Segmentally Variable Genes: A New Perspective on Adaptation

Zheng, Yu, Richard J Roberts, Simon Kasif. "Segmentally Variable Genes:A New Perspective on Adaptation" PLoS Biology 2(4): 0452-0464. (2004)
https://hdl.handle.net/2144/3038
Boston University
Segmentally Variable Genes: A New Perspective on Adaptation

Yu Zheng1*, Richard J. Roberts2, Simon Kasif1,3
1 Bioinformatics Graduate Program, Boston University, Boston, Massachusetts, United States of America, 2 New England Biolabs, Beverly, Massachusetts, United States of America, 3 Department of Biomedical Engineering, Boston University, Boston, Massachusetts, United States of America

Genomic sequence variation is the hallmark of life and is key to understanding diversity and adaptation among the numerous microorganisms on earth. Analysis of the sequenced microbial genomes suggests that genes are evolving at many different rates. We have attempted to derive a new classification of genes into three broad categories: lineage-specific genes that evolve rapidly and appear unique to individual species or strains; highly conserved genes that frequently perform housekeeping functions; and partially variable genes that contain highly variable regions, at least 70 amino acids long, interspersed among well-conserved regions. The latter we term segmentally variable genes (SVGs), and we suggest that they are especially interesting targets for biochemical studies. Among these genes are ones necessary to deal with the environment, including genes involved in host-pathogen interactions, defense mechanisms, and intracellular responses to internal and environmental changes. For the most part, the detailed function of these variable regions remains unknown. We propose that they are likely to perform important binding functions responsible for protein–protein, protein–nucleic acid, or protein–small molecule interactions. Discerning their function and identifying their binding partners may offer biologists new insights into the basic mechanisms of adaptation, context-dependent evolution, and the interaction between microbes and their environment.

Introduction

Microbes occupy almost every habitable niche in the biosphere, highlighting their enormous capability for adaptation and survival. This adaptive ability has been refined during millennia of evolution and has resulted in genes that evolve at very different rates. Some, such as housekeeping genes that code for the essential biochemical functions of the organism, are now evolving rather slowly. Others that have to defend against potentially lethal attack by viruses or toxins adapt to varying environmental conditions, often evolve more rapidly (Murphy 1993; Moxon and Thaler 1997; Jordan et al. 2002). Pathogenic microbes, for example, face stringent tests of their adaptive potential because of the escalating efficiency of the host-defense mechanisms (Moxon and Thaler 1997). In the arms race between pathogens and their hosts, both sides try to improve their overall fitness by deploying sophisticated strategies to generate genetic variability (Elena and Lenski 2003). Sequence divergence during rapid evolution can take many forms. Some genes change throughout their entire sequences, resulting in apparently lineage-specific genes that lack clear similar sequences in current versions of GenBank. Others show a mosaic pattern of one or more variable regions interspersed within conserved regions. This latter group is the subject of this paper and we refer to them as segmentally variable genes (SVGs). For the purpose of the current analysis, we define such variable regions as having a minimum length of 70 amino acids, which would permit them to fold into independent domains. This distinguishes them from most nonfunctional interdomain segments, which are usually shorter and whose principal function depends on length rather than specific sequence content.

An example of an SVG family is provided by the cytosine-5 DNA methyltransferases (Posfai et al. 1989). These enzymes typically form parts of restriction-modification systems, which are key components of an important bacterial defense mechanism to protect against phage attack and other unwanted infiltration of foreign DNA (Cheng 1995). These methyltransferases catalyze the addition of a methyl group from S-adenosylmethionine to the 5-position of cytosine and contain a highly variable region of more than 90 amino acids that is responsible for specific DNA sequence recognition (Figure 1A; Posfai et al. 1989; Cheng 1995; Lange et al. 1996). A detailed examination of the three-dimensional (3D) structure of the variable region suggests that it folds into an independent domain, which has been shown to bind to DNA (Cheng et al. 1993). The flanking sequences are highly conserved because they are responsible for the chemistry of methylation, which is common to all members of the family. Variability in this family has arisen because there is a need for great variation in the DNA sequences being recognized so that the specific pattern of methylation becomes a marker to distinguish innate DNA from foreign DNA.

To the best of our knowledge, there has been no systematic attempt to identify, catalog, and classify similar SVGs in the...
sequenced microbial genomes. Nor has any attempt been made to find potentially common functions among genes displaying this property. Since it is known that many genes involved in defense mechanisms, such as the DNA methyltransferases and the antigens exposed on the surface of bacteria, show such variability (Roche et al. 2001), it is tempting to speculate that one might identify host-defense genes based on this property. Thus, the regional variability might reflect the influence of diversifying selection pressure that could come from constant interaction with other fast-evolving molecules in the environment. Could such genes be the predominant members of the SVG families? Or do other genes, such as those involved in basic energy metabolism and synthesis, show similar variability? In this paper we provide an initial systematic analysis. We describe our findings about the distribution of SVGs and the potential function achieved by segmental variability.

Results
Classification of Genes into Three Broad Groups
We carried out a classification of the genes in 43 fully sequenced microbial genomes (see Table S1 for a full name list). A Web site (http://geneva.bu.edu) is also provided with results for several selected genomes, including Escherichia coli, Helicobacter pylori, Neisseria meningitidis, and several others. Each gene is accompanied with schematic diagrams from Pfam (Bateman et al. 2002), BLOCKS (Henikoff et al. 1999), PRINTS (Attwood et al. 2003), and the nongapped BLAST (Altschul et al. 1990) analyses.

For each genome, the full proteome is compared with the nonredundant GenBank sequence set using nongapped BLAST (see Materials and Methods for the parameters used). Based on the degree of conservation or divergence among similar genes in different species, we classify them into three broad groups. Lineage-specific genes are defined as genes with no significantly similar hits from other species in the
current GenBank (E-value cutoff, 1E-5). SVGs are defined as genes containing at least one highly variable region, containing more than 70 amino acids, interspersed among well-conserved regions. In any single SVG family, the length of the variable region can differ only within a certain range (see Materials and Methods for more details). In this paper, regions are considered to be variable if no sequence similarity can be detected against possible homologous genes, where the overall homology is determined by the conserved portions. The rest of the genes in the genome are considered as fully conserved genes. Naturally, this initial soft classification is somewhat dependent on specific thresholds and will be biased by the current state of GenBank and the quality of the annotation.

