counteracted by HIV-1 Vif [21,22]. It is possible that Vif might be fully functional and active in interactions that involve multiple targets.

Cells have evolved an astonishing number of mechanisms to protect themselves from viral invaders, however, viruses have developed stealthy mechanisms to escape these inhibitory pathways. The battle between cells and retroviruses is an ancient and on-going one. APOBEC3G, Hck and Sp140 are not the only recently identified cellular defenders in this battle. Gao *et al.* [23] have reported on ZAP, a protein that targets and destroys viral mRNAs, and Hatziioannou *et al.* [24] have described the antiretroviral factors Fv1 and Lv1, which target the capsid proteins of incoming retroviral cores of the viral DNA. Though overall we should not expect these new findings to translate to antiviral treatment for patients in the near future, they certainly bring us a giant step ahead and make biomedical scientists more hopeful in the battle against pathogenic viral infections.

References

1 Mariani, R. *et al.* (2003) Species-specific exclusion of APOBEC3G from HIV-1 virions by Vif. *Cell* 114, 21–31
2 Zhang, H. *et al.* (2003) The cytidine deaminase CEM15 induces hypermutation in newly synthesized HIV-1 DNA. *Nature* 424, 94–98
3 Harris, R.S. *et al.* (2003) DNA deamination mediates innate immunity to retroviral infection. *Cell* 113, 803–809
4 Lecossier, D. *et al.* (2003) Hypermutation of HIV-1 DNA in the absence of the Vif protein. *Science* 300, 1112
5 Mangeat, B. *et al.* (2003) Broad antiretroviral defense by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* 424, 99–103
6 Fisher, A.G. *et al.* (1987) The *sor* gene of HIV-1 is required for efficient virus transmission *in vitro*. *Science* 237, 888–893
7 Strebel, K. *et al.* (1987) The HIV 'A' (*sor*) gene product is essential for virus infectivity. *Nature* 328, 728–773
8 Gabuzda, D.H. *et al.* (1992) Role of Vif in replication of human immunodeficiency virus type 1 in CD4 + T lymphocytes. *J. Virol.* 66, 6489–6495
9 von Schwedler, U. *et al.* (1993) Vif is crucial for human immunodeficiency virus type 1 proviral DNA synthesis in infected cells. *J. Virol.* 67, 4945–4955
10 Zhang, H. *et al.* (2000) Human immunodeficiency virus type 1 Vif protein is an integral component of an mRNP complex of viral RNA and could be involved in the viral RNA folding and packaging process. *J. Virol.* 74, 8252–8261
11 Khan, M.A. *et al.* (2000) Human immunodeficiency virus type 1 Vif protein is packaged into the nucleoprotein complex through an interaction with viral genomic RNA. *J. Virol.* 76, 8252–8261
12 Dettenhofer, M. *et al.* (2000) Association of human immunodeficiency virus type 1 Vif with RNA and its role in reverse transcription. *J. Virol.* 74, 8938–8945
13 Madani, N. and Kabat, D. (1998) An endogenous inhibitor of human immunodeficiency virus in human lymphocytes is overcome by the viral Vif protein. *J. Virol.* 72, 10251–10255
14 Simon, J.H.M. *et al.* (1998) Evidence for a newly discovered cellular anti-HIV-1 phenotype. *Nat. Med.* 4, 1397–1400
15 Sheehy, A.M. *et al.* (2002) Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* 418, 646–650
16 Stopak, K. *et al.* (2003) HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Mol. Cell* 12, 591–601
17 Sheehy, A.M. *et al.* (2003) The antiretroviral enzyme APOBEC3G is degraded by the proteasome in response to HIV-1 Vif. *Nat. Med.* 9, 1404–1407
18 Marin, M. *et al.* (2003) HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nat. Med.* 9, 1398–1403
19 Yang, S. *et al.* (2001) The multimerization of human immunodeficiency virus type 1 (HIV-1) Vif protein: a requirement for Vif function in the viral life-cycle. *J. Biol. Chem.* 276, 4889–4893
20 Yang, B. *et al.* (2003) Potent suppression of viral infectivity by the peptides that inhibit multimerization of human immunodeficiency virus type 1 (HIV-1) Vif proteins. *J. Biol. Chem.* 278, 6596–6602
21 Hassaine, G. *et al.* (2001) The tyrosine kinase Hck is an inhibitor of HIV-1 replication counteracted by the viral Vif protein. *J. Biol. Chem.* 276, 16885–16893
22 Madani, N. *et al.* (2002) Implication of the lymphocyte-specific nuclear body protein Sp140 in an innate response to human immunodeficiency virus type 1. *J. Virol.* 76, 11133–11138
23 Gao, G. *et al.* (2002) Inhibition of retroviral RNA production by ZAP, a CCCH-type zinc finger protein. *Science* 297, 1703–1706
24 Hatziioannou, T. *et al.* (2003) Restriction of multiple divergent retroviruses by Lv1 and Ref1. *EMBO J.* 22, 385–394

