Deletional Bias across the Three Domains of Life

Chih-Horng Kuo and Howard Ochman

Department of Ecology & Evolutionary Biology, University of Arizona

Elevated levels of genetic drift are hypothesized to be a dominant factor that influences genome size evolution across all life-forms. However, increased levels of drift appear to be correlated with genome expansion in eukaryotes but with genome contraction in bacteria, suggesting that these two groups of organisms experience vastly different mutational inputs and selective constraints. To determine the contribution of small insertion and deletion events to the differences in genome organization between eukaryotes and prokaryotes, we systematically surveyed 17 taxonomic groups across the three domains of life. Based on over 5,000 indel events in noncoding regions, we found that deletional events outnumbered insertions in all groups examined. The extent of deletional bias, when measured by the total length of insertions to deletions, revealed a marked disparity between eukaryotes and prokaryotes, whereas the ratio was close to one in the three eukaryotic groups examined, deletions outweighed insertions by at least a factor of 10 in most prokaryotes. Moreover, the strength of deletional bias is associated with the proportion of coding regions in prokaryotic genomes. Considering that genetic drift is a stochastic process and does not discriminate the exact nature of mutations, the degree of bias toward deletions provides an explanation to the differential responses of eukaryotes and prokaryotes to elevated levels of drift. Furthermore, deletional bias, rather than natural selection, is the primary mechanism by which the compact gene packing within most prokaryotic genomes is maintained.

Introduction

The genome sizes of cellular organisms span at least six orders of magnitude (Gregory 2005), but the evolutionary and functional basis of this variation remains unclear. Early studies detected relationships between genome size and several phenotypic traits, such as generation time (Bennett 1972), cell and nuclear volume (Cavalier-Smith 1982), duration of mitosis and meiosis (Bennett 1987), embryonic developmental time (Jockusch 1997), and plant seed or leaf size (Chung et al. 1998). Based on such correlations, genome size was hypothesized to be under selective constraints (Gregory 2002); however, comparative studies in bacteria have failed to support such adaptive view as a general explanation. For example, although the streamlined genomes of bacteria are often regarded as an adaptation for rapid cell growth, bacterial replication rates are not correlated with genome size either within (Mikkola and Kurland 1991; Bergthorsson and Ochman 1998) or among species (Mira et al. 2001; Froula and Francino 2007). The only exception appears to be from nutrient-limited marine bacteria (Dufresne et al. 2005; Giovannoni et al. 2005), which may be under selection for reduced cell volume.

It has recently been posited that the overall size and structure of genomes are determined mainly by a nonadaptive, population-level process, namely random genetic drift (Lynch and Conery 2003). Because the accumulation of slightly deleterious mutations is facilitated by an increase in drift, lineages with relatively small effective population sizes (e.g., mammals) tend to have large genomes due to the proliferation of transposable elements and the lengthening of introns (Lynch and Conery 2003; Lynch 2006a). In contrast, lineages with relatively large population sizes (e.g., most free-living bacteria) would be expected to have more streamlined genomes on account of more effective selection against unnecessary or slightly deleterious sequences, which limits the accumulation of selfish and noncoding DNA (Lynch 2006b). Although this model provides a straightforward and seemingly unifying explanation for the evolution of genome size across all life forms, it does not explain the variation in the most genetically diverse group of organisms on the planet. Contrary to the predictions of this model, the strength of drift is "negatively" correlated with genome size in Bacteria (Kuo et al. 2009; Novichkov et al. 2009), with those bacteria the lifestyles of which cause the most dramatic reductions in effective population size having the most reduced genomes (Moran and Plague 2004; Nakabachi et al. 2006).

Because genetic drift facilitates the fixation of slightly deleterious mutations, the difference between the effects of drift on the size of eukaryotic and bacterial genomes is most likely rooted in the mutational input. Previous studies that examined small-scale indels (ranging from single to several hundred nucleotides) in pseudogenes or other nonfunctional elements revealed that deletions prevail over insertions across a wide range of taxonomic groups, including Archaea (von Passel et al. 2007), Bacteria (Andersson JO and Andersson SGE 2001; Mira et al. 2001), nematodes (Robertson 2000), insects (Petrov et al. 1996, 2000; Petrov and Hartl 1998; Bensasson et al. 2001), and mammals (Graur et al. 1989). Mutation accumulation experiments on a few model organisms offer a slightly different view: A preponderance of deletions has been observed in the bacterium *Salmonella enterica* (Nilsson et al. 2005), whereas insertions outnumbered deletions in the nematode *Caenorhabditis elegans* (Denver et al. 2004). Based on these observations, a general mutational bias toward deletions coupled with the genome-wide effects of genetic drift have been hypothesized as the major factors contributing to genome size evolution (Petrov et al. 2000; Petrov 2002; but see Gregory 2003, 2004; Vinogradov 2004).

Unfortunately, the extent to which eukaryotes and prokaryotes differ with respect to their deletional bias is unclear, mainly because the methods used to identify indels vary widely across studies and the taxon sampling in individual studies was rather limited. By taking advantage of the large collection of genome sequences available, we examined a diverse set of lineages to directly compare the impact of mutational input on genome evolution across the domains of life.