In Figure 2 we show a scatter plot of the three classes of genes in the \textit{H. pylori} genome in two-dimensional (2D) space, where the x-axis shows the length of the variable region and the y-axis shows the number of possible homologs for each gene. Lineage-specific genes (filled square in Figure 2) by definition naturally cluster on the x-axis. Most of the genes in this group are still annotated as unknown. A few genes with annotated functions in this group, such as the outer-membrane protein family in \textit{H. pylori} (Tomb et al. 1997), only appear in this organism and contribute to its unique biology. A second group contains fully conserved genes (filled triangle in Figure 2) with only short variable regions. It is in this class that most “housekeeping” genes fall. Examples include the subunits of ATP synthetase F1 (atpD, atpA, atpG) and ribosomal proteins such as rps4 (Figure 2), etc. The third group contains the SVGs (filled diamond in Figure 2). A few examples in this group are labeled with their names in Figure 2 and will be discussed later. In Table 1 we list the number of genes in each category for a representative set of microbial genomes (see Table S1 for a full list).

SVGs are subdivided into different types depending on whether they have one, two, or more variable regions. The number of genes with a single variable region is much larger than the number of genes with multiple ones. In Figure 1A we show the variation profile of an SVG containing one variable region. The variability lies in its DNA recognition domain (approximately 140 amino acids), which in this case recognizes the DNA sequence CCTC (Lin et al. 2001). In Figure 1B we give an example with two variable regions. It is the virulence-associated protein homolog VacB from \textit{H. pylori}, which has variable regions at both its N-terminus (approximately 200 amino acids) and C-terminus (approximately 100 amino acids). \textit{VacB} has been shown to encode a 3'–5' exoribonuclease and is necessary for expression of virulence (Cheng and Deutscher 2002). The conserved central region (approxi-
Table 1. Classification of Genes into Three Broad Categories for a Representative Set of Microbial Genomes

| Microorganism                      | Total | Variable Domain = 1 | Variable Domain = 2 | Variable Domain > 2 | Fully Conserved Genes | Lineage-Specific Genes |
|------------------------------------|-------|---------------------|---------------------|---------------------|-----------------------|------------------------|
| **Bacteria**                       |       |                     |                     |                     |                       |                        |
| Helicobacter pylori                | 255 (16%) | 210 | 37 | 8 | 958 (62%) | 353 (22%) |
| Neisseria meningitidis            | 254 (12%) | 226 | 22 | 6 | 1373 (67%) | 440 (21%) |
| Streptococcus pneumoniae          | 259 (12%) | 220 | 34 | 5 | 1485 (71%) | 357 (17%) |
| Escherichia coli                  | 561 (13%) | 488 | 66 | 7 | 3581 (83%) | 147 (3%) |
| Bacillus subtilis                 | 535 (13%) | 476 | 49 | 10 | 3214 (78%) | 351 (9%) |
| Streptomyces coelicolor           | 1111 (14%) | 940 | 146 | 25 | 5821 (74%) | 965 (12%) |
| Mycoplasma genitalium             | 99 (20%) | 85 | 11 | 3 | 380 (79.8%) | 1 (0.2%) |
| Coxiella burnetii                 | 206 (10%) | 182 | 19 | 5 | 1148 (59%) | 655 (33%) |
| Methanococcus jannaschii          | 178 (10%) | 146 | 27 | 5 | 1263 (71%) | 329 (19%) |
| Pyrococcus abyssi                 | 204 (12%) | 179 | 23 | 2 | 1511 (85%) | 50 (3%) |
| Aeropyrum pernix                  | 259 (10%) | 219 | 33 | 7 | 1039 (38%) | 1401 (52%) |
| Archaea                            |       |                     |                     |                     |                       |                        |
| Methanococcus jannaschii          | 178 (10%) | 146 | 27 | 5 | 1263 (71%) | 329 (19%) |
| Pyrococcus abyssi                 | 204 (12%) | 179 | 23 | 2 | 1511 (85%) | 50 (3%) |
| Aeropyrum pernix                  | 259 (10%) | 219 | 33 | 7 | 1039 (38%) | 1401 (52%) |
| Archaeoglobus fulgidus            | 236 (15%) | 210 | 22 | 2 | 911 (71%) | 1101 (14%) |

Table 1 for the entire table.
DOI: 10.1371/journal.pbio.0020081.t001

mately 400 amino acids (Pfam domain: RNB) defines a group of homologs distributed in a number of microbial genomes (Zuo and Deutscher 2001). Note that the C-terminal region is variable, and its E. coli homolog contains RNA-binding motifs (Zuo and Deutscher 2001). Although the detailed physiological roles of VacB remain unknown (Cheng and Deutscher 2002), the variable regions may contribute to the determination of substrate specificity of VacB in the RNA quality-control process that eliminates defective ribosomal RNA (rRNA) molecules in different species.

The number of SVGs increases as genome sizes vary, from 0.5 MB (Mycoplasma genitalium) to 8.6 MB (Streptomyces coelicolor) (Table 1). For most microorganisms included, the proportion of SVGs varies in the range of 10%–20%. The number of lineage-specific genes, on the other hand, does not appear to correlate with the genome size. Instead, it is influenced by the content of the database. For instance, a “minimal” genome, Methanococcus jannaschii, has a relatively high content of SVGs (20%) and a low percentage of lineage-specific genes (0.2%). However, when a closely related species, M. genitalium, is excluded from the database, its proportion of lineage-specific genes rises to 14%, while the proportion of SVGs remains unchanged. In general, the genomic proportion of SVGs is less affected by the database content.

Case Studies of SVGs and Functional Implication of Variability

In the following sections, we have selected several SVG families to demonstrate the functional implication of segmental variability.