0966-842X/$ - see front matter
doi:10.1016/j.tim.2004.02.004

Genome Analysis

Gene duplication and biased functional retention of paralogs in bacterial genomes

Dirk Gevers, Klaas Vandepoele, Cedric Simillion and Yves Van de Peer

Bioinformatics and Evolutionary Genomics, Ghent University/Flanders Interuniversity Institute for Biotechnology (VIB), Technologiepark 927, B-9052 Ghent, Belgium

Gene duplication is considered an important prerequisite for gene innovation that can facilitate adaptation to changing environments. The analysis of 106 bacterial genome sequences has revealed the existence of a significant number of paralogs. Analysis of the functional classification of these paralogs reveals a preferential enrichment in functional classes that are involved in transcription, metabolism and defense mechanisms. From the organization of paralogs in the genome we can conclude that duplicated genes in bacteria appear to have been mainly created by small-scale duplication events, such as tandem and operon duplications.

Microbial genomes have a considerable fraction of genes that are homologous to other genes within the same genome [1,2]. There are basically two ways through which

Corresponding author: Yves Van de Peer (Yves.VandePeer@psb.Ugent.be).

Glossary

HOMOLOGS: genes ascribed to the same gene family indicating that they have a common ancestry.
ORPHANS: single copy genes without any homolog.
ORTHOLOGS: homologous genes from different genomes ascribed to the same gene family, evolved by speciation.
PARALOGS: homologous genes within a genome ascribed to the same gene family, created by duplication.
PARANOME: collection of paralogous genes.
STRAIN-SPECIFIC EXPANSIONS (SSE): gene families of paralogs without any orthologs, and consequently restricted to one strain.
XENOLOGS: homologous genes within a genome ascribed to the same gene family, obtained by horizontal gene transfer.

such intra-genome HOMOLOGS (see Glossary) can arise: (i) through duplication, where both gene copies are called PARALOGS, and (ii) by acquiring similar genes from outside sources through horizontal gene transfer, where both gene copies are called XENOLOGS. Although the latter case is not a duplication mechanism *sensu strictu*, such similar genes can act as if they were actual paralogs. The process by which paralogs are generated usually depends on repeated sequences within the genome, such as the widely distributed rRNA genes, insertion elements, or other transposable elements or repetitive element (REP) sequences [3,4]. Gene duplication occurs as an evolutionary response in bacteria that have been exposed to different selection pressures, such as starvation conditions and thermal stress [5–7]. When the selective pressure is removed the duplicates can be rapidly lost, thereby forming a reversible adaptive mutation that alters gene dosage without really altering genetic information. In addition to this short-term evolutionary advantage, gene duplication and consequent functional divergence is considered an important evolutionary step towards diversity in the functional repertoire of an organism. This presumably enables the organism to adapt to varying environmental conditions and to broaden the phenotypes that it expresses [8,9]. Together with horizontal gene transfer and gene loss, gene duplication is also considered an important process that shapes prokaryotic genomes [10–12]. Detailed analysis of the duplication events that have occurred is essential for understanding the evolutionary dynamics of prokaryotic genomes. To date, studies on gene duplication in prokaryotes have only been performed on a limited number of genomes, and have either been focusing on the degree of duplication in individual genomes [1,2] or on the possible causes for functional divergence of bacterial paralogs [9]. Here, we discuss the duplication history of 106 complete bacterial genomes, the occurrence of tandem and block duplications, and the functional composition of the PARANOME.

Box 1. Detection of paralogs and block duplications

The detection of duplicated genes is based on determining homologous genes within a genome. To obtain candidate homologs, protein sequences are compared against each other in a sequence similarity search, using BLASTp [27]. One way to select 'suitable' homologs is to apply an E-value cut-off, as used in previous analyses [1,9]. In our approach, two protein sequences are called homologous when they share more than 30% overall sequence identity (recalculated to a similarity along the entire amino acid sequence, as suggested elsewhere [28], and not as a similarity over the alignable region) and have an alignable region of more than 150 amino acids. For matching sequences with an alignable region of less than 150 amino acids, a cut-off curve based on homology-derived secondary structure prediction identity is used to determine whether the two sequences are homologous [29]. Based on the obtained list of intra-genome homologs, the degree of duplication per genome was determined. Annotated phage- and transposon-related sequences and homologs to such elements were excluded from the analysis.