Key words: genome evolution, genome size, mutational spectra, organismal complexity, indels.

E-mail: hochman@email.arizona.edu.

Genome. Biol. Evol. Vol. 2009:145–152.
doi:10.1093/gbe/evp016
Advance Access publication June 27, 2009

Table 1
Summary of Taxon Sampling

Domain[a]	Phylum	Genus	Sampled Genomes[b]	NCBI Genome ID
A	Crenarchaeota	*Sulfolobus*	((*S. tokodaii*, *S. acidocaldarius*), *S. solfataricus*)	((246, 13935), 108)
A	Euryarchaeota	*Methanococcus*	((*M. maripaludis* S2, *M. maripaludis* C6), *M. vannielii*);	((10632, 19639), 17889)
B	Actinobacteria	*Mycobacterium*	((*M. tuberculosis*, *M. marinum*), *M. avium*)	((15642, 16725), 88)
B	Chlorobi	*Chlorobium*	((*C. limicola*, *C. phaeobacteroides*), *C. phaeovibrioides*)	((12606, 12609), 12607)
B	Cyanobacteria	*Synechococcus*	((*S.* sp. CC9605, *S. sp.* CC9902), *S.* sp. WH 8102)	((13643, 13655), 230)
B	Firmicutes	*Bacillus*	((*B. subtilis*, *B. amyloliquefaciens*), *B. licheniformis*)	((76, 13403), 12388)
B	Proteobacteria	*Bartonella*	((*B. quintana*, *B. henselae*), *B. bacilliformis*)	((44, 196), 16249)
B	Proteobacteria	*Rickettsia*	((*R. prowazekii*, *R. typhi*), *R. canadensis*)	((43, 10679), 12952)
B	Proteobacteria	*Wolbachia*	((*W. w*Mel, *W. w*Pip), *W. w*Bm)	((272, 30313), 12475)
B	Proteobacteria	*Neisseria*	((*N. meningitidis* FAM18, *N. meningitidis* 053442), *N. gonorrhoeae*)	((255, 16393), 23)
B	Proteobacteria	*Geobacter*	((*G. metallireducens*, *G. sulfurreducens*), *G. uraniireducens*)	((177, 192), 15768)
B	Proteobacteria	*Buchnera*	((*B. aphidicola* APS, *B. aphidicola* Sg), *B. aphidicola* Bp)	((245, 312), 256)
B	Proteobacteria	*Escherichia*	((*E. coli* K12, *E. coli* EDL933), *E. coli* CFT073)	((225, 259), 313)
B	Spirochetes	*Borrelia*	((*B. turicatae*, *B. recurrentis*), *B. burgdorferi*)	((13597, 18233), 3)
E	Chordata	*Homo*/*Pan*	((*Homo sapiens*, *Pan troglodytes*), *Pongo pygmaeus*)	NA
E	Arthropoda	*Drosophila*	((*D. sechellia*, *D. simulans*), *D. melanogaster*)	NA
E	Ascomycota	*Saccharomyces*	((*S. cerevisiae*, *S. paradoxus*), *S. mikatae*)	NA

NA, Not applicable.
[a] A: Archaea; B: Bacteria; E: Eukaryota.
[b] Parentheses denote the phylogenetic grouping of taxa in standard Newick tree format.

Materials and Methods

Prokaryotes: Archaea and Bacteria

To assemble data sets for examining deletional bias in these microbial taxa, we selected sets of three genomes (representing separate strains or species depending on the particular taxonomic group) from each group. In order to infer neutral indels in highly degraded pseudogenes, we required the divergence level among the three lineages to be low enough to achieve unambiguous alignments but to have incurred an ample number of indels. All of the archaeal and bacterial genomes used in this study were downloaded from NCBI GenBank (Benson et al. 2008) on 4 December 2008. The Genome Project IDs are listed in table 1. Data parsing and processing were performed with a set of custom Perl scripts written with Bioperl modules (Stajich et al. 2002).

For each group, we began by identifying single-copy orthologs that are shared among all three genomes. These conserved single-copy genes served as anchors to delineate orthologous noncoding regions from which indels could be identified. Sets of orthologous genes were recovered with OrthoMCL (Li et al. 2003), which is a clustering algorithm largely based on all-against-all BlastP (Altschul et al. 1990) hits and has been shown to perform well by a benchmarking study (Hulsen et al. 2006). As a conservative inference of orthology, the BlastP *e*-value cutoffs were set at 1×10^{-15}.

After identifying conserved single-copy genes, we screened the genome for lineage-specific pseudogenes, recognized as protein-coding regions that are disrupted or truncated in only one of the three taxa and are flanked by two conserved single-copy genes. To ensure the quality of alignments, we also required the pseudogene regions between the two conserved flanking genes to be at least 50 bp in length.