Outer-membrane signal transduction genes/sensor histidine kinases. In prokaryotes, two-component signal-transducing systems are common and consist of a histidine kinase (HK) and a response regulator. Most HKs are membrane-bound, homodimeric proteins with an N-terminal periplasmic sensing domain and a C-terminal cytoplasmic kinase domain. HKs usually possess a highly variable sensing domain (usually over 150 amino acids), while the cytoplasmic kinase domain is quite conserved. By diversifying the sensing domain, microorganisms can develop different two-component modules to respond to different signals and interact with small molecules from the exterior. Figure 3 displays the distance matrix calculated from the sensing domains and the kinase domains from a group of highly similar HK genes. As shown in Figure 3, sensing domains are much more diverse than the kinase domains. Moreover, the two regions show distinct clustering patterns, of which only the one for the conserved kinase domains is close to the phylogenetic relationship inferred from 16S rRNA sequences (data not shown). Significant homologies in the sensing regions can only be found in closely related species (e.g., Ralstonia solanacearum [Rs] and Ralstonia metallidurans [Rm] in Figure 3), suggesting rapid divergence after speciation. Other sensor genes involved in cell motility, e.g., genes encoding methyl-accepting chemotactic protein (MCP) (see tlpA, tlpC in Figure 2), are also highly variable in their N-terminal domains. In several bacteria, e.g., Vibrio cholerae, there is a greater number of segmentally variable MCP genes (approximately 40) than in other genomes (see the gene list of V. cholerae at http://geneva.bu.edu), which must correspond to its expanded ability to detect different chemical signals and find favorable environments. Although a few conserved motifs have been detected in the sensing region (Galperin et al. 2001), the exact sensing signals for most prokaryotic HKs are unknown.

Transporter genes and outer-membrane proteins. The biggest family of SVGs is cell envelope-related, including the ATP-binding cassette transporters (ABC transporters), outer-membrane proteins, and virulence-related gene prod-
**Figure 3.** 2D Representation of the Distance Matrix Computed from the Variable and Conserved Domains in a Group of Similar HKs

The upper triangle shows the variable domains, the lower one the conserved domains. Amino acid sequence distances are calculated by the PROTDIST program using the Dayhoff PAM matrix. The sequence from each species is the best match (E-value < 1E-10) in that genome to the query *E. coli* gene. Abbreviations for organisms: *Ec*, *Escherichia coli* K12; *Ps*, *Pseudomonas syringae* pv. syringae B728a; *Rm*, *Ralstonia metallidurans* DSM1221; *Rs*, *Ralstonia solanacearum* ; *Li*, *Listeria innocua*; *Tm*, *Thermotoga maritime*; *Ml*, *Mycobacterium leprae*; *Nt*, *Nostoc* sp. PCC 7120; *Ef*, *Enterococcus faecalis*; *Bs*, *Bacillus subtilis*; *Ne*, *Nitrosomonas europaea*; *Sy*, *Synechococcus* sp. PCC 7942; *At*, *Agrobacterium tumefaciens*. The PROTDIST program is included in the PHYLIP software package version 3.5 (Felsenstein 1989).

DOI: 10.1371/journal.pbio.0020081.g003

...ucts. For membrane proteins, since part of their sequences are exposed to the outside of the cell and interact directly with the environment, one might hypothesize that the variable portions have evolved rapidly to deal with the changing environmental conditions.

ABC transporters are essential for microorganisms because they import nutrients into the cell and export noxious substances and toxins out of the cell. A typical ABC transporter gene in a prokaryote genome has a conserved ATPase domain (approximately 150 amino acids) and a large (over 300 amino acids) variable integral membrane domain. Two examples from this group are the multidrug-resistance genes *hetA* and *spaB* shown in Figure 2. It is known that substrates interact with the specific binding sites inside the membrane domain (Holland and Blight 1999), which suggests that the variability in the membrane domain may have to do with substrate selectivity or with different transport kinetics. Moreover, outer-membrane transporters are binding targets for bacteriophages and bacterial toxins. For example, the vitamin B12 transporter BtuB (614 amino acids) is the receptor for bacteriophage BF23 and E-coli (Bradbeer et al. 1976; Mohanty et al. 2003). The crystal structure of BtuB in *E. coli* has been solved (Chimento et al. 2003). The variable region in *E. coli* BtuB overlaps with the 22-strand β-barrel (position 150–360), while the N-terminal hatch domain (position 6–132) and the extreme C-terminal TonB-box domain (position 550–614) are conserved among many homologs (Figure S1). The extracellular loops between contiguous strands in the β-barrel are displayed outside the cell (Chimento et al. 2003) and possibly serve as receptor sites for bacteriophages and toxins. The variability in these loops may be driven by attempts to defend against bacteriophages and interaction with different bacterial toxins.

**DNA/RNA-processing enzymes.** DNA/RNA processing enzymes form another large family of SVGs. Characteristic examples are the restriction and modification enzymes, where the DNA methylases have a variable region designed for DNA sequence recognition (Cheng 1995) and the restriction enzymes are almost completely variable. Here we discuss two other genes: DNA gyrase B (*gyrB*) and DNA topoisomerase A (*topA*), whose competing actions control the degree of DNA supercoiling (Tse-Dinh et al. 1997). Schematic alignments anchored by the conserved motifs from the BLOCKS database (Henikoff et al. 1999) for both enzymes are shown in Figure 4. The variable region in *GyrB* is an additional approximately 160 amino acids long segment that is only present in the gram-negative eu-bacteria (Figure 4B). Experiments probing the role of this region in *E. coli* *GyrB* have demonstrated its involvement in DNA binding, although the detailed function is unknown (Chatterji et al. 2000). We suspect that variability in this inserted domain may determine the specificity of the interaction between *GyrB* and DNA or suggest interaction with other molecules. It is intriguing to see that other gyrase(s) lacking this region are also functional.

For *TopA*, the N-terminal region of approximately 600 amino acids shows extensive sequence similarity while the C-terminal region (over 100 amino acids) is variable both in sequence content and in length (Figure 4A). The conserved N-terminal region of *TopA* has the catalytic function of relaxing negatively supercoiled DNA (Feinberg et al. 1999). The variable C-terminus of *TopA* sometimes contains multiple copies of zinc-binding motifs, although there are a few exceptions, e.g., *TopA* in *Mycobacterium tuberculosis* CDC1551; *Nt*, *Nostoc* sp. PCC 7120; *Ef*, *Enterobacter* *faecalis*; *Bs*, *Bacillus subtilis*; *Ne*, *N. europaea*; *Sy*, *Synechococcus* sp. PCC 7942; *At*, *Agrobacterium tumefaciens*. The PROTDIST program is included in the PHYLIP software package version 3.5 (Felsenstein 1989).