A block duplication is a genomic segment (containing multiple genes) rather than individual genes that have been duplicated. Such block or segmental duplications within a genome can be recognized as regions with statistically significant conserved gene content and order, referred to as colinearity. To detect a colinear genomic region, a 'map-based' approach can be applied, in which the dataset consists of all gene products and their position on a genomic sequence. For all gene products, pairs of homologous genes are determined as described above. The information on homologous genes is then stored in a so-called gene homology matrix (GHM) of n by n elements (n being the total number of genes in the genome), each non-zero element (x,y) being a pair of homologous genes (x and y denote the coordinates of these genes) (Figure I). In the matrix, colinear regions are represented as diagonal lines, whereas tandem duplications are manifested as purely horizontal or vertical lines, inversions can be detected by looking at the organization of the elements, and gaps in diagonal regions refer to gene loss in one of the duplicated segments. A colinear region is therefore defined in the matrix representation as several points showing diagonal proximity. Previously, we developed the software tool ADHoRe (automated detection of homologous regions) for detecting block duplications and applied it to eukaryotic genomes [30–32].

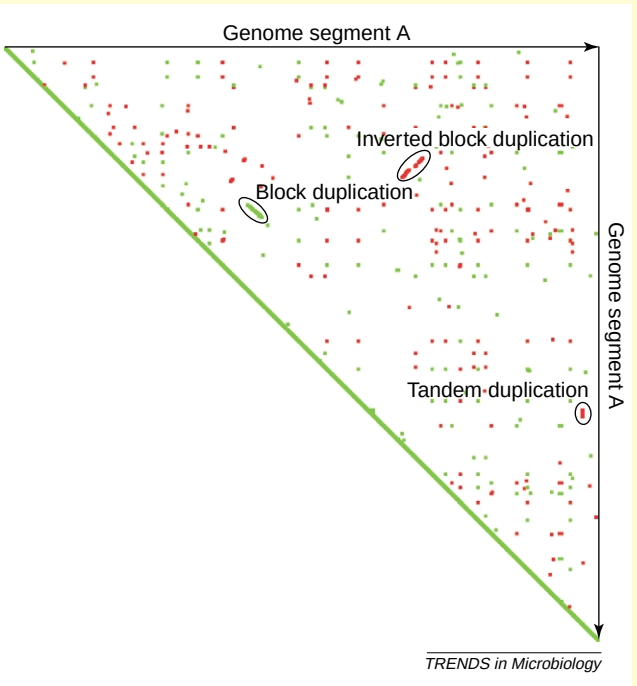

Figure I. A part of the gene homology matrix of *Salmonella typhimurium* LT2. A small part of the *S. typhimurium* LT2 genome was compared against itself using the ADHoRe algorithm [30]. Only half of the gene homology matrix is shown, as both halves are identical. Green and red dots indicate pairs of homologous genes that are in the same or inverted orientation, respectively.