The rationale for focusing on pseudogenes in Archaea and Bacteria is based on the fact that the ancestral state of such regions can be confidently inferred from the conserved, single-copy homologs that are uninterrupted in the other two genomes; thus, even a high rate of recombination, as observed among some closely related bacteria (Touchon et al. 2009), would not affect our classification of each event as either an insertion or a deletion. Considering that the first indel to disrupt an open reading frame might not be neutral, we examined only those pseudogenes that contained at least three indels. Pseudogenes that have incurred this number of indels are often unrecognizable by sequence-similarity searches but can be readily identified using our synteny-based approach. Importantly, this method also eliminates the latent bias toward detecting deletions when using full-length open reading frames to search for fragmented pseudogenes.

Orthologous regions that were identified by the described approach were aligned in Muscle (Edgar 2004) using default parameters. To improve alignment quality, sequence alignments incorporated the entire region including the adjacent flanking genes, which were not subjected to indel analysis. Indels specific to one taxon were identified by a custom Perl script; all indels were then manually curated by visual inspection in Jalview (Waterhouse et al. 2009), and poorly aligned regions were excluded.

Eukaryotes: Primates, Flies, and Yeasts

Data sets of indels for the three groups of eukaryotes were constructed and analyzed using approaches similar to those used for Archaea and Bacteria, with the differences noted below:

Table 2
Summary of Indel Statistics

Domain[a]	Genus	Noncoding Regions Examined		Insertional Events		Deletional Events	
		Number	Total Length (bp)[b]	Number	Total Length (bp)	Number	Total Length (bp)
A	*Sulfolobus*	14	9,210	20	173	69	2,015
A	*Methanococcus*	21	12,925	45	178	180	7,514
B	*Mycobacterium*	28	17,017	21	121	136	6,089
B	*Chlorobium*	19	6,440	11	33	123	7,090
B	*Synechococcus*	14	4,761	15	64	79	3,334
B	*Bacillus*	50	15,682	79	318	339	15,195
B	*Bartonella*	34	30,775	123	1,433	370	11,582
B	*Rickettsia*	16	17,606	64	345	158	2,823
B	*Wolbachia*	10	10,915	31	471	35	507
B	*Neisseria*	13	9,728	24	216	84	4,016
B	*Geobacter*	22	3,401	9	35	121	7,516
B	*Buchnera*	29	23,111	105	676	377	9,334
B	*Escherichia*	12	4,909	8	9	59	8,223
B	*Borrelia*	18	6,3532	17	59	199	9,335
E	*Homo/Pan*	136	1,182,162	235	3,343	412	1,610
E	*Drosophila*	170	235,213	204	1,372	385	2,582
E	*Saccharomyces*	99	167,335	374	980	801	2,168

[a] A: Archaea; B: Bacteria; E: Eukaryota.
[b] Sequence length in focal lineages before alignment.

(1) Due to the lack of robust (or any) gene annotations in several of the eukaryotic genomes available from GenBank, we obtained each of the three eukaryote data sets from alternate databases. Data on primate genomes, including human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), and orangutan (*Pongo pygmaeus*), were retrieved from Ensembl (Hubbard et al. 2009) release 52; *Drosophila* genomes, including *Drosophila melanogaster*, *Drosophila sechellia*, and *Drosophila simulans*, were downloaded from FlyBase (Tweedie et al. 2009) version FB2009_01; the *Saccharomyces* data set, including *Saccharomyces cerevisiae*, *Saccharomyces mikatae*, and *Saccharomyces paradoxus*, was extracted from the *Saccharomyces* Genome Database (Christie et al. 2004) on 27 January, 2009.
(2) To minimize the effects of paralogs in the identification of single-copy orthologs, we applied a more stringent *e*-value cutoff of 1×10^{-25} in the BlastP step.
(3) The organization of most eukaryotic genomes makes it problematic to identify pseudogenes and their corresponding orthologs, so we focused instead on other classes of noncoding regions, that is, introns or intergenic regions that can be readily aligned among species. Because of the low level of recombination among species and the availability of well-established phylogenies, we utilized an outgroup to infer ancestral states and to establish the polarity of all indels that are specific to only one of the two ingroup lineages. For primates and *Drosophila*, we selected single-copy genes with exactly one intron in all three species because the orthology among such introns can be established unequivocally. We imposed lower and upper limits on intron lengths because indels in extremely short introns may not be neutral and extremely long introns might prove difficult to align. For primates, we examined introns that were 1–20 kb in length in all three species; for *Drosophila*, we set the range to 0.2–10 kb. When examining introns, we included the two flanking exons (instead of genes) to ensure quality of the alignments.
(4) Due to the paucity of introns in the *Saccharomyces* genomes, we examined the intergenic regions that are flanked by two conserved single-copy genes. Because regulatory elements might constitute a significant fraction of short intergenic regions (and thus the indels are more likely to have a fitness effect and not represent neutral events), we excluded intergenic regions shorter than 600 base pairs (bp) in any of the three species considered.