DOI: 10.1371/journal.pbio.0020081.g003
replication errors, and conserved motifs can be observed (Derbyshire et al. 1995). However, other polymerases in the same family that lack such proofreading activity show much sequence divergence in this region (Derbyshire et al. 1995). The exact reason why sequence variability is observed in these polymerases is unknown.

Another interesting family is the aminoacyl-tRNA synthetases (AARS) (Ibba and SölII 2000). This family of genes is well known for its precision in substrate selection. The molecules known to interact with AARS include tRNA, amino acids, and ATP. Since the same amino acids and ATP molecules are found in all organisms, variability inside the AARS sequences must relate to the recognition and interaction with the tRNAs. Correspondingly, each AARS usually contains a conserved domain for catalysis and acceptor helix interaction and a nonconserved domain that interacts with the variable distal parts of its substrate tRNA (Schimmel et al. 1993). For instance, in bacterial-type prolyl-tRNA synthetase (ProRS), the N-terminal catalytic domain (approximately 200 amino acids) and the C-terminal anticodon-binding domain (approximately 150 amino acids) are highly conserved, while a less conserved region of about 180 amino acids is inserted between them (Figure S2). This variable domain shows similarity to the YbaK domain, which is thought to be involved in oligonucleotide binding (Zhang et al. 2000). Sporadic conserved residues in this region of E. coli ProRS are known to be involved in the posttransfer editing for mischarged Ala-tRNAPro (Wong et al. 2002). ProRS is also known to possess an inherent ability to mischarge cysteine (Ahel et al. 2002). Partial deletion of this variable region of E. coli ProRS results in a lower rate of proline acylation to cysteine acylation (Ahel et al. 2002), suggesting a possible role.
of substrate discrimination in this region. Thus, the variability in this inserted domain of ProRS appears to contribute to substrate recognition and the editing function of the enzyme. Intriguingly, ProRS in *Methanococcus jannaschii*, which does not have this inserted region, also possess editing abilities (Beuning and Musier-Forsyth 2001). As a result, there is a possibility that this region may have another unknown function, e.g., interaction with other undetected molecules.

**Carbohydrate active enzymes.** Variable regions exist in carbohydrate metabolizing enzymes, such as glycosyltransferases (GTs) and glycoside hydrolases (GHs), which respectively catalyze the biosynthesis of diverse glycoconjugates and their selective cleavage (Bourne and Henrissat 2001). Many pathogens express outer-membrane glycosylated oligosaccharides, which closely interact with the host environment (Saxon and Bertozzi 2001). For example, they even mimic host cell surface glycoconjugates to evade immune recognition (Persson et al. 2001). Both GTs and GHs have been classified into subfamilies based on sequence similarity (Bourne and Henrissat 2001). Structural studies on bacterial GTs from different subfamilies always reveal two-domain molecules, such as LgtC (Persson et al. 2001), GtfB (Mulichak et al. 2001), MurG (Hu et al. 2003), and SpsA (Charnock and Davies 1999), with one domain responsible for donor molecule (usually nucleotide-diphospho-sugar) binding and the other domain involved in acceptor sugar molecule binding. These genes exhibit great variability in the acceptor-binding domains and conservation in the donor-binding domains (see Figure S3 for the example of GtfB), which agrees with the relatively limited types of donor species (usually UDP/TDP-sugar) and their conserved binding modes, but a diversity of acceptor molecules (LgtC: lactose; GtfB: vancomycin aglycone; MurG: N-acetyl muramylpentapeptide; SpsA: unknown). Owing to the lack of homology in the acceptor binding domains, the substrate specificities encoded by these regions for most GTs are still unknown.

**Transcriptional regulators.** Prokaryotic transcriptional regulators form another large group of SVGs. Transcription regulators are usually two-domain proteins with one binding to DNA and one binding to ligand. The DNA-binding domains, which usually interact with DNA via helix–turn–helix, zinc-finger, or other modes, are more conserved than ligand-binding domains. Based on the characteristic conserved DNA-binding domains, transcriptional regulators can be classified into many different families (Nguyen and Saier 1995; Rigali et al. 2002). Even within each family, the ligand-binding domains are variable. For instance, the C-terminal regions involved in effector molecule binding and oligomerization (E-b/O) inside the GntR transcriptional regulator family are highly variable both in sequence content and in size (Rigali et al. 2002). The variability in the effector molecule-binding domains enables the transcriptional regulators to sense the presence of diverse ligands and signal the regulation of the downstream genes or operons accordingly. As in most previous cases, these variable regions remain functionally uncharacterized.

**Hypothetical genes.** In addition to genes with functional annotations, our method identifies a number of SVGs with unknown or hypothetical annotations in each genome (*H. pylori*: 17 genes; *N. meningitidis*: 32 genes; *V. cholerae*: 69 genes, etc.; see http://geneva.bu.edu for the full list). In contrast to lineage-specific hypothetical genes, these hypothetical genes contain conserved domains, which suggest their functional importance. Although most of the conserved domains in these hypothetical genes have currently unknown function, there are a few exceptions. Among them are the procaryotic mechanosensitive channel proteins, which respond to external osmotic pressure (Pivetti et al. 2003). Examples include the 343 amino acid long *E. coli* B1350 and 371 amino acid long *Bacillus subtilis* YhdY, both of which are currently annotated as “hypothetical.” However, they both have the characteristic domain of mechanosensitive proteins (Pfam domain: MS__channel). The central regions (approximately 150 amino acids) of these genes are conserved while both the N-terminal region (approximately 100 amino acids) and the C-terminal region (approximately 100 amino acids) are variable (see alignment in Figure S4). The conserved central region encodes three transmembrane segments, and the molecules are predicted to have their N-terminus outside and C-terminus inside the cell (Miller et al. 2003). Although the C-terminus is variable, the deletion experiments show that it is indispensable for stability and activity of this protein (Miller et al. 2003). It is tempting to hypothesize that the interacting partners for both N- and C-termini might vary in different organisms.