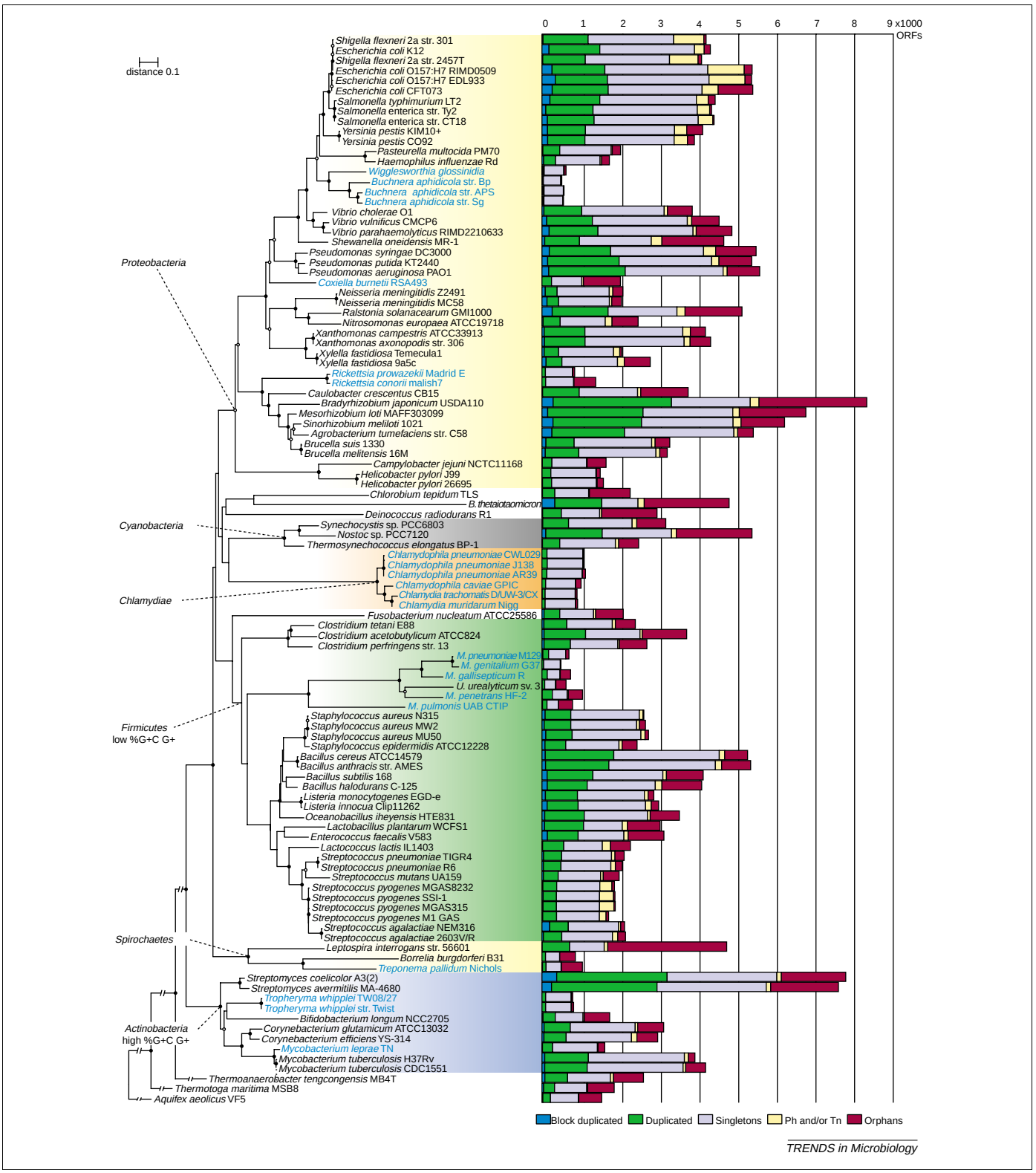

Figure 1. Prevalence of paralogs within 106 complete bacterial genome sequences. Each bar represents the size of the genome in terms of the number of open-reading frames (ORFs) and is divided into five categories: (i) genes occurring in block duplicated regions, (ii) duplicated genes (not in block), (iii) singleton genes with homologous genes in other genomes, (iv) phage- or transposon-related genes (Ph and/or Tn), and (v) orphans (genes without any homolog). The phylogenetic tree is based on 16S rRNA and was constructed by neighbour-joining as implemented in TREECON [33] and considering among-site rate variation [34]. Nodes that have bootstrap support above 50% and 75% are indicated with a white and black dot, respectively. Obligate intracellular organisms are shown in blue. See also our website for detailed numerical data (http://www.psb.ugent.be/bioinformatics/).

Furthermore, the paranome is discussed in relation to the whole bacterial proteome, in particular, with respect to strain-specific gene family expansions and the shared set of bacterial core genes. The approach used for detecting duplicated genes and block duplications is detailed in Box 1.

The bacterial paranome

The occurrence, genomic organization and functional annotation of homologous genes within each of the 106 bacterial genomes were analyzed. Each of the surveyed genomes exhibits a significant number of paralogs, ranging from 7% of its proteome for *Rickettsia conorii* to 41% for *Streptomyces coelicolor* A3(2) (on average 23.5 ± 8.7; see Figure 1). As shown previously using a smaller dataset of 23 genomes [1,2], our data confirm that the genome size is strongly correlated with the number of paralogs in a linear regression ($R^2 > 0.94$); additional material can be viewed on our website (http://www.psb.ugent.be/bioinformatics/) under the section entitled 'Supplementary Material'. The total combined bacterial paranome consists of 87 866 proteins (i.e. 29% of the proteome under study). The largest group of paralogs (>50 genes) encodes ABC-type transporters, transcriptional regulators or dehydrogenases. In the case of *Escherichia coli* K12, for which a list of predicted horizontally acquired genes is available [13], ~16% of the 1425 intra-genome homologs were estimated to have been acquired by horizontal gene transfer. Therefore, all other genes are assumed to have evolved by gene duplication.

Regarding the organization of paralogous genes within the genomes, it was found that 15% of the paranome consists of tandem duplicates, and 9.5% is located in block duplicated segments. *Streptococcus agalactiae* NEM316 showed a maximum of 9.31% of its genes in block duplications, whereas for 16 genomes block duplications could not be found at all. The majority of these genomes (i.e. 14 out of 16) are from obligate intracellular organisms, which are characterized by an overall strong reduction in genome size (Figure 1). An analysis of the block duplicated regions showed that the majority resembles the typical bacterial operon size (three to four genes). In the case of *E. coli* K12, one of the few organisms for which operons have been identified experimentally [14], the detected block duplicated regions were checked for similarity to previously described operons. All *E. coli* block duplicated segments were categorized into 20 groups (one group containing between two to seven homologous blocks), of which ten contain at least one known operon [Figure 2; and see also our website (http://www.psb.ugent.be/bioinformatics/) for more examples]. Evolution of the operon content by insertion, deletion and/or tandem duplication is observed in several groups. Furthermore, these ten groups can be regarded as the result of paralogous operon acquisition, because none of the operons was predicted to have been introduced into the *E. coli* genome by horizontal transfer, with the exception of the nickel transport operon [13]. This horizontally acquired operon was shown to be similar to six other block duplicated segments, of which three are known operons encoding peptide uptake, and dipeptide and oligopeptide transport.