Results

We sampled 17 broadly divergent taxonomic groups, each containing an extensive collection of genome sequences (table 1), and for each group, we selected three lineages that are closely related such that orthologous noncoding regions can be unambiguously aligned. The alignments allowed us to infer the exact boundaries and ancestral state of indels within these noncoding regions, which together provide robust estimates of the mutational input of base pair- to kilobase-sized insertions and deletions to these genomes. Note that because we focused on pseudogenes that had accumulated multiple indels in archaea and bacteria and on long noncoding regions in eukaryotes, the overwhelming majority of indel events can be considered neutral and therefore represent the background pattern of mutations in these genomes.

Our results revealed a pervasive bias toward deletions in all taxonomic groups examined, although the extent of bias was substantially lower in the eukaryotic lineages considered (fig. 1). Deletions outnumber insertions in all groups examined, with the extremes observed in Bacteria: The ratio of insertions to deletions ranges from a low of 0.07 in *Geobacter* to nearly 0.9 in *Wolbachia* (fig. 1 and table 2). With the exception of Primates, all sampled

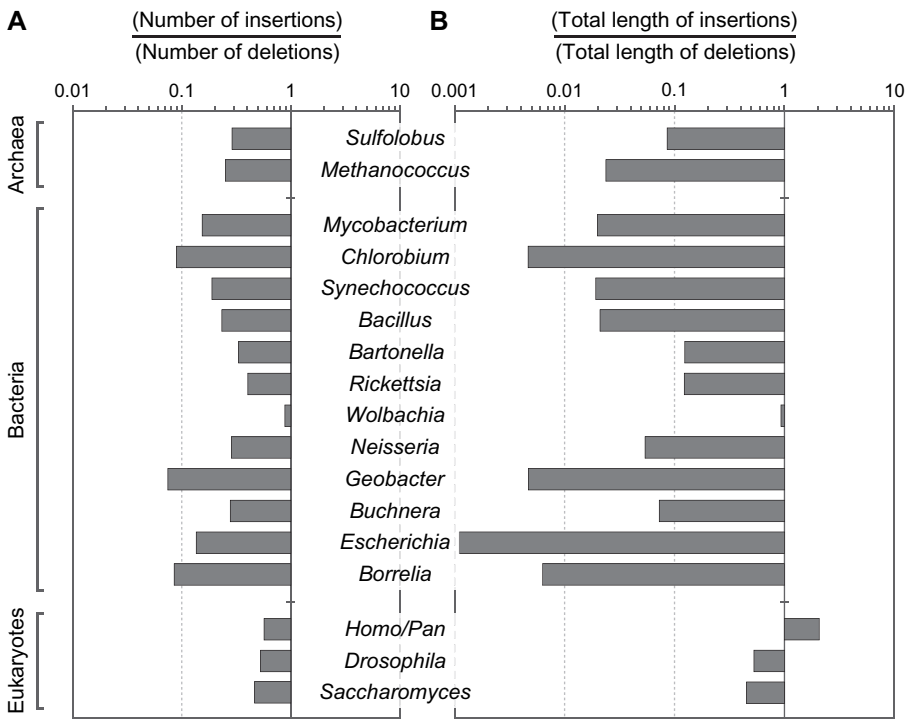

FIG. 1.—Extent of indel bias in cellular genomes. (*A*) Ratios of deletion to insertion events. A ratio of less than one indicates a bias toward deletions. (*B*) Indel bias based on the total length of DNA gained and lost. A ratio of less than 1 indicates a bias toward DNA loss.

taxonomic groups experienced a net loss of DNA through small indels. For each bp removed from a genome through deletions, prokaryotes gained from 0.001 bp in *Escherichia coli* to 0.93 bp in *Wolbachia* through insertions; in contrast, *Saccharomyces* and *Drosophila* gained 0.45 bp and 0.53 bp, respectively, whereas primates gained 2.08 bp for each bp removed by deletions (fig. 1 and table 2). In fact, the observed biases toward deletions are likely to be underestimates: Several deletions were excluded because we required inference of the exact ancestral state, and in the majority of cases, the focal lineage possessed a deletion that was >50 bp, but the exact length of this deletion could not be established because a shorter indel was present in the corresponding region in the other two lineages.

There is a clear difference in the length distribution of indels among three domains, which contributes to the disparity in genomes sizes between prokaryotes and eukaryotes (fig. 2). In Archaea and Bacteria, deletions are more frequent, and on average longer, than insertions, which results in the strong bias toward DNA loss (fig. 2*A* and *B*). In contrast, the length distributions of insertions and deletions in eukaryotes are not markedly different, with the majority of observed indels in the 1–10 bp range (fig. 2*C*). Of the three eukaryotic groups, single bp indels account for 46% of the observed indels in primates, 40% in *Drosophila*, and over 60% in *Saccharomyces* (supplementary fig. S1, Supplementary Material online).