**Functional Classification of SVGs**

We are interested in probing the functional distribution of SVGs within a single genome. Are certain functional categories overrepresented? In Figure 5, we show a functional classification of SVGs in three microorganisms using 18 broad functional categories of the clusters of orthologous group (COG) database (Tatusov et al. 1997). We calculated the percentage (r in Figure 5) of SVGs within each functional class and the p-value of overrepresentation (Figure 5). Several functional categories are overrepresented (p-value < 0.01; see Figure 5 for details): (i) cell envelope biogenesis, outer membrane; (ii) DNA replication, recombination and repair; (iii) secondary metabolite biosynthesis, transport and catabolism; (iv) cell motility and secretion; (v) cell division and chromosome partitioning. Among them, only categories (i) and (ii) are overrepresented in all three genomes. Most functional categories involved in the basic metabolic processes are not significantly overrepresented or even underrepresented. The number of overrepresented categories and the order of significance differ from one genome to another, reflecting differences in genome content and presumably the relative importance of the different specific adaptations.

In Figure 6 we show the relative abundance of a set of SVG families in several microorganisms based on shared keywords in the annotations. The relative enrichments in several gene families for some microbes seem to correlate with the peculiarities of niche adaptation. In particular, *H. pylori* has more SVGs involved in cell motility and chemotaxis than other genomes with a similar genome size (*N. meningitidis, Streptococcus pneumoniae*). *H. pylori* is one of the few microbes that can colonize the highly acidic gastric environment (Tomb et al. 1997). The motility of *H. pylori* is crucial for its infectious capability and there is evidence that poorly motile strains are less able to colonize or survive in the host (O’Toole et al. 2000). *S. pneumoniae* has more carbohydrate-metabolizing enzymes, especially glycosyltransferases (GTs), which appear to be segmentally variable. The unique pattern of cell surface glycosylation in *S. pneumoniae* has been under
extensive investigation and plays an important role in pathogenesis (Tettelin et al. 2001). The GTs are responsible for making O-linked glycosylations on surface proteins, which coat the surface of the bacterium and interact with the host (Tettelin et al. 2001).

Gene Duplication and SVGs

Duplication followed by diversification is an efficient way of generating functional innovations (Prince and Pickett 2002). Regional sequence divergence has been observed between duplicated gene copies (Gu 1999; Dermitzakis and Clark 2001; Marin et al. 2001). We thus asked the following questions: (1) What is the distribution of paralogous genes in the set of SVGs in a single genome? (2) Is there a significant association between gene duplication and SVGs?

In Figure 7A, we show the distribution of paralogous genes among SVGs in several genomes. We consider paralogous genes to be similar genes in the same genome with a BLAST E-value less than 1E-5. As shown in Figure 7A, in H. pylori, N. meningitidis, and S. pneumoniae, the largest group of SVGs is the one with no paralogs. However, in E. coli, the largest group is the one with a single paralog. E. coli obviously has more paralogous genes in the SVG set, probably owing to a larger genome size by duplication. In Figure 7A (inset), we show the percentage of genes with different numbers of paralogs in each class for both segmentally variable and fully conserved genes in E. coli. Interestingly, over half of the fully conserved genes in E. coli do not have paralogs. There is a significant difference between the two distributions ($\chi^2$ test, $p$-value < 1E-5). In Figure 7B, we list the number of genes in a contingency table and test the significance using a $\chi^2$ test. For all genomes examined, there is a strong association between

---

**Table 1. Functional Classification of SVGs in Three Microorganisms**

| COG functional category | H. pylori   | N. meningitidis | S. pneumoniae |
|-------------------------|-------------|-----------------|--------------|
| Translation, ribosomal structure and biogenesis | 28/119     | 18/149         | 55/120       |
| Translation              | 5/23        | 8/57            | 22/197       |
| DNA replication, recombination and repair         | 38/108      | 1.3e-15         | 34/125       |
| Cell division and chromosome partitioning         | 6/20        | 0.16            | 8/20         |
| Posttranslational modification, protein turnover, chaperones | 15/58       | 1.8e-3          | 10/40        |
| Cell envelope biogenesis, outer membrane         | 38/107      | 1.1e-7          | 26/72        |
| Cell motility and secretion         | 22/73       | 0.91            | 7/28         |
| Inorganic ion transport and metabolism         | 10/40       | 0.05            | 8/59         |
| Signal transduction mechanisms         | 4/15        | 0.22            | 13/43        |
| Energy production and conversion         | 12/66       | 0.39            | 6/47         |
| Carbohydrate transport and metabolism         | 5/32        | 0.61            | 24/164       |
| Amino acid transport and metabolism         | 19/90       | 0.13            | 22/141       |
| Nucleotide transport and metabolism         | 4/34        | 0.83            | 3/81         |
| Coenzyme metabolism         | 12/63       | 0.32            | 7/71         |
| Lipid metabolism         | 5/37        | 0.74            | 6/43         |
| Secondary metabolites biosynthesis, transport and catabolism | 9/22        | 4.9e-3          | 6/27         |
| General function prediction only         | 18/108      | 0.20            | 26/101       |
| Function unknown          | 5/218       | 0.02            | 18/409       |

* M is the total number of genes in a COG broad functional category, and m is the number of SVGs within that category. r (r = m/M) is the proportion of SVGs in that category. The p-value is calculated using a hypergeometric distribution: let N = number of genes in the genome; n = number of SVGs identified; M = number of genes belonging to a particular category; m = number of SVGs belonging to a particular category:

\[
P = \sum_{i=0}^{m} \frac{M!}{i!(M-i)!} \frac{(N-M)!}{(N-i)!} \frac{N!}{n!} \left(\frac{i}{M}\right)^n \left(\frac{N-M}{N-i}\right) \left(\frac{n-i}{N-n+i}\right)^{N-n+i} \left(\frac{1}{N-n+i}\right).
\]

---

**Figure 6. Abundance of SVGs in Different Functional Categories in Five Microorganisms**

The approximate total gene number for each organism is as follows: H. pylori, 1,566 genes; S. pneumoniae, 2,094 genes; N. meningitidis, 2,065 genes; E. coli, 4,289 genes; B. subtilis, 4,100 genes. DOI: 10.1371/journal.pbio.0020081.g006
gene duplication and SVGs, suggesting an SVG is more likely to have originated from a duplicated gene.