The functional landscape of the paranome

For 48 genomes, a gene function description and a functional class could be assigned for the majority of the genes using the functional annotation provided for the so-called COGs (clusters of orthologous groups) at NCBI (http://www.ncbi.nlm.nih.gov/COG/) [15]. An overview of the functional landscape of the paranome for these 48 bacterial genomes is presented in Figure 3, which allows us to determine whether paralog retention is biased towards specific functional classes for each of the bacterial strains. It appears that the preferentially retained duplicated genes mainly belong to the functional classes that are associated with amino acid metabolism (class E, see Figure 3), transcription (class K), inorganic ion metabolism (class P), and to a lesser extent with carbohydrate metabolism (class G), defense mechanisms (class V), and energy production and conversion (class C). For the classes E, P and V the preferentially retained expansion can be explained by the amplification of transport proteins, which are known to be the largest families that contain the greatest numbers of paralogs in a single organism [16]. Other known large families include the transcriptional regulatory proteins and response regulators, which are members of class K. In addition, more organism-specific cases of paralog retention can be deduced from our analysis [see our website (http://www.psb.ugent.be/bioinformatics/) for a map that provides a detailed overview of all strains by class combinations]. For example, in the paranome of mycobacteria, two functional classes with an excess of retained duplicated genes are prominent, namely 'lipid transport and metabolism' (class I) and 'secondary metabolites biosynthesis, transport and catabolism' (class Q). The fatty acid metabolism group is in agreement with the complex nature of the mycobacterium cell wall [17] and might reflect adaptive evolution of the bacterial cell surface. In the *Mycoplasma* and *Ureaplasma* genomes there is an enrichment in defense proteins (class V), such as ABC-type antimicrobial peptide transporters, multidrug efflux pumps, restriction endonucleases and restriction modification systems. A similar increase in ABC-type multidrug transporters is observed for *Rickettsia prowazekii* Madrid E, which confirms previous findings [18]. The biased retention of duplicated motility genes (class N) and chemotaxis genes (class T) in *Borrelia burgdorferi*, together comprising more than 6% of its proteome, appear to be biologically linked. Because *B. burgdorferi* lacks recognizable virulence

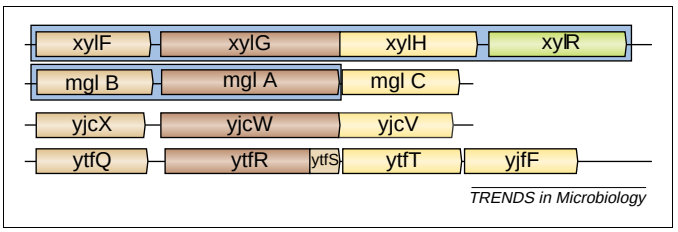

Figure 2. Operon duplication and evolution in the *Escherichia coli* K12 genome. Four block duplicated segments with conserved gene content and order are shown. Each gene family is represented by a different color, and known operons are marked with a blue box. Clear cases of operon content evolution by deletion, gene fission and tandem duplication are observed in this group of sugar transport operons. Other examples of operon duplications found in the *E. coli* genome can be found on our website (http://www.psb.ugent.be/bioinformatics/).