Archaea and Bacteria

The level of bias toward deletions varies considerably among the prokaryotic genomes examined (fig. 1), allowing us to test two hypotheses concerning the role of deletional bias in genome evolution. First, as the bias toward deletions increases, one expects a more rapid deterioration of nonfunctional regions, resulting in the more compact packing of genes within a genome. Consistent with this hypothesis, gene density (i.e., the proportion of a genome that consists of annotated genes) among prokaryotes is significantly correlated with strength of deletional bias (fig. 3*A* $r = -0.76$, $P = 0.0015$). Second, in that overall genome size in prokaryotes is largely a function of the number of genes in the genome (Mira et al. 2001; Giovannoni et al. 2005; Kuo et al. 2009), we expect little association between the extent of deletional bias in noncoding regions and overall genome size. Because this association borders the conventional significance threshold (fig. 3*B* $r = -0.52$, $P = 0.054$), a more extensive taxon sampling would be necessary to further test this hypothesis.

We note that the genera with the weakest biases toward deletions are members of the alphaproteobacteria (i.e., *Bartonella*, *Rickettsia*, and *Wolbachia*). Although each of these groups forms obligate associations with eukaryotic hosts, it is unlikely that this lifestyle alone or the age of the association with their respective hosts can explain the observed pattern. The extent of deletional bias in other obligate pathogens (e.g., *Borrelia* and *Neisseria*) and endosymbionts (i.e., *Buchnera*) span much of the observed range. Therefore, diminished biases toward deletions are probably taxonomic characteristic of this bacterial group.

Eukaryotes

The mutational input in *Saccharomyces* is dominated by small indels. Among 1,175 indels recognized in 99

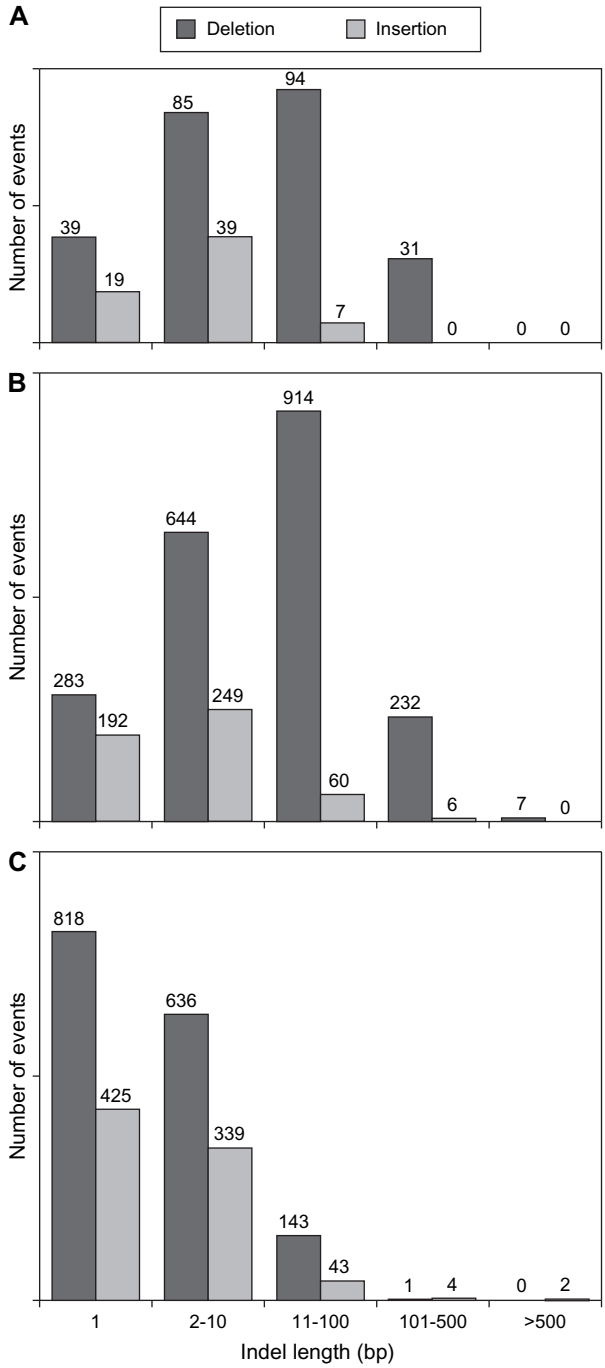

FIG. 2.—Length distribution of small indels across the three domains of life. (*A*) Archaea, (*B*) Bacteria, and (*C*) Eukaryotes.

intergenic regions, the longest insertion was only 65 bp and the longest deletion was 73 bp. Although the length distributions of insertions and deletions do not differ in *Saccharomyces* (supplementary fig. S1, Supplementary Material online), deletions outnumbered insertions, resulting in a net loss of DNA.

In contrast to prokaryotes and *Saccharomyces* (both of which lack long insertions), transposable elements provide a major source of DNA gains in primates and *Drosophila*. Although our analyses in these two eukaryotes were restricted to orthologous introns, which favor the identification of shorter indels, we detected one 1,102-bp insertion in *P. troglodytes* genome (containing two LINE and one SINE) and one 703-bp insertion in *D. sechellia* (containing a FB4 element). Despite their rare occurrences, these insertions of transposable elements offset the loss of DNA through frequent small deletions; and in fact, in the case of the primates, such rare long insertions are sufficient to result in a net gain of DNA in introns.