Here we give an interesting example where one paralogous copy of a gene is segmentally variable and the other copy is fully conserved. In *H. pylori* strain 26695, gene products of HP1299 (253 amino acids) and HP1037 (357 amino acids) both have a conserved domain (approximately 250 amino acids; Pfam: Peptidase_M24) that is characteristic of the methionyl aminopeptidase (*map*) family (metalloprotease family M24) (Rawlings and Barrett 1995). HP1299 is fully conserved in a number of microbes and is homologous to the *E. coli* *map* gene (Figure S5), while the product of HP1037 has an extra N-terminal region (approximately 100 amino acids) that is variable among its similar genes (Figure S6). Additionally, HP1037 is annotated as a conserved hypothetical gene. The five residues found in the *E. coli* *map* that are involved in cobalt (Co$^{2+}$) binding (Asp-97, Asp-108, His-177, Glu-204, Glu-235; Rawlings and Barrett 1995), are conserved in both genes by examining the multiple alignment. These findings suggest that HP1037 might also encode a *map* activity and that its variable N-terminal region might be involved in additional functional roles, e.g., interactions with other molecules. In *Saccharomyces cerevisiae*, there are two *map* genes and both have an extra N-terminal region compared to the *E. coli* *map* gene. One copy of the yeast *map* gene contains zinc-finger motifs in the N-terminal region that are indispensable for in vivo function (Li and Chang 1995). A functional role involving interaction with the ribosome has also been suggested for this N-terminal domain (Vetro and Chang 2002). In most prokaryotes, it has been assumed that there is only one copy of the *map* gene. The SVG family exemplified by HP1037 may represent another family of *map* genes in prokaryotes.

**Discussion**

A major fraction of bioinformatics research on sequence analysis has focused on the conserved regions in proteins, trying to hypothesize the role of the protein by identifying sequence motifs that have been shown experimentally to correlate with a specific function. Some work has gone into cataloging the groups of lineage-specific proteins that show no similarity to other proteins in GenBank (Galperin and Koonin 1999), but there the route to assigning function usually needs experimental approaches requiring biochemistry or genetics or more rarely by determining the crystal structure of the gene product (Zhang et al. 2000). Unfortunately, current bioinformatics methods are only occasionally helpful in suggesting where to begin such studies.

In this paper we have initiated an effort to identify SVGs, which contain both well-conserved regions and highly variable regions. By looking carefully at a few specific examples where functional information is available from experimental data, we find that the variable region often seems to play a key role in mediating interactions with other molecules, both large and small. Sometimes the variable portions are involved in biological processes with a component of interaction between the cell and agents from the external environment. For instance, the DNA methyltransferases are part of a defense system that recognizes and clears invading foreign DNA; membrane-bound sensory HKs and mechanosensitive ion channels, etc., monitor changes of living conditions. Sometimes the variable portions are involved in intracellular processes that appear to have lineage-specific features. Thus, the variable regions inside DNA GyrB and several types of AARSs probably determine the specificity of substrate recognition. The detailed factors...
that introduce the molecular variability may go well beyond our explanations here and likely vary from case to case. Some variable regions may have diverged a long time ago and are now kept constant, while others may keep changing. In all of these cases, SVGs are exceptionally worthy targets of further experimental investigation, and such investigations can be greatly aided by the presence of the conserved regions that may suggest a preliminary function to be tested.

Why might certain genes contain these variable regions? Could they be simply relics left over during evolution and now serve no purpose? Are they just “pseudo-segments” with no function? There are several lines of evidence that support the hypothesis that when variable regions have been retained, they indeed serve a function. First, several studies have shown that deletions are, on average, more frequent than insertions (Halliday and Glickman 1991). As a result, if a region is evolving under weak functional constraints, it tends to get smaller over time (Lipman et al. 2002). Second, in a special case, one can imagine that when a variable region occurs at the C-terminus of a protein and is not being selected, it is likely to suffer random mutations including nonsense mutations or insertions/deletions that cause a shift in reading frame. Thus, we searched GenBank release 136.0 for examples of genes that matched the conserved region of an SVG, but in which the C-terminus was missing or much shorter. The DNA sequences downstream of such hits were examined for similarity to the variable region in the query gene. Of the 83 SVGs with a C-terminal variable region in H. pylori, none of them had hits with a disrupting stop codon in the variable region; 20 of them have hits with genes showing insertions/deletions that cause frame shifts in the variable region. However, the real number is likely to be much fewer, since, based on previous work, many of them may be the results of sequencing errors (Posfai and Roberts 1992).

In other cases, we find that some proteins have lost the variable segment in a subset of genomes. For instance, in ProRSs, the variable segment is absent in archaea; in GyrB, the variable segment is absent in the Gram-positive bacteria. Clearly in those cases the organisms can get by without the variable domain, although they may have a compensating function in a different gene. But this again does not imply that the variable region has no function in those genes that have retained it.

SVGs are distinct from sequences with shuffled domains (Doolittle 1995) in that the variable region is bounded by the same sets of conserved portions, while domain shuffling usually manifests itself in a different sequential order of conserved domains. We also hypothesize that the variable regions in SVGs are not the result of multiple domain fusion events, each resulting in an insertion of a different sequence into the protein. This hypothesis is supported by the fact that the fused domains are often conserved across multiple organisms (Marcotte et al. 1999). Additionally, our procedure requires that the variable regions are of similar length within a family of proteins, which are also restricted to conserved length distributions. This filter suggests a mutational mechanism that originated from an ancient protein. Indeed, it is possible that originally the variable region was a result of a single or possibly relatively few ancient fusion events, but this paper does not focus on the evolutionary origin of SVGs.

Another prediction from our observations is that the variable regions are excellent candidates to bind substrates or partner macromolecules. They may be extremely helpful in discovering the networks of protein–protein or protein–nucleic acid interactions within a cell. Bioinformatics may even be able to help in this endeavor by finding genes that seem to have coevolving variable regions as a result of such interactions. Experimental data from techniques such as the yeast two-hybrid system or microarrays may provide evidence for interactions that can involve two variable regions.