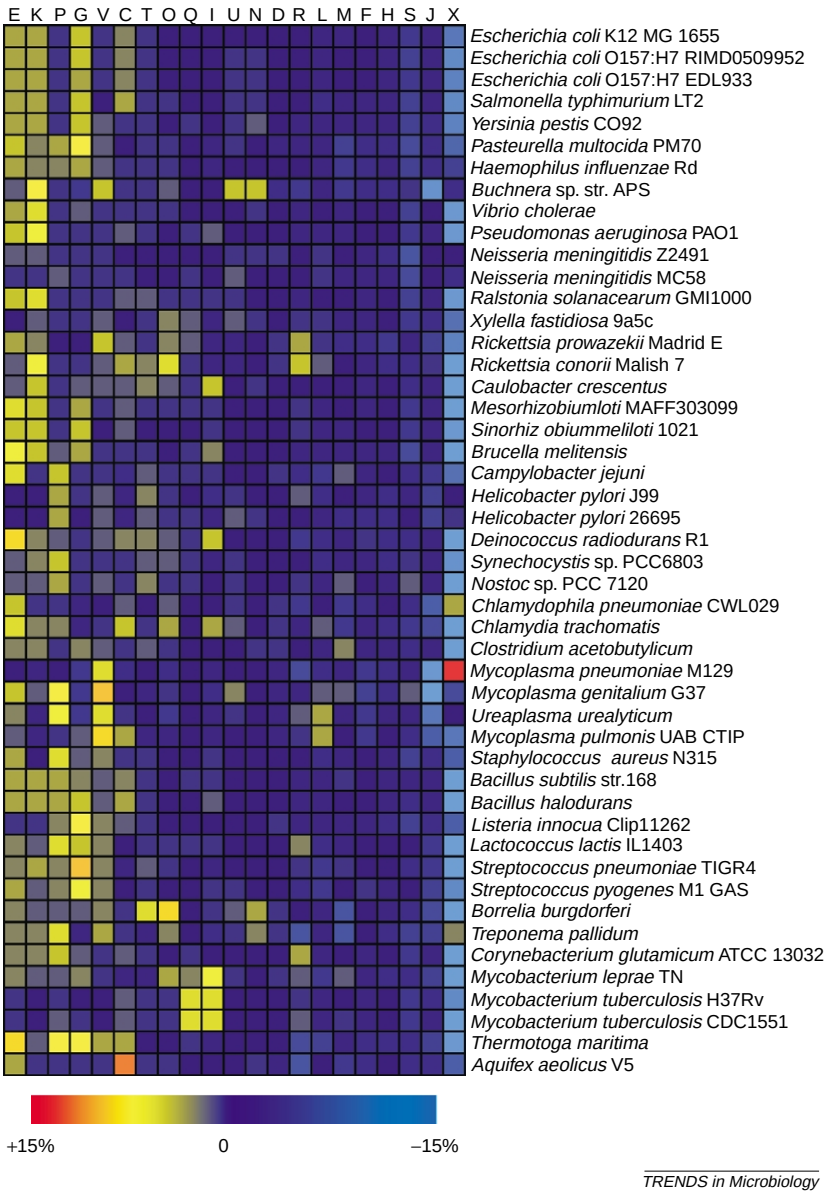

Figure 3. The functional landscape of the paranome. Preferentially retained duplicated genes according to their functional classes are shown for those genomes for which the functional classification is available; this can be found at NCBI (http://www.ncbi.nlm.nih.gov/COG/). The functional distribution of the paranome in terms of percentages is subtracted by the functional distribution of the single copy genes in terms of percentages, revealing the functional classes that have relatively more (yellow to red), or less (gradient of blue) paralogs compared to single copy genes. The functional classes have been ordered according to relative average contribution. See also our website (http://www.psb.ugent.be/bioinformatics/) for more information on each strain–functional class combination. The different functional classes are abbreviated as follows: C, energy production and conversion; D, cell cycle control, cell division and chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination and repair; M, cell wall/membrane/envelope biogenesis; N, cell motility; O, posttranslational modification, protein turnover and chaperones; P, inorganic ion transport and metabolism; Q, secondary metabolites biosynthesis, transport and catabolism; R, general function prediction only; S, function unknown; T, signal transduction mechanisms; U, intracellular trafficking, secretion and vesicular transport; V defense mechanisms; X, none of the others. Note that all genes or open-reading frames (ORFs) annotated as being transposon- or phage-related have been removed from this analysis.

factors, its ability to migrate to distant sites in the tick and mammalian host is probably dependent on a robust chemotaxis response, as reported previously by Fraser and co-workers [19]. These authors suggest that multiple chemotaxis genes can be differentially expressed under varied physiological conditions, or that different flagellar systems exist that require different chemotaxis systems. Interestingly, a large number of *Mycoplasma pneumoniae* M129 paralogs in class X occur in a limited number of strain-specific expansions. As most of these genes are described as (expressed) proteins with unknown function, further research is required to determine their role in the organism's lifestyle.

Strain-specific expansions

To analyze the taxonomic distribution of gene families, the annotated protein sequences from the 106 bacterial genomes (equal to 319 334 protein sequences) were subjected to a inter-genome homology search analogous to the intra-genome homology search described in Box 1. On the basis of these homology relations, gene families were constructed across genomes using a single-linkage

clustering algorithm. Looking at the taxonomic distribution of the gene families (Figure 4), a total of 2121 families (5651 proteins) consist of genes that are restricted to one strain (i.e. paralogs without any ORTHOLOGS) and consequently should be considered STRAIN-SPECIFIC EXPANSIONS (SSE). Such SSE-families were found in nearly all the genomes analyzed, although in different numbers. Among the strains with the highest number of SSE-families are *Bacteroides thetaiotaomicron* VPI-5482, *Bradyrhizobium japonicum*, *Nostoc* sp. PCC 7120, and *Ralstonia solanacearum* GMI1000. The expansion level of the SSE-families (i.e. the number of members in the family) ranges from 2 to 77, although the majority has a size smaller than 3. The family with the highest expansion level was found in *Bacteroides thetaiotaomicron* VPI-5482, and encodes outer membrane proteins that are probably involved in nutrient-binding. The majority of the other families (~40%) consists of members that are labeled 'protein of unknown function', and is an interesting dataset for further research. In theory, these families are similar to the lineage-specific expansions (in the strictest sense) described by Jordan and co-workers [8], who suggested that such families were either derived *de novo* in the current strain or that they diverged to such an extent that significant sequence similarity to homologs in other strains is no longer readily apparent. Such families probably resemble an adaptive evolution of a strain to its environment.