Discussion

The mutational input of insertions and deletions to a genome, as measured either by the number of events or the total length of DNA segments, is inherently biased toward deletions across a wide range of taxonomic groups representing the three domains of life. With the exception of alphaproteobacteria, the deletional biases in prokaryotes were at least one order of magnitude higher than those observed in eukaryotes (fig. 1). Although the prevalence of transposable elements in primates and *Drosophila* contribute to this difference, the indel pattern in *Saccharomyces* suggests that eukaryotic genomes have lower intrinsic rates of DNA loss through small indels. In spite of the limitation on taxon sampling imposed by the current availability of eukaryotic genome sequences, the strong differences in mutational input observed between prokaryotes and eukaryotes have played the major role in shaping the genome size and organization within these two groups.

Limitation on Taxon Sampling

Despite recent increases in sequence databases, the availability of genome sequences from closely related lineages remains the limiting factor in making reliable comparisons among divergent taxa. In addition to requiring a set of three genomes for each group, our analyses also demanded that their divergence levels be within a fairly narrow range, low enough to allow confident alignments of noncoding regions, yet sufficiently high to allow for the accumulation of indels. Such requirements limited the number of lineages that could be sampled, and therefore, we are presently unable to extend the generality of our findings to plants or protozoans.

Mutational Input at Larger Scales

To ascertain the mutational input to a genome, the present study focused on small indels occurring in noncoding regions; however, there are several classes of large-scale mutations that can contribute to genome size evolution. For example, whole-genome duplications are a major evolutionary force in many eukaryotic groups (Kellis et al. 2004; Adams and Wendel 2005; Dehal and Boore 2005; Aury et al. 2006), and alternatively, in prokaryotes, large-scale deletions have been detected in both experimental and comparative analyses (Moran and Mira 2001; Nilsson et al. 2005). Because such changes are accompanied by large changes in gene content, they often have a substantial effect on organismal fitness and highly variable fixation probabilities, and therefore, their

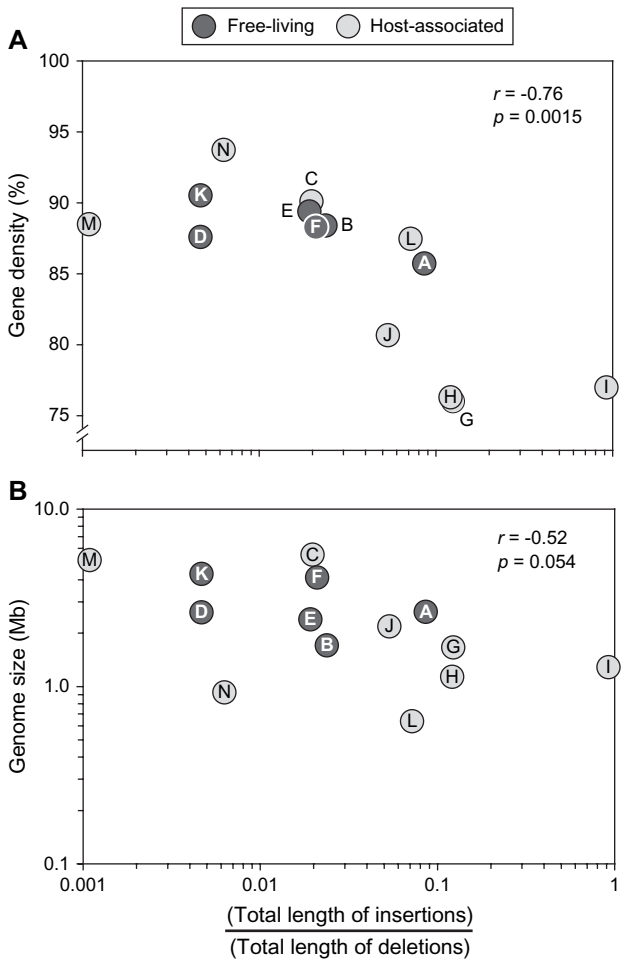

FIG. 3.—Correlation between indel bias and genomic features in prokaryotes. (*A*) Gene density. (*B*) Genome size. Data points are labeled as follows: A, *Sulfolobus*; B, *Methanococcus*; C, *Mycobacterium*; D, *Chlorobium*; E, *Synechococcus*; F, *Bacillus*; G, *Bartonella*; H, *Rickettsia*; I, *Wolbachia*; J, *Neisseria*; K, *Geobacter*; L, *Buchnera*; M, *Escherichia*; and N, *Borrelia*.

incidence cannot fully portray the underlying pattern of mutational events.