Much additional bioinformatics work will be needed to explore fully the potential of this method in hypothesizing function. For instance, the size limits we have arbitrarily imposed on the variable region should be tested systematically. In our relatively simple formulation presented here, the length of the variable region and the number of proteins in the same family that do not have an alignment to the variable region are the primary factors in determining its statistical significance. Methods using other sequence analysis tools, such as multiple alignment and sequence profiles, may provide alternative ways to identify segmental pattern of variability. A fundamental problem is to differentiate random evolutionary drift from positive selection correlated to functional requirements. Although one might expect that the N- and C-termini may be more variable than the regions in the middle, our data suggest that variable regions in SVGs are not preferentially located in either end (data not shown). We have also examined the amino acid composition, codon usage, and GC content in the variable regions and the conserved regions of the same SVG. While there is no significant deviation of amino acid composition and GC content between the two regions in general, codon usage appears to be biased in the variable regions (data not shown).

SVGs usually account for 10%–20% of the total genes in a microbial genome. Currently, we think of the class of lineage-specific genes as being the key factor that distinguishes one strain or species from another. The class of SVGs that we have defined in this paper must now be added to this collection of lineage-specific genes by virtue of the unique segments that constitute their variable regions. They also appear to provide functional elements that help to differentiate among strains and species. This point is well illustrated by considering the restriction-modification systems. Here, the DNA methyltransferases, which have a variable region responsible for DNA recognition, are members of the SVG class. With the help of their companion restriction endonucleases, which typically appear as lineage-specific genes, they serve to keep foreign, unmodified DNA sequences from entering the genome. In this case, the synergy of function provided by members of the two classes highlights the key role that both sets of genes must play in defining the individuality of a strain or species.

Our analysis to date is limited to prokaryotes and archaea where SVGs are transcribed and translated as contiguous genomic segments. In eukaryotes, alternative RNA splicing introduces substantial additional complexity into the interpretation of gene structure and protein product, thereby rendering impossible the simple analysis we have applied here. It is tempting to consider alternative splicing as a highly evolved control mechanism to introduce the variability we find in the SVGs and thereby achieve the functional diversity necessary for cell survival under different conditions. In eukaryotes, alternatively spliced exons can be introduced in response to the functional demands of different cell types by merely juggling protein coding regions in the genome.
by creating a SVG structure. If this view is correct, then it reinforces and highlights the importance of these SVGs to the workings of the cell.

In this paper we have provided an initial glimpse of SVGs, which appear to provide an important genetic layer in the adaptation of cells to novel environments and hazardous pathogens. We have focused attention on the biological significance of these genes, especially those that have highly diverged segments. We are currently trying to develop a more refined classification of these genes so as to explore the functional significance of the variability. We would like to know whether extreme variability is required for diverse function or whether more modest variation is sufficient. Such questions require that we can first distinguish positive selection acting on these variable regions from neutral evolution leading to gene decay and eventual loss. Since the variable regions we report are often not amenable to current tools available for alignment, we are exploring new methods that will help us to assess whether positive selection is driving the evolution of these genes.

In summary, we have identified an extremely useful way of classifying genes that leads to the identification of those with a high priority for both experimental and computational research.

Materials and Methods

Our method for detecting SVGs includes several steps: (1) identification of similar genes followed by query-anchored multiple alignment using nongapped BLAST (Altschul et al. 1990); (2) taxonomy clustering of similar genes to avoid bias; (3) detection of segmental variability.

Identification of similar genes. Given a gene, we start by searching for all its similar genes in the nonredundant database (GenBank release 136.0, 15 June 2003) using nongapped BLAST (Altschul et al. 1990). We use the nongapped BLAST because the gapless high scoring pairs (HSPs) reported are rather conservative. The gapped BLAST, however, tends to extend HSPs over variable regions, which has been observed in several examples (e.g., DNA-recognition domain in cytosine-specific methyltransferase; data not shown). Two criteria are used to define close similarity. First, the E-value is less than 1E-10. Here we use a strict E-value threshold to avoid possible functional divergence among the homologs. Accordingly, we use the BLOCKS and PRINTS databases are relatively conservative in (Henikoff et al. 1999), and PRINTS (Attwood et al. 2003) databases. The BLOCKS and PRINTS databases are relatively conservative in defining motifs. However, the Pfam domain seems to include the variable region within the conserved region, as shown in Figure 1A.

Supporting Information

Data Deposit We provide a static collection of segmentally variable genes at our Web site, http://geneva.bu.edu. SVGs for several representative genomes are listed there. For SVG lists in other genomes, please request more information from Y. Zheng at geneva.bu.edu. All the case examples mentioned throughout the paper and Supporting Information have been compiled into one Web page, http://geneva.bu.edu/paper03.html, with hyperlinks. Readers can follow each hyperlink to access additional information from Pfam, BLOCKS, PRINTS, COG, and nongapped BLAST for each gene.

Figure S1. Multiple Alignment of BtuB and Homologs

Conservation score is plotted under the alignment (ClustalX). The conserved portions are as follows: N-terminal domain, extreme C-terminal domain, and a segment between N-terminal and C-terminal domain. The variable domain (between N-terminal and C-terminal) overlaps with the transmembrane 22-strand β-barrel regions. Found at DOI: 10.1371/journal.pbio.0020081.sg001 (2.69 MB EPS).

Figure S2. Query-Anchored Alignment of ProRS

The query protein is H. pylori ProRS. The blue segments are HSPs reported by nongapped BLAST. The yellow segments are the variable
Figure S5. Multiple Alignment for *H. pylori* Gene HP1299

It is the methionine aminopeptidase (type Ia map). This is an example of a fully conserved gene.

Figure S6. Multiple Alignment for *H. pylori* Gene HP1037

It is currently annotated as “conserved hypothetical protein.” The N-terminal region is variable. The conserved C-terminal domain is characteristic of methionine aminopeptidase.

Table S1.

| Gene   | Accession | Species | Organism | Description |
|--------|-----------|---------|----------|-------------|
| HP1037 | (2314181) | *H. pylori* | *Helicobacter pylori* | Conserved hypothetical protein |

References

Abel I, Stathopoulos C, Ambrogelly A, Sauerwald A, Toogood H, et al. (2002) Characterization of the methyltransferase domain of the adaptive layer of the microbial proteome. Mol Biol Evol 19: 3733–3742.

Cheng ZF, Deutscher MP (2002) Purification and characterization of the repressor of the methionine aminopeptidase operon. J Biol Chem 277: 47443–47448.

Abumara A, Rokitta D, Yehuda R, Ben-Yosef D, et al. (2000) Characterization of the methyltransferase domain of the adaptive layer of the microbial proteome. J Biol Chem 275: 47443–47448.