The bacterial core genes

The taxonomic distribution also revealed a subset of gene families that is found in all 106 bacterial genomes (51 families, 11 749 proteins), which can be considered as representing house-keeping genes (Figure 4). These core genes encode ribosome proteins, translation elongation factors, tRNA-synthetases, ABC-type transporters, topoisomerases, polymerases and ATP/GTPases; a detailed list can be found on our website (http://www.psb.ugent.be/bioinformatics/). The largest family contains 5357 genes and corresponds to the ABC-type transport superfamily, which is known to be the largest group of paralogous genes in bacterial and archaeal genomes [20,21]. In addition, there are numerous gene families that are nearly ubiquitous, this is largely because of erroneous gene annotations in some organisms and to a lesser degree as a result of non-orthologous gene displacement (unrecognizable homology at the sequence level) [22]. The inclusion of those gene families with members in more than 95% of the genomes results in an additional 64 families (7686 proteins) that might be considered as part of the core. These families were found to encode similar functions as the core families *sensu strictu*, and in addition, chaperonins, proteases, DNases and DNA ligase. More than half of these core families contain more than one gene per strain, and to some extent have therefore been subject to gene duplication. Those core families that have only one member per strain and are true orthologs might be considered most valuable for inferring bacterial phylogeny and for complementing traditional 16S rDNA-based phylogenies [23,24]. When the bacterial core families obtained in our study were compared with the 205 ortholog families obtained in a multigene phylogeny approach that had been applied to the γ-Proteobacteria [25], 70 families overlapped. This means that 45 bacterial core families from our analysis do not have unambiguous orthologous relationships among its members, and that 135 gene families that are conserved in the γ-Proteobacteria are not appropriate for phylogenetic studies on a broader taxonomic scale.

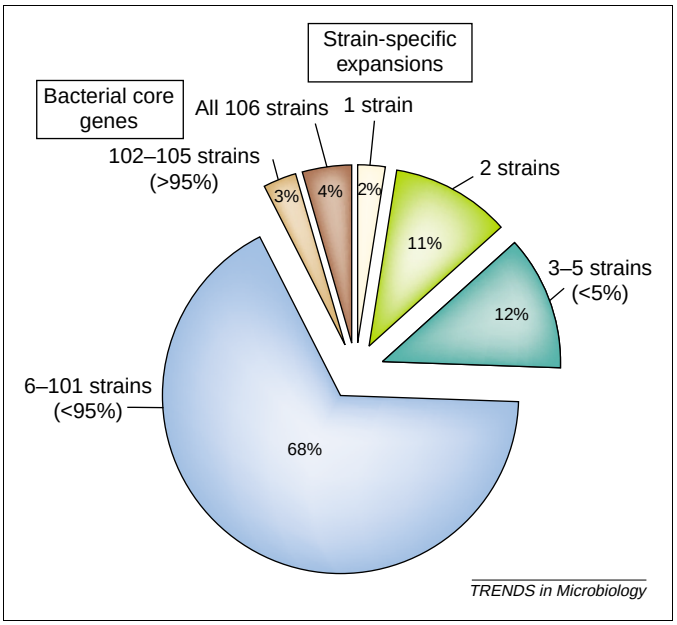

Figure 4. Taxonomic distribution of gene families. For each of the gene families, the number of distinct strains in which the family was represented was determined. The results were divided into six categories (slices), going from gene families represented in only one strain (strain-specific expansions) up to gene families represented in all 106 strains (the bacterial core genes). The size of each category is expressed as the percentage of the composite proteome that it encloses (indicated within the slice).

In contrast to the ubiquitous gene families, a subset of the proteome (57 174 proteins) does not cluster into gene families and can be considered as strain-specific genes (also referred to as ORPHANS) within our dataset (Figure 1). These orphans were subjected to a homology search against all known proteins that were not included in our dataset (i.e. non-genomic proteins or recently added genomic proteins in Swiss-Prot and TrEMBL). Using the stringent criteria described previously, a subset of 3484 proteins (6.1%) was found to show significant homology with other bacterial genes. One might expect that the number of genes without any homolog will decrease over time as more and more complete genomes are sequenced. However, it has been shown that, up to now, there has been a constant rate of new gene family discovery, suggesting that protein sequence space remains largely unexplored [26].

Concluding remarks

Most of the duplicated genes in bacteria appear to have been created by small gene duplication events. Evidence

for large-scale gene duplications, such as those observed in eukaryotic genomes, could not be detected in any of the 106 bacterial genomes investigated. Nevertheless, paralogous genes comprise a significant fraction of the bacterial genome coding capacity. Interestingly, there appears to be a clear correlation between the number of retained duplicates and the functional class to which they belong. In particular, genes that are involved in adapting to a constantly changing environment appear to be preserved, which again shows the importance of gene duplication for biological evolution.