Despite the strong bias toward deletions in most prokaryotic genomes, the constant influx of novel genes though lateral gene transfer (Garcia-Vallvé et al. 2000; Gogarten et al. 2002; Lerat et al. 2005) will offset the frequent deletions in noncoding regions and can even lead to large increases in genome size. These newly acquired genes seem to represent the most fluid portion of prokaryotic genomes and are the primary contributor to the observed differences in genome size and gene contents among closely related taxa (von Passel et al. 2008; Kuo and Ochman 2009; Touchon et al. 2009).

Transposable elements represent a special class of mutations that can greatly influence the genome size. For example, data from available genome sequences indicate that the quantity of transposable elements is the major determinant of genome size in eukaryotes (Gregory 2005). Because the proliferation of transposable elements is generally viewed as deleterious, the number of transposable elements within a genome is hypothesized to be under the control of purifying selection and a decrease in effective population size would inevitably lead to genome expansion (Lynch and Conery 2003). However, bacteria appear to be an exception to this rule (Kuo et al. 2009; Novichkov et al. 2009), possibly due to strong deletional biases, as observed in this study. Whereas transposable elements and insertion sequences are observed to proliferate during the initial stage of drift-associated genome reduction in bacteria, these elements are eventually eliminated and are virtually absent from the highly reduced genomes of bacterial symbionts (Moran and Plague 2004).

Evolution of Genome Organization

Among the starkest differences in genome architecture between prokaryotes and eukaryotes is the variation in gene density. The lower bound of gene density in prokaryotic genomes appears to be ~50%, and the vast majority of lineages having a gene density of well over 80% (Kuo et al. 2009). In contrast, the eukaryotic genomes that have been sequenced to date encompass a very wide distribution (Gregory 2005), ranging from about 90% in the microsporidian *Encephalitozoon cuniculi* (Katinka et al. 2001) to less than 2% in humans (International Human Genome Sequencing Consortium 2004). Much of the variation in gene density in eukaryotes is due to the prevalence of transposable elements and introns, whose fixation probability is in turn controlled by the balance between selection and drift (Lynch and Conery 2003; Gregory 2005).

The association between genome size and effective population size among eukaryotes has lead to the hypothesis that elevated levels of drift are the main cause of genome expansion in eukaryotes (Lynch and Conery 2003). Intriguingly, bacteria exhibit the opposite trend, such that genome reduction usually coincides with an increase level of genetic drift (Kuo et al. 2009; Novichkov et al. 2009). Our results suggest that this difference between prokaryotes and eukaryotes is due in large part to the mutational input of insertions and deletions to a genome. With a strong bias toward deletions, DNA segments that do not contribute to organismal fitness in prokaryotic genomes are likely to be purged, even in the absence of selection. And because drift promotes the fixation of slightly deleterious mutations, which are likely to instigate gene inactivation in gene-rich prokaryotic genomes, a reduction in effective population size (e.g., by switching from a free-living to an obligate endosymbiotic lifestyle) can lead to the loss of function in many nonessential genes. Subsequently, these newly formed noncoding regions are removed through the mutational bias toward deletions, thereby maintaining high gene densities.

Although nonadaptive processes, such as biases in mutational input and genetic drift, appear to be dominant forces that influence the evolution of genome size, natural selection will also govern the overall size of some genomes. The genome reduction observed in certain marine bacteria has been attributed to selection for decreased cell volume and energetic efficiency in light of limiting nutrients (Dufresne et al. 2005; Giovannoni et al. 2005). And aside from the selective constraints imposed on the proliferation of transposable elements (Lynch and Conery 2003) and introns (Lynch 2006a), other mutations with extremely small

effects on organismal fitness will be influenced by selection when population sizes are sufficiently large. For example, *E. coli* is thought to have a large effective population size when compared with other bacteria (Kuo et al. 2009), and this species also exhibits the strongest deletional bias among the prokaryotes examined in this study (fig. 1 and table 2). These observations, along with the relatively rapid removal of pseudogenes in *E. coli* (Lerat and Ochman 2004), might be taken to indicate that positive selection is operating on small-scale deletions to foster the elimination of pseudogenes. Although pseudogenes are generally considered to be selectively neutral, this suggests the possibility that the presence of pseudogenes incurs some detrimental effects, such as the energetic costs associated with their transcription and translation or the potential hazard of synthesizing anomalous proteins.

Supplementary Material

Supplementary figure S1 is available at Genome Biology and Evolution online (http://www.oxfordjournals.org/our_journals/gbe/).

Funding

This work was supported by the National Institutes of Health grant [GM56120 to H.O.].

Acknowledgment

We thank B. Nankivell for administrative assistance and preparation of the figures.

Literature Cited

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Curr Opin Plant Biol. 8:135–141.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Andersson JO, Andersson SGE. 2001. Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes. Mol Biol Evol. 18:829–839.

Aury JM, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. Nature. 444:171–178.

Bennett MD. 1972. Nuclear DNA content and minimum generation time in herbaceous plants. Proc R Soc Lond B Biol Sci. 181:109–135.

Bennett MD. 1987. Variation in genomic form in plants and its ecological implications. New Phytol. 106:177–200.

