs with *Buchnera aph.*, and these mostly represent the indispensable housekeeping genes. Among these genes, 190 have high SCCI in *E. coli.* Only 45 genes are shared between *W. brevipalpis* and *E. coli* but not *Buchnera aph.* and have high SCCI in *E. coli*; they code mostly for proteins participating in the synthesis of products involved in cellular processes, cell structure, fatty-acid metabolism, and, especially, biosynthesis of cofactors. Only 45 genes are shared between *Buchnera aph.* and *E. coli* but not *W. brevipalpis* and have high SCCI in *E. coli.* They code for proteins involved in biosynthesis of amino acids, in glycolysis, for components of the PTS system, and for nicotinamide adenine dinucleotide dehydrogenase subunits. These findings are in agreement with the analysis of Zientz et al. (2004), who reported these metabolic pathways as being the main differences between the two species.

By comparing *M. genitalium* with *Buchnera aph.* we found 189 homologous genes, and more than two thirds of these genes (129) had high SCCI value in *E. coli.* Comparison of the firmicutes *M. genitalium* with *B. subtilis* yielded 330 homologous genes, of which 143 are core genes in *B. subtilis.*

Table 5. No. of enzymes belonging to the functional genomic core of 27 organisms

Enzymes	Aci	Bha	Bsu	Bth	Bba	Cdi	Efa	Eca	Eco	Hin	Lpl	Lla	Mac	Pmu	Plu	Pab	Sty	Sat	Son	Sfl	Sag	Smu	Spn	Spy	Syn	Vch	Ype
Synthases	18	7	8	23	9	16	10	20	23	15	12	16	23	13	21	19	22	22	22	26	14	14	18	14	7	19	13
Dehydrogenases	8	9	6	16	8	11	8	8	7	10	8	9	17	8	10	10	7	6	13	6	8	9	9	7	4	8	8
Transferases	5	4	6	9	2	3	7	9	10	11	12	13	12	7	6	7	10	13	3	10	14	17	22	13	1	9	9
Kinase/mutase	3	4	5	10	2	7	6	7	12	8	7	7	4	8	8	1	11	11	6	11	8	6	8	8	5	8	8
Isomerases	3	3	3	3	2	1	4	5	8	5	5	7	3	3	5	1	6	6	4	7	3	3	6	5	2	4	6
Reductases			3	3	1	1	3	9	7	8	2	3	8	7	5	9	3	5	5	8	5	5	2	2	3	9	4
Transporters	1	8			6			2	1	6	5	6	3	5		3	1	2	2	3	7	5	14	4	1	8	4
Phosphatase/ phosphorilase	2	1		3	1	2	1	2	3	5		1		3	4	2	4	4	3	3	4	1	2	5		2	2
Helicases					1	1	1	2	1		1	1	1		1		1	1		1	1	1		1	1	1	1
Hydratases/ dehydratases	3	2	1			1		2	2	2	2	2	1	2	1		1	1	1	2	1	2		1	2		
Oxidases	2		1	1	1	1		1						1	1		2	2	2	1			1			1	1
Lyases	2			1				2	1	2				2	1		2	3	1	1		1	1			1	
Proteases			1	1	1	2	1	2	1	1	1	1			1		1	2		3	1	2	1		1		
DNA-gyrases					1	2		2	1								2	2		1							1
Hydrolases	1					2			1				1		1	1	1			1		2		1	2		
Ligases			1	1							2		1		1						1			1		1	
Hydrogenases													3			1				1							

Essential Cellular Functions in Archaea

The clear split between bacterial and eukaryotic components of archaeal genomes (Koonin et al. 1997; Makarova et al. 2003) suggests verification of the origin of the proteins contained in the functional genomic core of archaeal genomes. Among the 200 most biased proteins in *P. abyssi*, 54 occured in bacteria, 35 in eukaryotes, 80 in both bacteria and eukaryotes, and 23 exclusively in archaea (the remaining 8 were not classified in COG). Proteins having only eukaryotic homologues are involved in translation and associated functions, and those with only bacterial homologues are involved in metabolic functions, confirming, for proteins lying in functional cores, the observation made by Koonin (2003) for minimal gene sets. Based on the splitting of protein origins, no conclusion can be drawn on the origin of archaeal species.

Conclusion

Comparison among functional cores is useful in understanding differences among organisms. We observed that there is no ubiquitous set of genes appearing as part of any universal core; instead there is much variability within functional classes, with some functions (or genes) more represented across species than others. Informational and operational genes are both represented in functional genomic cores, and all functional classes are represented by at least a few genes that are largely spread among species. Translational and transcriptional classes contain many more of these largely spread genes than other functional classes (Table 3). These observations correspond closely to the outcome of the analysis proposed in Mushegian and Koonin (1996) on *M. genitalium* and *H. influenzae*, and one could argue that any genome comparison between two sufficiently distant organisms sharing the same environmental conditions and physiology could lead to the detection of those genes that are essential for survival. Although in principle this is true, the comparative approach demands biological knowledge of the two organisms in question and, most of all, a clear definition of "distance" between organisms, whereas our method output minimal sets of genes on a single genome without the need for comparative study or any biological knowledge of the organism.