Operons are the principal form of gene organization and regulation in prokaryotes. Previously, comparative analysis of prokaryotic genomes has shown that only a few operons are conserved across large evolutionary distances, suggesting that operons are relatively unstable throughout evolution. In the bacterial paranome we found only a minority as potential retained operon duplications (i.e. blocks of three to four genes), whereas the majority of the duplicated genes occur as single retained genes. The question remains as to whether or not the block duplications that were identified are under a strong selective pressure that preserves gene co-expression and co-regulation in duplicate, or alternatively, whether these represent recently duplicated gene strings that have been spared from rearrangements and disruption.

Acknowledgements

This study was carried out under financial support from the BOF (project nr 01110803). We would like to thank Eric Bonnet and Francis Dierick for their contribution to this work. The institute for the Encouragement of Scientific and Technological Research in the Industry (IWT) is acknowledged by C.S. and K.V. for their predoctoral fellowship. We would also like to thank four anonymous referees for valuable comments.

References

1 Tekaia, F. and Dujon, B. (1999) Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J. Mol. Evol.* 49, 591–600
2 Coissac, E. *et al.* (1997) A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol. Biol. Evol.* 14, 1062–1074
3 Achaz, G. *et al.* (2003) Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* 164, 1279–1289
4 Rocha, E.P.C. (2003) An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Res.* 13, 1123–1132
5 Riehle, M.M. *et al.* (2001) Genetic architecture of thermal adaptation in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A.* 98, 525–530
6 Caporale, L.H. (2003) Natural selection and the emergence of a mutation phenotype: an update of the evolutionary synthesis considering mechanisms that affect genome variation. *Annu. Rev. Microbiol.* 57, 467–485
7 Yamanaka, K. *et al.* (1998) The cspA family in *Escherichia coli*: multiple gene duplication for stress adaptation. *Mol. Microbiol.* 27, 247–255
8 Jordan, I.K. *et al.* (2001) Lineage-specific gene expansions in Bacterial and Archaeal genomes. *Genome Res.* 11, 555–565
9 Hooper, S.D. and Berg, O.G. (2003) On the nature of gene innovation: duplication patterns in microbial genomes. *Mol. Biol. Evol.* 20, 945–954
10 Kunin, V. and Ouzounis, C.A. (2003) The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 13, 1589–1594
11 Snel, B. *et al.* (2002) Genomes in flux: the evolution of Archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25
12 Hughes, D. (2000) Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biol* 1, REVIEWS0006 (http://genomebiology.com/2000/1/6/reviews/0006)
13 Lawrence, J.G. and Ochman, H. (1998) Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. U. S. A.* 95, 9413–9417
14 Itoh, T. *et al.* (1999) Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol. Biol. Evol.* 16, 332–346
15 Tatusov, R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41
16 Saier, M.H. and Paulsen, I.T. (1999) Paralogous genes encoding transport proteins in microbial genomes. *Res. Microbiol.* 150, 689–699
17 Tekaia, F. *et al.* (1999) Analysis of the proteome of *Mycobacterium tuberculosis in silico*. *Tuber. Lung Dis.* 79, 329–342
18 Andersson, S.G. *et al.* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133–140
19 Fraser, C.M. *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* 390, 580–586
20 Tomii, K. and Kanehisa, M. (1998) A comparative analysis of ABC transporters in complete microbial genomes. *Genome Res.* 8, 1048–1059
21 Tatusov, R.L. *et al.* (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.* 6, 279–291
22 Koonin, E.V. *et al.* (1996) Non-orthologous gene displacement. *Trends Genet.* 12, 334–336
23 Daubin, V. *et al.* (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* 12, 1080–1090
24 Stackebrandt, E. *et al.* (2002) Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* 52, 1043–1047
25 Lerat, E. *et al.* (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the gamma-proteobacteria. *PLoS Biol* 1, E19
26 Kunin, V. *et al.* (2003) Myriads of protein families, and still counting. *Genome Biol.* 4, 401
27 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
28 Li, W.H. *et al.* (2001) Evolutionary analyses of the human genome. *Nature* 409, 847–849
29 Rost, B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.* 12, 85–94
30 Vandepoele, K. *et al.* (2002) The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. *Genome Res.* 12, 1792–1801
31 Simillion, C. *et al.* (2002) The hidden duplication past of *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 13627–13632
32 Vandepoele, K. *et al.* (2003) Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15, 2192–2202
33 Van de Peer, Y. and De Wachter, R. (1997) Construction of evolutionary distance trees with TREECON for Windows: accounting for variation in nucleotide substitution rate among sites. *Comput. Appl. Biosci.* 13, 227–230
34 Van de Peer, Y. *et al.* (1996) A quantitative map of nucleotide substitution rates in bacterial rRNA. *Nucleic Acids Res.* 24, 3381–3391

0966-842X/$ - see front matter
doi:10.1016/j.tim.2004.02.007