Bensasson D, Petrov DA, Zhang DX, Hartl DL, Hewitt GM. 2001. Genomic gigantism: DNA loss is slow in mountain grasshoppers. Mol Biol Evol. 18:246–253.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. 2008. GenBank. Nucleic Acids Res. 36:D25–30.

Bergthorsson U, Ochman H. 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. Mol Biol Evol. 15:6–16.

Cavalier-Smith T. 1982. Skeletal DNA and the evolution of genome size. Annu Rev Biophys Bioeng. 11:273–302.

Christie KR, et al. 2004. *Saccharomyces* Genome Database. SGD. provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. Nucleic Acids Res. 32:D311–D314.

Chung J, Lee JH, Arumuganathan K, Graef GL, Specht JE. 1998. Relationships between nuclear DNA content and seed and leaf size in soybean. Theor Appl Genet. 96:1064–1068.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol. 3:e314.

Denver DR, Morris K, Lynch M, Thomas WK. 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. Nature. 430:679–682.

Dufresne A, Garczarek L, Partensky F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. 6:R14.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Froula JL, Francino MP. 2007. Selection against spurious promoter motifs correlates with translational efficiency across bacteria. PLoS One. 2:e745.

Garcia-Vallvé S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. Genome Res. 10:1719–1725.

Giovannoni SJ, et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 309:1242–1245.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19:2226–2238.

Graur D, Shuali Y, Li WH. 1989. Deletions in processed pseudogenes accumulate faster in rodents than in humans. J Mol Evol. 28:279–285.

Gregory TR. 2002. Genome size and developmental complexity. Genetica. 115:131–146.

Gregory TR. 2003. Is small indel bias a determinant of genome size? Trends Genet. 19:485–488.

Gregory TR. 2004. Insertion-deletion biases and the evolution of genome size. Gene. 324:15–34.

Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. Nat Rev Genet. 6:699–708.

Hubbard TJP, et al. 2009. Ensembl 2009. Nucleic Acids Res. 37:D690–D697.

Hulsen T, Huynen M, de Vlieg J, Groenen P. 2006. Benchmarking ortholog identification methods using functional genomics data. Genome Biol. 7:R31.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature. 431:931–945.

Jockusch EL. 1997. An evolutionary correlate of genome size change in plethodontid salamanders. Proc R Soc Lond B Biol Sci. 264:597–604.

Katinka MD, et al. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature. 414:450–453.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 428:617–624.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. DOI: 10.1101/gr.091785.109.

Kuo CH, Ochman H. 2009. The fate of new bacterial genes. FEMS Microbiol Rev. 33:38–43.

Lerat E, Ochman H. 2004. Psi-Phi: exploring the outer limits of bacterial pseudogenes. Genome Res. 14:2273–2278.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3:e130.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Lynch M. 2006a. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol. 60:327–349.
Lynch M. 2006b. The origins of eukaryotic gene structure. Mol Biol Evol. 23:450–468.
Lynch M, Conery JS. 2003. The origins of genome complexity. Science. 302:1401–1404.
Mikkola R, Kurland CG. 1991. Is there a unique ribosome phenotype for naturally occurring *Escherichia coli*? Biochimie. 73:1061–1066.
Mira A, Ochman H, Moran NA. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. 17:589–596.
Moran NA, Mira A. 2001. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. Genome Biol. 2:research0054.1–research0054.12.
Moran NA, Plague GR. 2004. Genomic changes following host restriction in bacteria. Curr Opin Genet Dev. 14:627–633.
Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. 2006. The 160-kilobase genome of the bacterial endosymbiont *Carsonella*. Science. 314:267.
Nilsson A, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JCD, Andersson DI. 2005. Bacterial genome size reduction by experimental evolution. Proc Natl Acad Sci USA. 102:12112–12116.
Novichkov PS, Wolf YI, Dubchak I, Koonin EV. 2009. Trends in prokaryotic evolution revealed by comparison of closely related bacterial and archaeal genomes. J Bacteriol. 191:65–73.
Petrov DA. 2002. Mutational equilibrium model of genome size evolution. Theor Pop Biol. 61:531–544.
Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol Biol Evol. 15:293–302.
Petrov DA, Lozovskaya ER, Hartl DL. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature. 384:346–349.
Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. Science. 287:1060–1062.
Robertson HM. 2000. The large *srh* family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res. 10:192–203.
Stajich JE, et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12:1611–1618.
Touchon M, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5:e1000344.
Tweedie S, et al. 2009. FlyBase: enhancing *Drosophila* Gene Ontology annotations. Nucleic Acids Res. 37:D555–D559.
Vinogradov AE. 2004. Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? Curr Opin Genet Dev. 14:620–626.
von Passel MWJ, Marri PR, Ochman H. 2008. The emergence and fate of horizontally acquired genes in *Escherichia coli*. PLoS Comput Biol. 4:e1000059.
von Passel MWJ, Smillie CS, Ochman H. 2007. Gene decay in archaea. Archaea. 2:137–143.
Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics. 25: 1189–1191.

George Zhang, Associate Editor

Accepted June 20, 2009