Within a functional core, some genes have uncharacterized function. Because of the putative essentiality of core genes, such uncharacterized sequences are likely to be relevant for understanding basic molecular mechanisms governing the cell, in agreement with the discussion in Hutchison et al. (1999). Comparison of different functional cores might be used in guiding functional annotation of uncharacterized core genes. For instance, among uncharacterized genes for organisms with a less-known life-style than *E. coli* (which shares 66 functional classes with at least 8 other organisms; Table 3), one is likely to find examples of nonorthologous gene displacement, that is, proteins that are distantly related or nonorthologous but that share a function with those in the *E. coli* functional core (Koonin 2003). In the same spirit, the absence of expected enzymes for an organism (Table 5) allows proposition of uncharacterized genes as participating in putative metabolic pathways. These suggestions are likely to be helpful in guiding experiments, especially for those microbes about which little is known of their living environment and whose annotation is poor.

Supplementary Material

Supplementary Table 1: A list of 200 most biased genes with position, SCCI value and functional annotation for the 27 organisms listed in the legend of Table 3.

Supplementary Table 2: Complete set of data allowing the construction of Table 3. A list of proteins is associated to each functional class.

Supplementary Table 3: List of genes with only a predicted function (classified as R and S in COG). Genes have been organized in COG classes.

Supplementary Table 4: A list of 27 organisms for which max and min SCCI value of genes in functional cores, mean and standard deviation of SCCI distribution over the whole genome, sum of mean and standard deviation used to determine the size of the corresponding functional cores are given.

Acknowledgments. We thank A. Zinovyev and C. Vaquero for insightful discussions and H. Isambert, A. Policriti, and L. Trojan for critical reading of the manuscript. This work was supported by a grant from the Fondation Pour la Recherche Médicale.

References

Mushegian AR, Koonin EV (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268–10273

Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, Bult CJ, Kerlavage AR, Sutton G, Kelley JM, Fritchman RD, et al. (1995) The minimal gene complement of *Mycoplasma genitalium*. Science 270:397–403

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty DA, Merrick JM, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. Science 269:496–512

Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV (2003) Comparative genomics of the archaea (euryarchaeota): Evolution of conserved protein families, the stable core, and the variable shell. Genome Res 9:608–628

Nesbø CL, Boucher Y, Doolittle WF (2001) Defining the core of non-transferable prokaryotic genes: The euryarchaeal core. Mol Evol 53:340–350

Harris JK, Kelley JT, Spiegelman GB, Pace NR (2003) The genetic core of the universal ancestor. Genome Res 13:407–412

Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2004) Universal trees based on large combined protein sequence data sets. Nat Genet 28:281–285

Koonin EV (2003) Comparative genomics, minimal gene sets and the last common ancestor. Nat Rev Microbiol 1:127–136

Charlebois RL, Doolittle WF (2004) Computing prokaryotic gene ubiquity: Rescuing the core from extinction. Genome Res 14:2469–2477

Itaya M (1995) An estimation of the minimal genome size required for life. FEBS Lett 362:257–260

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Anderson KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al. (2003) Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA 100:4678–4683

Hutchison CA, Peterson Gill SN, Cline RT, White O, Fraser CM, Smith HO, Venter JC (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. Science 286:2165–2169

Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, Hutchison III CA, Smith HO, Venter JC (2006) Essential genes of a minimal bacterium. Proc Natl Acad Sci USA 103:425–430

Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, Mekalanos JJ (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. Proc Natl Acad Sci USA 99:966–971

Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, Somera AL, Kyrpides NC, Anderson I, Gelfand MS, Bhattacharya A, et al. (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG 1655. J Bacteriol 185:5673–5684

Hashimoto M, Ichimura T, Mizoguchi H, Tanaka K, Fujimitsuk, Keyamura K, Ote T, Yamakava T, Yamazaki Y, Mori H, Katayama MS, Kato T (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. Mol Microbiol 55:137

Salama NR, et al. (2004) Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. J Bacteriol 186:7926–7935

Ji Y, et al. (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. Science 293:2266–2269

Forsyth RA, et al. (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. Mol Microbiol 43:1387–1400

Thanassi JA, et al. (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. Nucleic Acids Res 30:3152–3162

Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, et al. (1997) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. Science 285:901–906

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, et al. (2005) Functional of the *Saccharomyces cerevisiae* genome. Nature 418:387–391

Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, et al. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. Nature 421:231–237

Nishi K, Dabbs ER, Schnier J (1988) DNA sequence and complementation analysis of a mutation in the rplX gene from *Escherichia coli* leading to loss of ribosomal protein L24. J Bacteriol 163:890–894