Cheng ZF, Deutscher MP (2002) Purification and characterization of the repressor of the methionine aminopeptidase operon. J Biol Chem 277: 47443–47448.

Abumara A, Rokitta D, Yehuda R, Ben-Yosef D, et al. (2000) Characterization of the methyltransferase domain of the adaptive layer of the microbial proteome. J Biol Chem 275: 47443–47448.
fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. J Mol Biol 293: 381–399.

Hu Y, Chen L, Ha S, Gross B, Falcone B, et al. (2003) Crystal structure of the MurGUDP-GlcNAc complex reveals common structural principles of a superfamily of glycosyltransferases. Proc Natl Acad Sci U S A 100: 845–849.

Ibba M, Soll D (2000) Aminoacyl-tRNA synthesis. Annu Rev Biochem 69: 617–650.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Microevolutionary genomics of bacteria. Theor Popul Biol 61: 435–447.

Lange C, Wild C, Trautner TA (1996) Identification of a subdomain within DNA-(cytosine-C5)-methyltransferases responsible for the recognition of the 5’ part of their DNA target. EMBO J 15: 1445–1450.

Li X, Chang YH (1995) Amino-terminal protein processing in Saccharomyces cerevisiae is an essential function that requires two distinct methionine aminopeptidases. Proc Natl Acad Sci U S A 92: 12357–12361.

Lin LF, Posfai J, Roberts RJ, Kong H (2001) Comparative genomics of the restriction-modification systems in Helicobacter pylori. Proc Natl Acad Sci U S A 98: 2749–2754.

Lipman DJ, Souvorov A, Koonin EV, Panchenko AR, Tatusova TA (2002) The relationship of protein conservation and sequence length. BMC Evol Biol 2: 20.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein–protein interactions from genome sequences. Science 285: 751–753.

Marin I, Fares MA, Gonzalez-Candelas F, Barrio E, Moya A (2001) Detecting changes in the functional constraints of paralogous genes. J Mol Evol 52: 17–28.

Miller S, Bartlett W, Chandrasekaran S, Simpson S, Edwards M, et al. (2003) Microevolutionary genomics of bacteria. Theor Popul Biol 61: 435–447.

Moorhead D, Mattei MA, Pachter L, Schuh AL, et al. (2000) Representation of pyrimidine dimers in X-ray crystal structures of DNA. Proc Natl Acad Sci U S A 97: 12145–12150.

Mullighan CG, Young DS, Dohner H, Peabody JS, et al. (2004) The t(4;11)(q21;q23) t(6;9)(p23;q34) translocations in acute myeloid leukemia frequently alter the NOTCH1 gene. J Clin Invest 113: 340–348.

Muruganujan A, Azevedo RB, Boore JL, Lopez R, et al. (2008) OrthoVenn: Venn diagrams for multiple ortholog sets. Nucleic Acids Res 36: W552–W557.

Muroi T, Takei N, Goto M, Ohmi H, et al. (2000) The crystal structure of the Escherichia coli OmpC osmoporin reveals a striking similarity with aquaporins. EMBO J 19: 2695–2702.

O’Toole PW, Lane MC, Porwollik S (2000) Helicobacter pylori motility. Microbes Infect 2: 1207–1214.

Persson K, Ly HD, Dieckmann M, Waack AR, Withers SG, et al. (2001) Crystal structure of the retaining galactosyltransferase LgtC from Neisseria meningitidis in complex with donor and acceptor sugar analogs. Nat Struct Biol 8: 166–175.

Pivetti CD, Yen MR, Miller S, Busch W, Tseng YH, et al. (2003) Two families of mechanosensitive channel proteins. Microbiol Mol Biol Rev 67: 66–85.

Posfai J, Roberts RJ (1992) Finding errors in DNA sequences. Proc Natl Acad Sci U S A 89: 4698–4702.

Posfai J, Bhagwat AS, Posfai G, Roberts RJ (1989) Predictive motifs derived from cytosine methyltransferases. Nucleic Acids Res 17: 2421–2435.

Prince VE, Pickett FB (2002) Splitting pairs: The diverging fates of duplicated genes. Nat Rev Genet 3: 827–837.

Raven JD, Barrett AJ (1995) Evolutionary families of metallopeptidases. Methods Enzymol 248: 183–238.

Rigali S, Nantiola P, Giannotta F, Dusart J (2002) Subdivision of the hexin-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. J Biol Chem 277: 12507–12515.

Roche A, McFadden J, Owen P (2001) Antigen 43, the major phase-variable protein of the Escherichia coli outer membrane, can exist as a family of proteins encoded by multiple alleles. Microbiology 147: 161–169.

Saxon E, Bertozzi CR (2001) Chemical and biological strategies for engineering cell surface glycosylation. Annu Rev Cell Dev Biol 17: 1–25.

Schimmel P, Giege R, Moras D, Yokoyama S (1993) An operational RNA code for amino acids and possible relationship to genetic code. Proc Natl Acad Sci U S A 90: 8765–8768.

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 651–657.

Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, et al. (2001) Complete genome sequence of a virulent isolate of Streptococcus pneumoniae. Science 293: 498–506.

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388: 539–547.

Tse-Dinh YC, Qi H, Menzel R (1997) DNA supercoiling and bacterial adaptation: Thermotolerance and thermoresistance. Trends Microbiol 5: 325–326.

Vetro JA, Chang YH (2002) Yeast methionine aminopeptidase type 1 is ribosome-associated and requires its N-terminal zinc finger domain for normal function in vivo. J Cell Biochem 85: 678–688.

Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, et al. (2003) Database resources of the National Center for Biotechnology. Nucleic Acids Res 31: 28–35.

Wong FC, Beuning PJ, Nagan M, Shiba K, Musier-Forsyth K (2002) Functional role of the prokaryotic proline-tRNA synthetase insertion domain in amino acid editing. Biochemistry 41: 7108–7115.

Zhang H, Huang K, Li Z, Banerjee L, Fisher KE, et al. (2000) Crystal structure of YabK, protein from Haemophilus influenzae (H1434) at 1.8 A resolution: Functional implications. Proteins 40: 86–97.

Zuo Y, Deutscher MP (2001) Exoribonuclease superfamilies: Structural analysis and phylogenetic distribution. Nucleic Acids Res 29: 1017–1026.