The dissection of natural selection and neutral processes remains a core problem for molecular evolutionary biologists. One of the longest-standing controversies concerns the causes of genome base composition, notably the variation in the sum of G and C content (GC) between 17% and 75% in bacteria. Sueoka argued very early that GC content variation is driven by mutational biases and, as this bias affects non-synonymous sites, protein evolution might also be largely driven by neutral forces [1]. Later, Muto and Osawa showed that 4-fold degenerate positions in codons exhibit the largest range of GC content (GC4), whereas the non-degenerate second codon positions (GGC) exhibit the narrowest (Figure 1) [2]. As the footprint of genomic GC variation is most evident in those sites under the least selective constraint for amino acid composition, it has become accepted that GC content variation is primarily driven by neutral mutational effects and has little adaptive relevance [2].

Two papers in the current issue of PLoS Genetics aim to test whether the variation in bacterial genomic GC content results directly from mutation biases. Far from observing variation in mutational patterns concordant with the range of GC content, Hildebrand et al. [3], and Hershberg and Petrov [4] independently point to a strong and consistent AT pressure on bacterial genomes, whereby de novo GC → AT mutations arise much more commonly than the reverse. Hershberg predicts that most bacterial genomes, if left entirely vulnerable to mutation, would approach an equilibrium GC content of 20%–30%, close to the highly reduced genomes of endosymbionts [5]. Discounting a rather implausible scenario whereby nearly all diverse GC-rich taxa are converging towards a low GC content, one is forced to conclude that the excess A and T generated by mutation bias (AT pressure) is lost over time. If so, mutational patterns are not strongly shaping genomes after all, and something else is keeping GC contents up.

Hildebrand and co-workers analyze polymorphism data from 149 phylogenetically diverse species corresponding to a wide range of GC content. A major strength of this analysis is that it tests for a number of possible confounders that might explain the excess of GC → AT changes, including variation in mutation rates, sequencing errors, and violations of the infinite sites assumption. The proportion of GC → AT changes that are GC → AT (Z) is almost always >0.3, and is positively correlated with GC4. This means that AT pressure is strongest in GC-rich genomes. For the most GC-poor genomes, the ratio is reversed (Z<0.5), but this might result from violation of the infinite sites assumption at extreme GC content. In fact, the extreme AT-rich genomes of Buchnera do have Z = 0.5 [6].

Hershberg and Petrov exploit full genome data of five very recently evolved “clonal pathogens”, presumably under relaxed selection, allowing precise detection of mutational patterns. This more limited dataset includes no extreme GC-poor genomes. On the other hand, the availability of a large number of SNPs and of an outgroup allows the comparison of patterns within and between species. Consistent with the results of Hildebrand et al., Hershberg and Petrov find an excess of GC → AT mutations in synonymous, non-synonymous, and intergenic sites. Comparisons with the outgroup species suggest this is not caused by loss of repair genes, and that it abates over greater phylogenetic distances (i.e., between “species”). This pattern is similar to that previously found in E. coli [7], and reflects the action of purifying selection (or a process that mimics selection) preferentially removing AT-enriching mutations over time. Hershberg and Petrov’s study also highlights the significance of weaker purifying selection in newly emerged pathogens, as shown in Shigella strains [7]. Strikingly, they find no evidence for a correlation between predicted GC contents at mutational equilibrium and extant base composition, suggesting that mutational bias might have no role in shaping genome composition. Hildebrand et al. show a similar qualitative bias, but predicted equilibrium values vary between 5% and 90% GC. As methods and datasets differ in the two studies, further analyses will be required to shed light on this issue.

Taken together, the evidence for a common mutational pressure towards low GC is clear. The process maintaining base composition in GC-rich genomes must be very strong, because a genomic GC content of 75% corresponds to a GC4 of nearly 100% (Figure 1). This represents a ~70% gap with Hershberg and Petrov’s predicted mutational equilibrium. Two distinct processes might be at work: biased gene conversion (BGC) and natural selection.

In certain eukaryotes, BGC results from recombination between heterologous sequences preferentially removing AT polymorphisms [8]. Contrary to sexual eukaryotes, allelic recombination in bacteria requires horizontal transfer. As a result, rates of recombination between, and even
that are under selection, e.g., codon usage bias, regulatory signals, etc. Naturally, this is an idealized view of genomes that code for many additional overlapping signals.

Patterns are AT-rich relative to genome composition, and there are no neutral positions. The composition of 4-fold degenerate positions results from selection for GC content, the mutational pressures, which show additional AT pressure (grey area) [19]. In the selectionist view (blue), the selection patterns are modified in bacteria that lose repair genes, such as mutators, which show additional AT pressure (grey area) [19]. In the selectionist view (blue), the composition of 4-fold degenerate positions results from selection for GC content, the mutational patterns are AT-rich relative to genome composition, and there are no neutral positions. Naturally, this is an idealized view of genomes that code for many additional overlapping signals that are under selection, e.g., codon usage bias, regulatory signals, etc.

doi:10.1371/journal.pgen.1001104.g001

Figure 1. The GC composition of genomes is strongly correlated with second codon (GC2) and 4-fold degenerate positions (GC4) [2]. Second codon positions show low variability due to purifying selection on non-synonymous changes. 4-fold degenerate positions vary between 5% and 97% GC among published genomes. In the classical neutral scenario (red), 4-fold degenerate positions are nearly neutral and their composition results essentially from mutational patterns. These patterns are modified in bacteria that lose repair genes, such as mutators, which show additional AT pressure (grey area) [19]. In the selectionist view (blue), the composition of 4-fold degenerate positions results from selection for GC content, the mutational patterns are AT-rich relative to genome composition, and there are no neutral positions. Naturally, this is an idealized view of genomes that code for many additional overlapping signals that are under selection, e.g., codon usage bias, regulatory signals, etc.

With different bacterial species are notoriously variable. Consistent with the action of BGC, ecologically isolated endosymbiots do not recombine and have extremely rich AT genomes [5], and regions of high recombination in E. coli are also GC rich [9]. Yet, Hildebrand et al. found qualitatively similar results when excluding taxa with evidence for recombination. Hershberg and Petrov mostly use nearly clonal genomes and still find a large gap between mutation patterns and genome composition. While available evidence suggests a weak role for BGC in the variation of GC content in bacteria, it is very difficult to completely rule out a role for BGC because it purges AT polymorphisms just like natural selection. As a result, recently emerged pathogens with an excess of AT polymorphisms experience both weakened selection and decreased recombination, both of which could potentially explain a decrease in GC content. More research is needed on the impact of BGC in bacterial genomes.

The alternative to BGC is that high GC contents are selectively maintained. Many explanations for GC content variation have been proposed (summarized in Table 1). GC content variation is most marked at synonymous and intergenic sites. Hence, any selective explanation for this variation forces us to turn the traditional concept of the “neutral site” on its head (Figure 1). In this new view, no single position is evolving neutrally in genomes. As a result, 4-fold degenerate positions are not the closest proxy to mutational patterns, but the result of selection for genomic GC content. If so, we are facing a seismic shift of paradigm in molecular evolution. Detection of adaptive features such as codon bias or amino acid frequencies currently rely on a background null hypothesis assumed to reflect neutrality. Neutral models are also the basis of coalescent-based studies of bacte-
patterns be universally linked to the same biological processes? The ever-expanding sequencing output should soon allow extensive comparative studies to shed a great deal of light on these mysteries.

References