Gerdes SY, Scholle MD, D'Souza M, Bernal MV, Baev A, Farrell M, Kurnasov OV, Daugherty MD, Mseeh F, Polanuger BM (2002) From genetic footprinting to Antimicrobial drug targets: Examples in cofactor biosynthetic pathways. J Bacteriol 184:4555–4572

Grantham R, Gautier C, Gouy M, Mercier R, Pave A (1980) Codon catalog usage and the genome hypothesis. Nucleic Acids Res 8:r49–r62

Sharp PM, Li W-H (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acid Research 15:1281–1295

Carbone A, Zinovyev F, Képés F (2003) Codon Adaptation Index as a measure of dominating codon bias. Bioinformatics 19:2005–2015

Carbone A, Madden R (2005) Insights on the evolution of metabolic networks of unicellular translationally biased organisms from transcriptomic data and sequence analysis. J Mol Evol 61:456–469

Carbone A, Képés F, Zinovyev A (2004) Codon bias signatures, organisation of microorganisms in codon space and lifestyle. Mol Biol Evol 22:547–561

Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? Mol Biol Evol 22:2147–2156

Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19:1720–1730

Daubin V, Gouy M, Perriuere G (2002) A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. Genome Res 12:1080–1090

Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. PLoS Biol 1:E19

Kreil PD, Ouzounis CA (2001) Identification of thermophilic species by the amino-acids composition deduced from their genomes. Nucleic Acids Res 29:1608–1615

Lynn DJ, Singer GA, Hickey DA (2002) Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res 30:4272–4277

Tekaia F, Yeramian E, Dujon B (2002) Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. Gene 297:51–60

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141–1153

Venter JC, Levy S, Stockwell T, Remington K, Halpern A (2003) A massive parallelism, randomness and genomic advances. Nature Genetics 33:219–227

Zimmer C (2003) Genomics. Tinker, tailor: Can Venter stitch together a genome from scratch? Science 299:1006–1007

Smith HO, Hutchison CA III, Pfannkoch C, Venter C (2003) Generating a synthetic genome by whole genome assembly: AX174 bacteriophage from synthetic oligonucleotides. Proc Natl Acad Sci USA 100:15440–15445

Rocha EP, Matic I, Taddei F (2002) Over-representation of repeats in stress response genes: A strategy to increase versatility under stressful conditions? Nucleic Acids Res 30:1886–1894

Koonin EV (2000) How many genes can make a cell: The minimal-gene-set concept. Annu Rev Genomics Hum Genet 1:99–116

Galagan JE, Nusbaum C, Roy A, Endrizzi MG, Macdonald P, FitzHugh W, Calvo S, Engels R, Smirnov S, Atnoor D, et al. (2002) The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. Genome Res 12:532–542

Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Querellou J, Ripp R, et al. (2003) An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol Microbiol 47:1495–1512

Silva PJ, van den Ban EC, Wassink H, de Haaker HCB , Robb FT, Hagen WR (2000) Enzymes of hydrogen metabolism in *Pyrococcus furiosus*. Eur J Biochem 267:6541–6551

Schut GJ, Zhou J, Adams MW (2001) DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus:* Evidence for a new type of sulfur-reducing enzyme complex. J Bacteriol 183:7027–7036

Ward DE, Kengen SW, Van der Oost J, De Vos WM (2000) Purification and characterization of the alanine aminotransferase from the hyperthermophilic Archaeon *Pyrococcus furiosus* and its role in alanine production. J Bacteriol 182:2559–2566

Ajdić D, McShan WM, McLaughlin RE, Savic G, Chang J, Carson MB, Primeaux C, Tian R, Kenton S, Jia H, et al. (2002) Genome sequence of *Streptococcus mutans UA159*, a cariogenic dental pathogen. Proc Natl Acad Sci USA 99:14434–14439

Illades-Aguiar B, Setlow P (1994) Studies of the processing of the protease which initiates degradation of small, acid-soluble proteins during germination of spores of *Bacillus species*. J Bacteriol 176:2788–2795

Gil R, Sabater-Muoz B, Latorre A, Silva FJ, Moya A (2002) Extreme genome reduction in *Buchnera* spp.: Toward the minimal genome needed for symbiotic life. Proc Natl Acad Sci USA 99:4454–4458

Akman L, Yamashita A, Watanabe H, Oshima K, Shiba T, Hattori M, Aksoy S (2002) Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. Nature Genet 32:402–407

van Ham RCHJ , Kamerbeek J, Palacios C, Rausell C, Abascal F, Bastolla U, Fernández JM, Jiménez L, Postigo M, Silva FJ, Tamames J, Viguera E, Latorre A, Valencia A, Morán F, Moya A (2003) Reductive genome evolution in *Buchnera aphidicola*. Proc Natl Acad Sci USA 100:581–586

Zientz E, Dandekar T, Gross R (2004) Metabolic interdependence of obligate intracellular bacteria and their insect hosts. Microbiol Mol Biol Rev 68:745–770

Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: Computer analysis of protein sequences predicts novel functions and suggests a chimeric origin of the archaea. Mol Microbiol 25:619–637